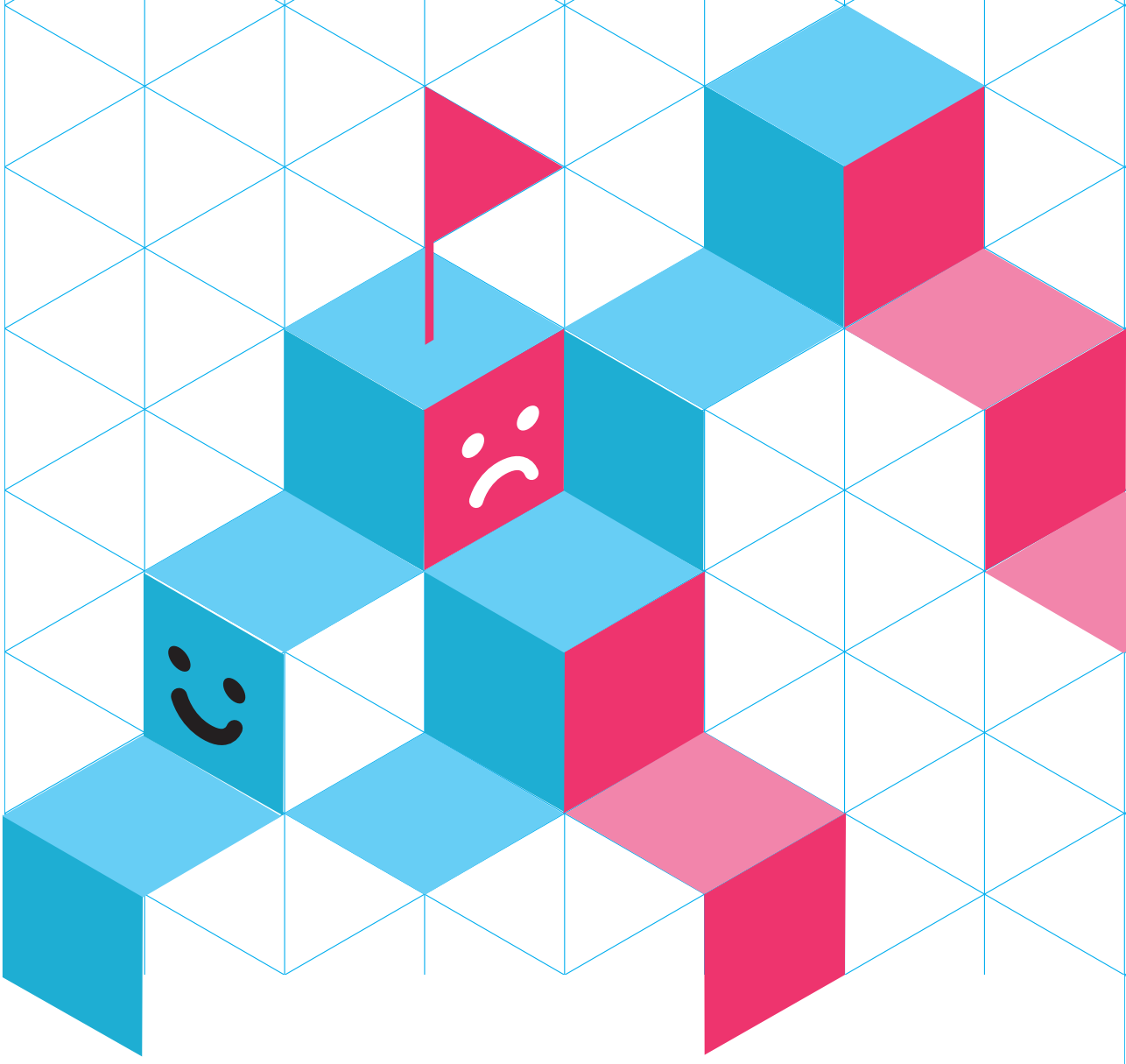


Modelling patterns of time and emotion in Twitter

#anticipointment

Florian Kunneman

The research was supported by the Dutch national research program COMMIT/.

SIKS Dissertation Series No. 2017-11
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover design by Monica Hajek

Printed and bound by Ipskamp Drukkers, Nijmegen

Modelling patterns of time and emotion in Twitter #anticipointment

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 21 maart 2017
om 16.30 uur precies

door
Florian Akky Kunneman
geboren op 7 juli 1987
te Monnickendam

Promotoren:	Prof. dr. Antal van den Bosch
	Prof. dr. Margot van Mulken

Manuscriptcommissie:

Prof. dr. Lidwien van de Wijngaert (Voorzitter)	
Prof. dr. Walter Daelemans	Universiteit Antwerpen (België)
Prof. dr. Hans Hoeken	Universiteit Utrecht
Prof. dr. Geert-Jan Houben	Technische Universiteit Delft
Dr. Barbara Plank	Rijksuniversiteit Groningen

Modelling patterns of time and emotion in Twitter #anticipointment

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Tuesday, March 21, 2017
at 16.30 hours

by
Florian Akky Kunneman
Born on July 7 1987
in Monnickendam (The Netherlands)

Supervisors:	Prof. dr. Antal van den Bosch
	Prof. dr. Margot van Mulken

Doctoral Thesis Committee:

Prof. dr. Lidwien van de Wijngaert (Chair)	
Prof. dr. Walter Daelemans	University of Antwerp (Belgium)
Prof. dr. Hans Hoeken	Utrecht University
Prof. dr. Geert-Jan Houben	Delft University of Technology
Dr. Barbara Plank	University of Groningen

Dankwoord

Dissertaties zijn het meest tastbare resultaat van jaren aan promotieonderzoek, en ik ben blij dat ik nu de laatste hand leg aan de mijne met dit dankwoord. Toch koester ik vooral de jaren waarin dit woord nog nauwelijks uitgesproken hoefde te worden. Ik heb genoten van mijn tijd als promovendus, maar dat kan niet eindeloos duren. *Real artists ship* is dan het adagium, zoals Steve Jobs dit ooit heeft verkondigd. Op deze plek wil ik graag een aantal mensen bedanken.

Te beginnen met degene die voornoemde motto met me gedeeld heeft. Antal, dank voor je inspiratie en alle tijd die je beschikbaar had om met me over onderzoek te praten. Je verfrissende kijk en ideeën hadden op mij altijd een enthousiasmerende werking, en het vertrouwen dat ik van je voel is een enorme steun in de rug. Het ga je goed als directeur van het Meertens; ik hoop dat we niet onze laatste guitar hero sessie hebben gehad.

Margot, ongeveer twee jaar geleden werd je gevraagd als mijn tweede promotor. Ondanks dat je als decaan bepaald geen zeeën van tijd had, pakte je deze rol zonder aarzelen op. Je vragen en theoretische kennis hebben me geholpen om met nieuwe invalshoeken naar de data te kijken. Ik heb genoten van onze meetings, en ben er niet rouwig om dat ze nog steeds voortgang vinden in het onderzoek naar sarcasme in tweets.

Thanks to the manuscript committee, Lidwien, Walter, Hans, Geert-Jan and Barbara, for the time that they made available to read the dissertation, their insightful comments and their presence at the defence. Thanks to Nelleke for joining the Corona at the defence.

Veel dank aan Erik Tjong Kim Sang voor het ontwikkelen en onderhouden van het TwiNL archief. Dit maakte het mogelijk om verschillende toepassingen van tweets te onderzoeken die met de reguliere routes niet mogelijk zouden zijn. Daarnaast wil ik degenen bedanken die tijd beschikbaar hebben gesteld voor het het annoteren van systeem output ter evaluatie.

Naast en binnen mijn promotieonderzoek heb ik met verschillende personen prettig samengewerkt. Christine, ik heb veel schik gehad in ons onderzoek

naar sarcasme en emotie. Dank voor de vermakelijke meetings en de goede samenwerking. Nelleke en Ali, dank voor de prettige samenwerking bij de 'time-to-event' onderzoeken. Rik, op basis van de events die mijn onderzoek opleverde heb je zeer nauwgezet onderzocht hoe ze per type gecategoriseerd kunnen worden, en zelfs een aanzet gegeven voor het voorspellen van demonstraties. Het was een prettige begeleiding, waarin je goed gebruik maakte van IRC. Erkan, thanks for developing the first version of Lama Events, and your lessons in Javascript and Django. It is great having you around. Monica, veel dank voor het mooie ontwerp van Lama events, het prikbord en natuurlijk het omslag van dit boek.

Ik ben verheugd met Ali en Louis aan mijn zijde als paranimfen. Ali, we werkten als PhD binnen hetzelfde project en werkpakket, en hebben beiden behoorlijk wat voetbaltweets geanalyseerd. Uiteindelijk hebben we onze eigen wegen kunnen vinden in de rijke Twitterdata. Dank voor de prettige samenwerking, de vele koffiepauzes en onze aanhoudende strijd om als eerste de deur te openen. Louis, we kennen elkaar sinds de Language en Speech Technology master. Gelukkig is het niet bij dit ene jaar gebleven. Je bent met je hulpvaardigheid, initiatieven en humor een fijne collega en vriend.

Tijdens mijn PhD heb ik de groep van Antal zien groeien van twee promovendi naar de kudde LaMa's die we nu zijn. Iris, je mag met je organisatievaardigheden, zorgzaamheid en onderhoudende beestenverhalen wat mij betreft de titel 'opperlama' dragen. Maarten, je bent het levende bewijs dat fysieke aanwezigheid geenszins noodzakelijk is. Dank dat je altijd klaarstaat op IRC om geduldig te helpen met technische zaken. Ik heb veel van je geleerd over programmeren. Wessel, onze gedeelde interesses voor gamen, twitterdata en bijtijdse lunch doorbreken soms de afstand tussen de 4e en 8e verdieping, en daar ben ik blij om. Maria, je maakt een verbluffende hoeveelheid vlieguren, maar de verhalen waarmee je terugkomt in kamer E4.06 zorgen altijd voor leven in de brouwerij. Alessandro, Ali, Antal, Eric, Folgert, Kelly, Kobus, Louis, Marten, Martin en Suzan, dank voor de goede sfeer en collegialiteit. Sorry dat ik dikwijls degene was die de lunchtijd heeft beknot.

De afdeling CIW is een behoorlijke ratjetoe als het over onderzoek gaat, maar als er maar wat te bakken valt zit het goed. Dank iedereen voor de warme sfeer.

Ik heb vier maanden mogen vertoeven bij de CLIPS groep in Antwerpen, met veel genoegen. Walter, dank dat je deze tijd gefaciliteerd hebt. Ik vond het fijn om met je over wetenschap te praten. Ben, het was bijzonder hoe je me

opgevangen hebt en geïntroduceerd in de wereld van het improv theater. Daardoor voelde ik me gelijk thuis. Chris, Enrique, Giovanni, Guy, Janneke, Mike, Robert en Stephan, dank voor de gezelligheid, Vedett IPA's en de onbegrensde humor.

Er zijn altijd genoeg gelegenheden geweest om het onderzoek neer te leggen. Dank aan de LST PI-group, de CIW PhD's, de Information Foraging groep, COMMIT/, de Infiniti PhD's, ATILA en de Faces of Science voor de mooie meetings en borrels.

Beste Eric, Hielke en Matthijs, Het doet me deugd dat we na de middelbare schooltijd nog zo een hecht contact hebben. Het blijkt moeilijk om de frequentie op pijl te houden, maar als we elkaar zien is het weer als vanouds. Ik hoop dat we op nog veel gelegenheden speeches mogen improviseren.

Lowlands was het periodiek terugkerende zomerse event tijdens mijn PhD, en vooral de mensen waarmee ik dit beleefde hebben het tot een mooie onderbreking van het onderzoek gemaakt. Hoewel de groep nogal in samenstelling varieerde, wil ik vooral Anke, Nicky, Thomas en Franka danken voor de mooie tijd in de Alpha, Bravo, Charlie, Grolsch en op camping 3.

Studentenvoetbalvereniging FC Kunde heeft me niet alleen veel ontspanning gebracht; ik heb er ook veel contacten aan overgehouden die me dierbaar zijn. Thom, Do en Mark, dank voor het samen hangen en de avonden uit. Ik hoop dat er nog vele volgen. Dennis, Nien, Bas, Fien en Roel, jullie waren fantas-tische bestuursleden. Altijd aanstellen. Laurens, Lars, Joep en Dennis, ik heb genoten van onze repetities en optredens vol glamour.

Lieve broer, dank dat je me altijd het goede voorbeeld hebt gegeven, zoals het nastreven van een PhD, en dat je altijd klaarstond om me te adviseren over wiskundige vraagstukken. Ik hecht veel waarde aan onze gesprekken over gamen en het leven, en kijk met graagte terug op onze avonden vol science fiction en Nintendo. Alle goeds aan jou, Kate, Tom en Ina.

Lieve Har en Tien, toen ik na mijn master aan het genieten was van het Nijmeegse studentenleven, hebben jullie me er toe aangezet eens werk te maken van de ambitie om promovendus te worden. Hier ben ik jullie nog altijd dankbaar voor. Jullie bijdrage tijdens de PhD is niet minder geweest. Veel dank voor de interesse in mijn onderzoek, de suggesties en de morele ondersteuning. Ik prijs me gelukkig met zulke liefdevolle en leuke ouders.

Lieve An, je bent het meest waardevolle dat me in de hele PhD-tijd en daarbuiten is overkomen. Je hebt het dappere besluit genomen om te stoppen als PhD, en hoewel ik je mis op de 4e zie ik dat je hier goed aan gedaan hebt. Ik ben

blij dat je nog steeds in de wetenschap werkzaam bent en we onze passie voor onderzoek kunnen delen. Je hebt me ontzettend ondersteund bij de totstandkoming van deze dissertatie. Met tijd en feedback, maar vooral met de nodige humor en vertier. Dank voor deze gemoedelijke basis.

Contents

Als we iets trachten te begrijpen, moeten we het telkens vereenvoudigen en reduceren, en we moeten vooral het prospect om alles te kunnen begrijpen opgeven, teneinde überhaupt nog iets te kunnen begrijpen. Volgens mij is dit van toepassing op alles wat mensen onderzoeken.

Zia Haider Rahman - In het licht van wat wij weten

Chapter 1

Introduction

Twitter[1], the social media platform, started in 2006 as a service to facilitate the sharing of personal updates with friends through SMS (Short Message Service). In accordance with the standard limit of 160 characters imposed on SMS messages, Twitter messages ('tweets') were bounded by a maximum of 140 characters, reserving 20 characters for a unique user address.[2] Twitter swiftly evolved into a medium through which users could 'follow' one another and in which messages were open for anyone to read. While the user base and volume of tweets grew considerably over time, the 140 character limit to date remains effective. Information elements beyond the basic lexical content of messages offer alternative clues to interpret a Twitter post. For instance, the contents of a tweet could be understood by the *point in time* at which it was posted, which is stored as metadata in a post. In addition, the *hashtag* ('#') surfaced as a convention to summarise the topic or emotion of a tweet. Not only are such contextual elements essential for communication through microblogs, they also make it possible to computationally model the relation between words in tweets and their context. An evident utility of such an effort is the automatic detection of real-world events and trending topics in the Twitter stream of messages, as well as a summary of related opinions and emotion. In this thesis, we study approaches to leverage temporal information and hashtags in tweets, in order to build a system that continuously generates overviews of public events, and that is sensitive to event aspects such as periodicity and emotion.

The overarching aim of this thesis is to expose the popular experiences and feelings towards events that can be disclosed from a high volume of tweets. This aim accumulates in '#anticipointment'. As a catchy portmanteau of anticipation

[1] `www.twitter.com`

[2] `http://www.140characters.com/2009/01/30/how-twitter-was-born/`

and disappointment, the word 'anticipointment' is used to either denote the positive expectation of a future event that is followed by a letdown[3] or the anticipation of disappointment itself. As a platform through which experiences are abundantly anticipated and evaluated, Twitter might exhibit anticipointment frequently. The two approaches proposed in this thesis, aimed at detecting events and emotion, ultimately converge into the automatic detection of public anticipointment. The thesis title also points to the primary building blocks of its studies: hashtags and time. In the title, hashtags are reflected by the hashtag marker that precedes 'anticipointment', and time is a necessary anchor for the concepts of anticipation and disappointment, as well as the transition between these emotions as part of anticipointment.

This thesis is divided into three parts: 'Time-based patterns', 'Hashtag-based patterns', and 'Expectations and retrospections on Twitter: Converging hashtags and time'. The first part comprises studies to infer the time interval between tweets and the start of an event, detect real-world events by leveraging either word frequencies over time or explicit time references in tweets, and identify periodic patterns from detected events. The second part explores the use of hashtags to model sarcasm and a variety of emotions. The last part combines the outcomes of the previous two sections into a computational analysis of the emotions before and after a diverse set of automatically detected events.

The studies in this thesis have been conducted as part of the Information Retrieval for Information Services project (INFINITI) within the COMMIT/ research community.[4] All studies have been applied to Dutch tweets harvested within the TwiNL framework.[5] Based on these studies, a system was assembled that processes Dutch tweets as they are posted. Its output is made available as the 'Lama Events' Web application.[6]

In the remainder of the introduction, I will motivate the studies that were conducted as part of this Thesis (Section 1.1), describe the underlying research questions and contributions (Sections 1.2 and 1.3) and provide an outline of the chapters (Section 1.4).

[3] http://nancyfriedman.typepad.com/away_with_words/2011/12/word-of-the-week-anticipointment.html

[4] http://www.commit-nl.nl/projects/information-retrieval-for-information-services

[5] http://www.ru.nl/lst/projects/twinl/

[6] http://lamaevents.cls.ru.nl

1.1 Motivation

1.1.1 Approach

This thesis is set in the Artificial Intelligence tradition to build an intelligent entity (Russell & Norvig, 1995), that in our case takes the form of an Information System. On the most basic level, Information Systems provide human end users with information of value in a given context. Such a system often has access to a voluminous database from which it distills information. The user itself is only presented with the output, and does not have to be proficient in the underlying software or hardware (Wasserman, 1980). We develop software to find information from a stream of tweets, in the form of events and some of their aspects, and provide the user with a sensible overview of automatically detected events.

In aiming to filter and transform the Twitter stream into useful information, our research connects to several fields. The first is the field of Time Series Analysis. As an area of Statistics, Time Series Analysis is a collection of methods that deal with observations made sequentially over time (Chatfield, 2013). In this thesis, the field is predominantly reflected in our effort to highlight words that display 'bursty' behaviour (Kleinberg, 2003), and that thereby may signify real-world events. Second, our research relates to the field of Social Sensing, which aims to make sense of distributed observations about the world. By consulting tweets to detect real-world events, these sensors enable 'social data scavenging', where the participants agree to the fact that their posts are in the public domain, while at the same time they are unaware of how their information is used (D. Wang, Abdelzaher, & Kaplan, 2015). A third field that our research connects to lacks a common denominator, but can be formulated as 'Mining the thought Web': the development of smart tools to interpret the millions of thoughts that are posted on Social Media platforms in real-time and connect them to subjects and events.[7] Within this pursuit, our studies mostly relate to Sentiment Analysis, which encompasses the computational detection of opinion, sentiment, and subjectivity in text (Pang & Lee, 2008). In comparison to Social Sensing, the focus of this field is on thoughts and feelings rather than events in the real world. We aim at mining for both types of information.

Despite its different fields of inspiration, the unifying characteristic of this thesis is that its studies lean heavily on 'Collective Intelligence' or 'The wisdom of the crowd'; the idea that a group as a whole knows more than any selection of experts (Surowiecki, 2005). An estimated total of 303 million tweets were posted

[7] http://techcrunch.com/2009/02/15/mining-the-thought-stream/

per day in January 2016[8] and an average of 732 thousand Dutch tweets could be crawled in the TwiNL framework per day in 2016.[9] Following the common practice on Twitter to share experiences and thoughts of personal importance, the collective of these posts arguably hosts a broad diversity of information about the world in real-time. Collective intelligence is leveraged to automatically detect the versatility and prominence of events and words to express emotions in Twitter. Essential elements in this pursuit are *contextual anchors*.

Contextual anchors can be defined as elements in a dataset that aggregate instances by a certain characteristic. Examples of anchors in Twitter are the geolocation that anchors tweets to their location, the user name that anchors tweets to their author, and the URL that relates tweets to information sources outside of Twitter. Anchors can be used to learn more about a target in a data-driven fashion. For example, by aggregating tweets that share a certain location from which they were posted, information about this location can be deduced from the remaining data in these tweets, such as the text. The contextual anchors that are explored in this thesis are time and hashtags.

Time is a useful anchor to detect events. While events can take many forms, we focus on events that take place within the limits of a single date. In accordance, tweets can be aggregated by their reference to the same date, and events can be identified as prominent words or phrases in these tweets. Effectively, these words or phrases might describe various types of events. We aim to detect events of any type, but do focus on significant events, e.g. events that might be reported in the news media (McMinn, Moshfeghi, & Jose, 2013) and that are hence of interest to a large group of people. We studied two approaches to link tweets to a date and thereby detect significant events.

Hashtags are a potentially useful basis to model the language of emotions and rhetorical figures. Twitter users commonly deploy them to refer to a topic or event, to explicate the valence of a tweet (#sarcasm) or to describe the emotion that they feel when writing a tweet. From a computational perspective, a hashtag can be leveraged to train a model of the words that link strongly to the underlying concept referred to by the hashtag. The model can henceforth be used to detect the presence of the concept automatically in tweets that do not include the hashtag. Such an endeavour is valuable, given the proportion in which hashtags are deployed on Twitter. About a quarter of all tweets[10] include

[8] `http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2`

[9] It should be noted that this average was a lot higher up to the summer of 2014, but has since then been in decline (see for the exact numbers in the case of TwiNL `http://145.100.59.120/counts/` or navigate from `http://twiqs.nl`).

[10] Based on 27.7 million, predominantly Dutch, tweets, extracted from TwiNL (see Section 1.1.3).

a hashtag. A data set based on tweets with hashtags may offer predictive information that can be applied to tweets without a hashtag. As an alternative to labelling examples of a target manually or specify the type of words to focus on, this approach makes it possible to model with little effort any concept that is sufficiently clearly denoted with a hashtag. In case a hashtag denotes an emotion or rhetorical figure, the words that surround this hashtag ideally reflect the prototypical contexts in which the emotion or rhetorical figure can be expressed, as well as the degree to which any word is linked to it.

While the framework of our studies is constituted by these data-driven approaches, knowledge-driven components are indispensable for their implementation. For example, in order to link tweets to a date and detect events, we formulate rules to specify the linguistic phrases that might refer to a date. In addition, the hashtags that we use to train an emotion or rhetorical figure are selected based on human knowledge. Hence, these studies manifest how knowledge-driven and data-driven components can best be combined in pursuit of information extraction.

1.1.2 Utility

The approaches that we study are aimed at the extraction of a broad overview of events and their characteristics from tweets as they are posted. A system with a broad and open scope has a number of benefits. First, its output is of potential interest to several target groups. It can assist tourists who want to know about popular events in a specific location and time frame, as well as organisers of an event or product release who want to monitor the popularity of their event in time, possibly in relation to competitive events. Intelligence agencies may also be interested in a system that can timely foresee a social gathering of potential threat to public security. Although the output is not designed for any specific target group, the flexibility of the system makes it possible to easily post-process and tune the extracted patterns to the demands of any target group.

A second benefit is the possibility to compare events on various dimensions, such as the degree of popularity, the level of aggression, or disappointment, and rank them accordingly. The quantity and diversity of events allow for ranking and highlighting the most extreme cases, providing insights on a broad level. In the current thesis, this opportunity is explored in studies on periodicity detection and the detection of emotion in tweets that relate to a large number of events.

Third, the system does not restrict the words that are used to refer to an event or express an emotion, and may thereby disclose output that is not found by any specialised system. In such specialised systems, the features to find a target are to a larger extent pre-specified. In contrast, our system makes no assumptions about the relative importance of input features and may discover strong predictors in features that we had not expected.

Finally, the information that is highlighted based on our data-driven approaches can be used to study the concept that they link to. For example, the use of sarcasm on Twitter can be studied after modelling the words that are used in the context of '#sarcasm'. Such a procedure complements existing manual corpus studies of sarcasm, or verbal irony (Burgers, van Mulken, & Schellens, 2011).

1.1.3 Data

Throughout this thesis we make use of the tweets that are collected within the TwiNL framework.[11] These tweets are collected continuously from the Twitter API from December 2010 onward on the basis of a seed list of Dutch words and a dynamic list of the most active Dutch Twitter users (Tjong Kim Sang & van den Bosch, 2013). The harvesting is limited; TwiNL harvests an estimated half of all Dutch tweets. Nonetheless, the benefit of TwiNL is that it keeps a record of tweet IDs since December 2010, while the Twitter search API currently only returns tweets that were posted in the most recent seven days.[12] The long span of tweets in TwiNL enables us to research periodic event patterns and makes it possible to collect many tweets by querying for certain hashtags. For example, querying TwiNL for tweets that mention #not in a window spanning December 2010 until January 2013 yields 353,758 tweets, and a search for tweets with #zinin ('looking forward to it') between December 2010 and October 2015 results in 606,310 tweets.

The use of Twitter data has two important implications. The first is that it might be tempting, based on the high number of tweets that are posted, to relate the patterns that are found in Twitter to the society as a whole. Studies have shown that the demographics of Twitter users do not reflect the demographics of a society (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). The high number of posts do allow, however, for relatively accurate measurements of public opinion and approximate, for example, the outcomes of an election (Tjong Kim Sang, 2011; DiGrazia, McKelvey, Bollen, & Rojas, 2013; Sanders & van den

[11] http://www.ru.nl/lst/projects/twinl/
[12] https://dev.twitter.com/rest/public/search

Bosch, 2013), but this should not be seen as a consistent quality (Gayo-Avello, 2012).

Second, Twitter is a volatile data source. The Twitter terms of service[13] state that users by default agree to post their messages in the public domain, which means that tweets can be read by anyone and harvested by third parties, as far as Twitter chooses to share its data. Users are allowed to remove tweets or make their account private, after which their tweets are invisible for any party outside their follower circle. Researchers are allowed to keep a database of tweets that they harvested, but cannot share tweets with other researchers. However, the unique IDs of tweets can be shared, based on which the content of the tweets can be queried from the Twitter API. Any tweet that has been removed or screened will not be returned from the API, which means that any research that is conducted based on Twitter posts can not be reproduced with the original dataset. In a study into indicators of tweet deletion, Petrovic, Osborne, and Lavrenko (2013) experimented with a data set of 75 million tweets, of which 2.4 million tweets (3.2%) were deleted within a month time. For this reason, our research is aimed at developing a system that processes tweets in real-time. The outcomes of our evaluations are meant to be descriptive of such a context. We do share tweet IDs of every study, as well as the information that has been extracted from them.[14]

1.2 Research Questions

The main problem statement that drives the studies in this thesis is the following:

> **PS:** How can we discover time-anchored events and their characteristics from a stream of Twitter messages?

As motivated in Section 1.1.1, the events that we target typically take place on a single day, and are thereby anchored in time. Research questions 1 and 2 address the detection of such events from Twitter. The remaining research questions drive the work to analyse event characteristics.

[13] `https://twitter.com/tos?lang=en`

[14] In each section that discusses a dataset in this thesis, we point to the URL at which we have placed the related tweet IDs and system output.

1.2.1 Part 1: Time-based patterns

The date at which future events take place might be shared in tweets that refer to it. Extracting this information is especially useful to learn about events that are not known to the general public, such as a demonstration or local party. Obviously, it is most useful to extract such information as early as possible. In a stream of tweets that refer to an event, by means of a common hashtag or event phrase, only part of the references to future dates might relate to the event itself. The objective is then to find the most optimal approach to extract the event date, in terms of speed and accuracy:

> **RQ 1:** Given a stream of tweets that refer to the same event, how accurately and early can we infer the number of days until the start of this event?

While a known event can be leveraged to obtain insights into temporal information in tweets, this temporal information can in turn be deployed as starting point to detect events. We compare two approaches to using temporal information for event detection. The first is to observe the time when tweets are posted. The second is to focus on the temporal information that might be given in the text of a tweet. To illustrate the two approaches, we plotted in Figure 1.1 the frequency of tweets that refer to a referendum held in the Netherlands on April 6, 2016, either by just mentioning the word 'referendum' or by mentioning the word 'referendum' along with a future reference to the date of this event. The plots are depicted in the scale of all tweets with this term. The date of the event can clearly be recognised as the point where the number of tweets is magnitudes higher than the frequency at any other point. This reflects that Twitter users are mostly inclined to tweet about an event as it occurs. Based on such behaviour, events can be detected by counting words that are mentioned on Twitter over time, and highlighting words that show a sudden increase. This word frequency, possibly indicating an event, coincides with the time at which the event takes place. We refer to event detection based on this direct indication as 'burstiness-based event detection'.

The line in Figure 1.1 that describes the number of tweets with a textual reference to the event date is barely visible, as it stays close to the horizontal axis. As a comparison, we depicted the same plots in the scale of this line in Figure 1.2. The frequency of all tweets is partly visible in this plot, before shooting through the roof about seven weeks before the event. The number of tweets that refer to the date of the event are more scattered over time and only show a subtle

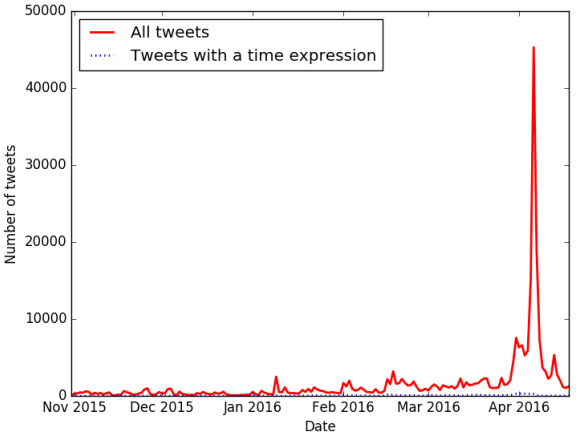


FIGURE 1.1: The number of tweets per day that mention 'referendum' and 'referendum' in combination with a reference to the future date of April 6, 2016, depicted in the scale of the former.

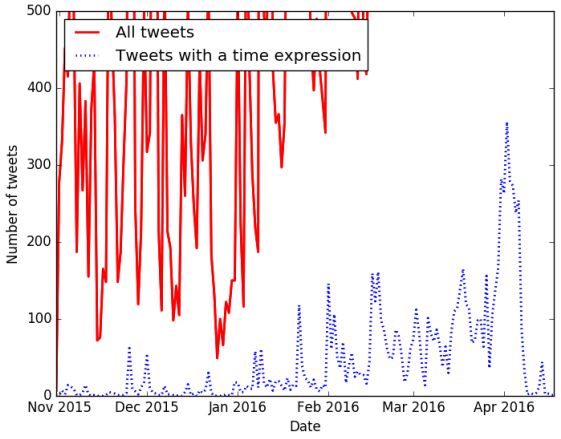


FIGURE 1.2: The number of tweets per day that mention 'referendum' and 'referendum' in combination with a reference to the future date of April 6, 2016, depicted in the scale of the latter.

increase as the event approaches, with a small peak close to the event. Rather than showing a pattern by their time stamp, it is the content of these tweets that reveals a dominant event date. This signal can be leveraged by scanning tweets for the presence of a time expression, following the study linked to RQ1, and subsequently finding words that potentially describe an event in the remainder of these tweets. We refer to this more indirect approach to event detection as 'time reference-based event detection'.

The development of both burstiness-based and time reference-based event detection systems is driven by the following two research questions:

RQ 2a: To what extent can events be detected from Twitter by means of

burstiness-based event detection?
RQ 2b: To what extent can events be detected from Twitter by means of time reference-based event detection?

Based on the effort to automatically detect events as part of RQ 2, we can subsequently apply event detection to the long span of tweets that is available in the TwiNL framework. The resulting set of events makes it possible to zoom out and move the attention from temporal information in tweets to temporal information in events. One follow-up task is to identify events that recur periodically, so that we can predict their next recurrence even without following the Twitter stream. We compare two approaches to identify periodic events in a bottom-up fashion. The first is based on the intervals in days between occurrences of similarly described events. The second is driven by calendar information. We ask the following research question:

RQ 3: Given a set of detected events from Twitter over an extended period of time, to what extent can periodic events be identified?

1.2.2 Part 2: Hashtag-based patterns

Shifting our focus from time to hashtags as anchor within the Twitter stream, we performed experiments to test the usefulness of hashtags in modelling the language of sarcasm and emotions. Although emotions have a more overt relation to events than sarcasm, sarcasm detection can potentially be used as a filter for sentiment detection. We argue that the usefulness of a hashtag depends on two factors. The first is whether a hashtag is deployed in a consistent way by Twitter users. For example, the hashtag '#funny' might denote a joke or a strange situation, or be used in a sarcastic way. A high ambiguity makes it difficult to learn the context of a hashtag, and renders the automatically learned detector useless. The second factor is whether the hashtag reflects the emotion that is present in the rest of the words, or whether it adds the concept to the message that in itself does not express the emotion. To illustrate, consider these example tweets:

1. Doing homework #boring
2. I would rather do something else, but unfortunately I have to do my homework #boring

In the first example, the phrase 'doing homework' does not convey the lack of excitement of the author about this activity. The boredom is purely reflected in the hashtag. In contrast, the textual part of the second example clearly conveys a certain boredom about the prospect of doing homework, which is only repeated or strengthened by the hashtag. In order to train a model of the emotion of boredom, usage of the hashtag as in the second example is obviously preferable over the first. As it is not clear beforehand how useful any hashtag is as training label, we empirically test the usefulness for several hashtags. This is related to the research question below:

> **RQ 4:** To what extent can figurative speech or emotion in tweets be detected based on hashtag-annotations?

1.2.3 Part 3: Expectations and retrospections on Twitter: Converging hashtags and time

We can combine the insights gained from answering RQ 2 and RQ 4 to detect the emotion in tweets that relate to automatically detected events. One application of such a combination is the detection of emotions that relate to expectations and actual experience. We set out to study occurrences of anticipointment by automatically detecting the emotion of positive expectation in tweets before events, and disappointment in tweets after events. By also including satisfaction, we can research whether a correlation exists between these three emotions:

> **RQ 5:** What is the strength of the correlation between the collective expression of positive expectation before events and disappointment and satisfaction after events on Twitter?

1.3 Thesis Contributions

This thesis aims for a number of contributions to the fields mentioned in Section 1.1. First, it presents three new tasks: identifying the number of days until an event starts from a stream of event tweets, detecting periodicity from Twitter events and comparing the emotion before and after events as expressed on Twitter. Second, it provides further insights on the widely studied tasks of event detection, sarcasm detection and emotion detection. Third, all studies in this thesis give insight into system performance in a realistic setting of streaming Twitter

posts, where the output is automatically ranked in order to select output for manual evaluation. Fourth, the components that were developed and tested in the separate studies are assembled into the Lama Events Web application[15] that processes tweets in real-time.

1.4 Thesis Outline

Part I comprises studies on time-based patterns. In Chapter 2 we start working with hashtags that are strongly linked to known football events, in order to examine the Twitter text that surrounds such a hashtag for information on the event date. Using both time features and word n-gram features, we compare the performance of a rule-based system and machine learning on accurately estimating the time-to-event at any point in a stream of hashtag-aggregated event tweets. We subsequently evaluate performance on a set of open-domain events. In Chapter 3, we start from two months in the total Twitter stream of TwiNL and set out to identify hashtags and event phrases that signify an event, based on the sudden rise in frequency of such tokens. In Chapter 4, we pursue the same goal, but identify events as the induced date of forward referring time expressions that link strongly to a certain word or phrase. In Chapter 5, we set out to identify periodic events by applying event detection on a span of over four years and looking for periodic intervals or calendar patterns related to similar event terms.

Part II features research on hashtag-based patterns. In Chapter 6, we select a combination of hashtags that might be deployed by Twitter users as a marker of sarcasm that flips the polarity of their message. We test the effectiveness of these hashtags as training label for machine learning on a total day of Twitter messages, and analyse the characteristics of sarcasm in tweets. In Chapter 7, we extend the procedure of Chapter 6 to a total of 24 hashtags that link to several emotions, and assess the usefulness of these hashtags as annotation label for emotion detection.

In Part III, we combine the lines of research in Part I and Part II in a study of the emotions before and after a wide range of events. We train a classifier to recognise positive expectation, disappointment and satisfaction, and apply it to a large number of pre-event and post-event tweets. For each of the emotions, an aggregate emotion score is derived per event based on the tweet classifications. We make use of this aggregate score to calculate the correlation between any of

[15] `http://lamaevents.cls.ru.nl`

the three emotions, highlight the events that manifest the most extreme patterns, and analyse the emotion for a sequence of known events.

Part I

Time-based patterns

> If it is true, though, there's really no such thing as the present, only past and future. Take this moment, for instance: by the time I talk about it, it's already in the past.
>
> Momo
>
> (from Michael Ende - Momo and the Thieves of Time)

Chapter 2

Timely identification of event start dates from Twitter

Based on: Kunneman, F., Hürriyetoğlu, A., Oostdijk, N. & van den Bosch, A. (2014). Timely identification of event start dates from Twitter. *Computational Linguistics in the Netherlands Journal*, *4*, 39-52.

Tweets that mention a future event might provide several textual clues to the start time of this event. We set out to infer the start date of an event based on a stream of Twitter messages related to this event. Taking hashtags or event name expressions as query terms, we gathered a certain number of tweets about an event and used clues in these tweets to estimate at what date the event will start. In addition to temporal expressions with knowledge-based and automatically generated estimations, we included word n-grams as potential markers. The estimation is performed by processing tweets in the order in which they were posted, either with a machine learning classifier or by taking a majority vote over the temporal expressions found in the set of tweets. Results show that temporal expressions are the strongest predictors. The majority-based and machine learning approaches attain equal performances when trained and tested on a single event type, football matches, with an average estimation error of 0.05 days; but when tested on a range of different events, the majority voting approach shows to be more robust than machine learning for this task, yielding high performance on all events. Still, per-event differences hint at a context in which machine learning might be beneficial.

2.1 Introduction

A substantial number of posts on Twitter refer to real-world events and report on what is currently happening. While most of these tweets are posted as the event unfolds, some refer to events that have not yet happened. We aim to leverage the latter type of posts in order to identify, in a continuous way and on unseen data, the dates at which future events take place.

A calendar of future events fed by the most recent information available on the social media may be of interest to various user groups. One such group is journalists who would like to know about future events that emerge first on social media. The idea of publishing future calendars with potentially interesting events has been implemented before and is available through services such as Zapaday[1], Daybees[2], and Songkick[3]. However, these services do not mine social media automatically; to our knowledge, based on the public interfaces of these platforms, these services perform directed crawls of (structured) information sources, and identify exact date and time references in posts on these sources. They also manually curate event information, or collect this through crowdsourcing. Using Twitter instead as source of information is of particular interest as it might carry information about events that are not picked up by these services, especially when they are informal, regional, or relatively small-scale.

Identifying event start dates from tweets is challenging, especially as Twitter users who mention a future event can be found to talk about various things, some of which are only very loosely related to the targeted event. Consider the following examples:

1. *Preparing the last few things for my beyoncé concert tomorrow*
2. *86 days to #WC2014 - He's the only man born on 18-March to have scored at the World Cup. Guess who? URL*
3. *The Walking Dead Creator Eagerly Anticipates Mario Kart 8 URL*

In examples 1 and 2, the time of the future event is explicitly mentioned. Yet, example 2 contains two temporal expressions ('86 days to' and '18-March'), but only the first has future reference. Example 3 also refers to a future event (the release of *Mario Kart 8*), but does not contain any information about the release date.

[1] `http://www.zapaday.com`
[2] `http://daybees.com/`
[3] `https://www.songkick.com/`

In this study we compare different methods to identify the start date of a future event, based on tweets referring to the event. We aim for the identification of start dates of open-domain events, but first experiment on a closed set of football events. In a second experiment, we extend the task to events of different types. While temporal expressions (henceforth referred to as TIMEXs) are an obvious feature type for this task, we also investigate the use of word n-grams in a machine learning approach to estimate the start time of an event.

2.2 Related Work

We study the early identification of an event start date from tweets referring to that event. Most of the previous studies on event mentions on Twitter focus on the tweets during or right after an event (Chakrabarti & Punera, 2011; Jackoway, Samet, & Sankaranarayanan, 2011; Becker, Iter, Naaman, & Gravano, 2012; Quezada & Poblete, 2013). In other research, the sudden bursts of tweets with common key terms ('trending topics') are leveraged to detect unknown events (Ozdikis, Senkul, & Oguztuzun, 2012; Qin, Zhang, Zhang, & Zheng, 2013; Weiler, Scholl, Wanner, & Rohrdantz, 2013; Valkanas & Gunopulos, 2013; Chunara, Andrews, & Brownstein, 2012; Zhou & Chen, 2013).

As regards the focus on forward references to events, the studies by Ritter, Mausam, Etzioni, and Clark (2012) and Weerkamp and de Rijke (2012) are most comparable to our research. Ritter et al. (2012) train on open-domain annotated event mentions in tweets in order to create a calendar of future and past events based on explicit date mentions and event phrases recognised by a trained tagger. Weerkamp and de Rijke (2012) study anticipation seen in tweets, and focus on personal activities in the very near future.

Time-to-event (TTE) estimation of football matches has been the topic of several studies. Kunneman and van den Bosch (2012) show that machine learning methods can differentiate between tweets posted before, during, and after a football match. Estimating the time to event of future matches from tweet streams has been studied by Hürriyetoğlu, Kunneman, and van den Bosch (2013) and Hürriyetoğlu, Oostdijk, and van den Bosch (2014), using local regression over word time series. In a related study, Tops, van den Bosch, and Kunneman (2013) use support vector machines to classify the time to event in automatically discretised categories. At best, the systems described in these studies are within a day off in their predictions, optimally 8 hours for Hürriyetoğlu et al. (2014), but they remain within a single type of event, football matches. We will take football matches as a first step, but then in addition move to events of different types.

2.3 Football Events

As a first step we carried out a controlled case study in which we focused on Dutch premier league football matches as a type of planned event. Football matches provide useful data, as they occur frequently, have a distinctive hashtag by convention ('#ajafey' for a match between Ajax and Feyenoord), and typically generate thousands to several tens of thousands of tweets per match.

2.3.1 Experimental Set-up

Data

As data for our experiments we selected 60 football matches played in the Dutch premier league. We harvested tweets from the TwiNL database. We selected the (average) top 6 teams of the league,[4] and queried all matches played between these teams in 2011 and 2012. For each query, the conventional hashtag for a match was used with a restricted six-week search space, viz. three weeks before the match until three weeks after, so as to avoid overlap with another match between the same two teams. The queries resulted in tweet streams for a total of 60 matches; a total of 703,767 tweets. Of these, 269,753 are posted before event time. The number of tweets per event ranged from 321 to 35,464. Retweets were removed, as they only duplicate data and may pass on a previous tweet with a different TTE. The resulting set without retweets contains 411,784 tweets, of which 140,060 are posted before event time.

Every tweet in our data set has a time stamp of the moment it was posted. Moreover, for each football match we know when it took place. This information is used to calculate for each tweet the actual time that remains to the event in terms of the number of days and the error in estimating the TTE.[5]

Features

In the feature space we experiment with different (combinations of) feature types. In line with the task of TTE identification, TIMEXs are the primary source of information that is leveraged. We also include word n-grams to see whether other types of information might contribute to the estimation accuracy. We deliberately omit domain information like the distribution of days at which football matches are played, in line with our aim of open-domain TTE estimation.

[4] In 2011 and 2012, these teams were Ajax, Feyenoord, PSV, FC Twente, AZ Alkmaar and FC Utrecht.

[5] The tweet IDs for all 60 events, along with their calculated time to event, can be downloaded from `http://dx.doi.org/10.17026/dans-z8e-3uqb`.

TIMEX	English translation	Median TTE in days based on training data
Weekeinde	Weekend	1
Komende zondagmiddag	Next Sunday afternoon	2
Nog maar een paar daagjes	Only a couple of days	4
Nog maar 2 weken	Only 2 weeks	14

TABLE 2.1: Examples of the median TTE in days for TIMEXS as calculated from training tweets.

For the extraction of TIMEXs we make use of the extensive list of Dutch TIMEXs that was compiled by Hürriyetoğlu et al. (2014)[6]. Although Heideltime (Strötgen & Gertz, 2010) has a module to extract Dutch TIMEXs, the list of TIMEXs compiled by Hürriyetoğlu et al. (2014) is more extensive and tuned to our research aims. In line with the approach of Hürriyetoğlu et al. (2014), we estimated the TTE of TIMEXs in two ways: trained and rule-based.

In the trained approach, the TTE estimation linked to a TIMEX is derived in a data-driven way, which we call 'Timelearn'. By collecting all occurrences of a TIMEX and the observed TTE in days during training, we can calculate the median of this set, which we then take as the TTE estimate attributed to the TIMEX. We excluded TIMEXs of which the observed TTE had a standard deviation higher than 2 (i.e. two days). Examples of trained TTE estimations are given in Table 2.1. Specific TIMEXs ('only 2 weeks') are indeed linked to the specified TTE based on the training data. The added value of this approach is the estimation for less specific TIMEXs (such as 'only a couple of days').

In addition to a data-driven estimation, the TTE from a TIMEX can be estimated manually by common-sense annotation. Hürriyetoğlu et al. (2014) distinguish two kinds of rules: dynamic and exact. Dynamic rules apply to TIMEXs of which the TTE is dependent on the point in time at which they are posted. A mention of a date should be linked to the date of posting to calculate the difference, and a mention of a weekday should be linked to the weekday at which it was mentioned. The example 'next Sunday afternoon', that is given in Table 2.1, would be calculated as the first Sunday from the day of tweeting (three days if Thursday is the day of posting). Exact TIMEXs imply a TTE that can be estimated without information on the moment at which they are posted. Examples are 'tomorrow' and 'another 12 days before'.

Because Hürriyetoğlu et al. (2014) focus on estimating the TTE in terms of the number of hours within a frame of eight days, the list of TIMEXs does not

[6]The list has been made available through `http://www.ru.nl/lst/resources/`.

include any expression that relates to longer periods of time. Therefore, we complemented the list with a number of additional expressions with a longer range in time. For example, expressions like 'nog [1-8] dagen' ('only [1-8] days') were extended to a range of 21 days, and also expressions like 'over 3 weken' ('in three weeks') were included.

The third type of features, word n-grams, were extracted after preprocessing of the data. All characters in the tweets were lowercased and user names and URLs were normalised into the dummy features 'USER' and 'URL'. The tweets were tokenised with Ucto[7] and surrounded by beginning- and end-of-tweet markers. We extracted word unigrams, bigrams, and trigrams from the tweets.

With these three feature types (trained and rule-based TTE estimates and word n-grams) we tested all possible permutations: the three feature types in isolation, combinations of two, and all three combined.

Representing time windows

Given a stream of tweets that refer to a specific event and that were identified on the basis of a common hashtag or event name query, we aim to extract the start date of the event as early as possible. Although the correct TTE differs depending on the time at which a tweet is posted, the task is to infer from every tweet that is encountered the date at which the event will take place. In order to smooth any noisy temporal information in tweets (such as a TIMEX that does not refer to the targeted event), we chose to aggregate tweets by means of a sliding window in time, rather than judging the start date from any single tweet. The length of the window can be defined in terms of time or in terms of the number of tweets. We chose the latter option, to ensure an approximately equal portion of information at each window. This means that the time period that a window encompasses may vary. Typically, the period will be shorter as the event start time draws nearer and more people start to tweet about the event.

Different window sizes can be chosen. Long windows might lead to more accurate estimations, but sampling a long window takes longer: with a window size of 100 tweets, the first estimation can only be made after this many tweets have been seen. In view of the scarcity of tweets that are posted many days before an event takes place, an estimation based on a large window might

[7] `https://languagemachines.github.io/ucto/`

be made only right before or even after the event start time. From this perspective, smaller windows are favourable. To test for the optimal window setting we alternated three different window sizes: 50, 20 and 10 tweets.

The steps by which a window slides forward can range from large steps (e.g. the step size being equal to the window size) to tweet-by-tweet. Overlapping steps lead to more frequent estimations and more training instances for machine learning (ML), while the computational load will be higher. We tried three step sizes, each a fraction of the window size: $1/5, 1/2$ and $1/1$. Thus, we applied our methods on nine window and step size combinations.

For training and testing, the windows of tweets were given a label based on the actual TTE of the last tweet in the window. Windows of which the last tweet was posted during or after the event time were given the label 'during' and 'after' respectively.[8] Thus, the labels are any TTE in days from 21 until the day of the event, as well as 'during' and 'after'. The features in a window were derived by aggregating the features of the separate tweets in the window. In the case of rule-based features, the date that was derived from such features was translated into the TTE at the time of the last post of the window.

It is possible that a window of tweets comprises several days. A TIMEX in a tweet that was posted earlier than the last tweet of a window would then give outdated information about the actual TTE at the estimation time. In such cases, we normalised the TIMEX features to the date of the last tweet in a window.

Prediction

We tested the different feature combinations and window settings on the 60 football events by means of 10-fold cross validation, training on the windows in 54 events and testing on the 6 remaining events. For each fold, we trained and tested a Support Vector Machines (SVM) classifier based on libsvm (Hsu, Chang, & Lin, 2003). During training, additional preprocessing was needed to facilitate the classifier. Given that the vast majority of tweets are posted right before, during and after an event, the number of instances per TTE label, or 'during' or 'after', is highly imbalanced. To avoid a classifier bias towards TTE close to event time, we balanced the number of instances per TTE label by a combination of under-sampling and over-sampling. This dual approach limits excessive duplication and removal of instances (Sappelli, Verberne, & Kraaij, 2013). After balancing the data set, we reduced the feature space by selecting the 10,000 most frequent features. We made use of error-correcting output codes (James

[8] 'During' and 'after' are included in the task, as in any realistic forecasting setting it is vital to establish that the event is not in the future, but is actually ongoing or has already happened.

& Hastie, 1998) to obtain a single classification for each window. The different parameters of SVM were tuned by splitting the training data into five folds and performing classifications based on ten random parameter combinations from a grid. The grid tested a linear, polynomial and RBF kernel, different values of C and $gamma$ and different degrees.

The TTE information that was inferred from TIMEXs on the basis of rules and training was used as input to the SVM classifier, but could also be used to make a direct estimation. We implemented a majority voting method that bases its estimation on the most frequent TTE that was derived from the TIMEXs in a window. This method was applied based on two feature combinations: the rule-inferred TTE and a combination of rules and trained TTEs. This results in nine different methods that are compared: SVMs with seven different feature combinations, and two methods that take into account the majority of TTE estimations per window.

Given that our task is to progressively estimate the TTE for the same event based on a stream of tweets referring to it, we can include the knowledge from earlier estimations in new estimations, by choosing the majority estimation over all windows. This way outlier estimations are overruled by the majority vote, ensuring a more robust system. We included this postprocessing step, referred to as 'History', as an additional experimental variant.

Evaluation

We evaluate the variants of our method by taking the number of days that estimations are off on average, and by measuring the accuracy of predicting the correct date. Estimation errors are evaluated by Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (Willmott & Matsuura, 2005). The formula for the MAE is given in equation 2.1:

$$\text{MAE} = \frac{1}{n}\sum_{n=1}^{n} f^i - e^i \tag{2.1}$$

The MAE sums for each estimate i the absolute number of days that it is off with the predicted TTE (e^i) from the actual TTE (f^i), and takes the average of all these differences.

The RMSE is calculated by the formula in equation 2.2:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(f^i - e^i)^2} \tag{2.2}$$

Window size	Step $1/5$	Step $1/2$	Step $1/1$	Av. TTE	Av. duration
10	1.14 (1.37)	1.43 (1.48)	1.83 (1.65)	-14	4 days
20	0.81 (1.03)	0.80 (0.97)	1.01 (1.34)	-11	8 days
50	0.45 (0.50)	0.55 (0.55)	0.63 (0.73)	-7	11 days

TABLE 2.2: MAE (with standard deviations) averaged over the 9 methods applied on all 60 events per window and step combination, versus actual average TTE at first estimation, and average duration of the first estimate, per window.

	Standard			History		
	Accuracy	MAE	RMSE	Accuracy	MAE	RMSE
Majority Rules	0.95	0.07	0.30	**0.97**	**0.05**	0.20
Majority Rules+TL	0.96	0.07	0.36	**0.97**	**0.05**	0.24
ML n-grams	0.74	0.79	1.96	0.85	0.34	0.78
ML Rules	0.65	2.21	3.88	0.88	0.86	1.42
ML TL	0.42	2.95	4.55	0.60	1.65	2.40
ML n-grams+Rules	0.76	1.32	3.10	0.90	0.29	0.82
ML n-grams+TL	0.73	0.98	2.21	0.83	0.59	1.28
ML Rules+TL	0.80	1.12	2.57	0.96	0.19	0.48
ML n-grams+Rules+TL	0.88	0.31	1.17	0.96	**0.05**	**0.19**

TABLE 2.3: Performance of different settings on 60 football events for a window size of 50 and a step size of 10 (Rules = Dynamic and exact Rules, TL = Timelearn, ML = Machine Learning, n-grams= Word n-grams).

The squared differences between the actual TTE (f) and the estimated TTE (e) are summed, and finally the square root of this sum is taken to produce the RMSE of the sequence of predictions.

MAE can be interpreted as the average number of days a method is off. RMSE penalises large errors more heavily. The overall MAE and RMSE for the 60 events is calculated as the average of all event MAE and RMSE scores respectively.

In addition to estimation errors, we calculated the accuracy, which scores the proportion of exact estimates. A system that classifies many windows as 'during' and 'after', in which case no error could be calculated, might have a low MAE and RMSE, but will have a poor accuracy.

2.3.2 Results

To obtain an impression of the quality of the multiple window and step sizes, we calculated the average MAE per window and step size across the applied methods, shown in Table 2.2. Estimation errors are within one day for window sizes 20 and 50. The largest window size of 50 tweets produces the lowest errors, although the first estimation is made only a week before the start of the event

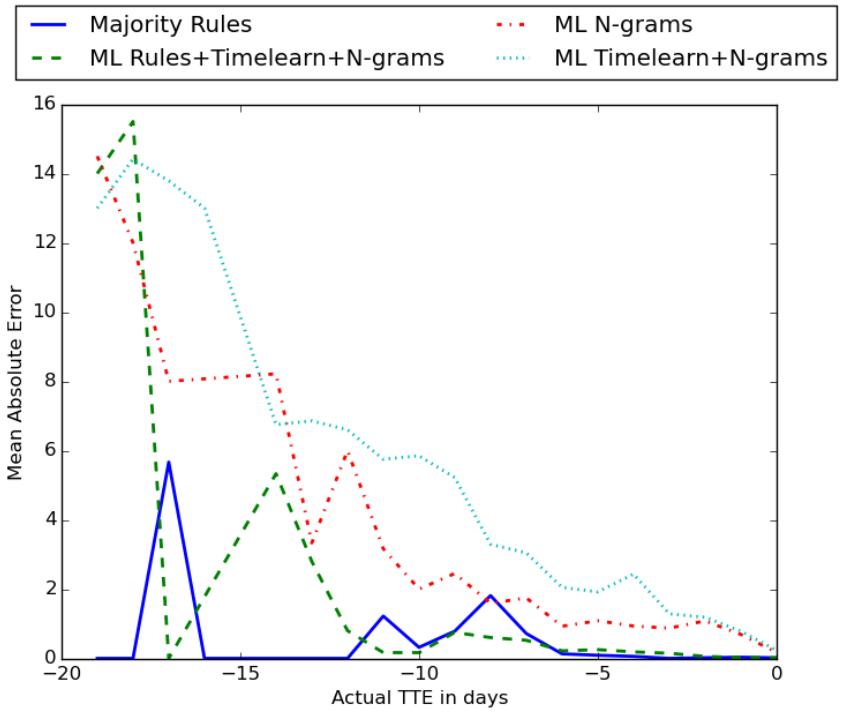


FIGURE 2.1: MAE per actual TTE in days averaged over the absolute errors for all 60 football events, for a window size of 50 and a step size of 10 (ML = Machine Learning).

rather than two weeks in case of a window size of 10. Smaller step sizes lead to better estimates.

To compare the methods, we select the window and step combination that leads to the best performance for most methods: a window size of 50 and step size of 10. The results for this combination are presented in Table 2.3. The incorporation of history knowledge when the final estimation is made, shown in the right half of the table, shows to be rather beneficial for the performance of any method. Surprisingly, SVMs are nearly always outperformed by either of the more straightforward majority voting methods, with a very high accuracy of 0.97 and a very low MAE of 0.05 in the history-sensitive variant. Apparently, the manually-set TTE estimations already provide sufficient information. The majority voting method combining the trained and rule-based TTEs offers no improvement over the method that only uses the rule-based TTEs.

The best ML method combines all three feature types, leading to a performance roughly equal to the majority voting methods. The RMSE, which is more sensitive to higher errors, is even better for this method. The two TIMEX feature types, Timelearn and Rules, are not effective by themselves when fed to the SVM classifiers, while word n-grams do lead to reliable estimations with a MAE of less than half a day.

To obtain insights into the performance related to actual TTEs, we plotted the MAE per actual TTE averaged over all 60 football events for a selection of four methods in Figure 2.1. The majority voting method obtains a flawless performance between an actual TTE of -17 and -13 days, while the ML method with all three feature types performs better between an actual TTE of -13 and -7 days. The other two methods lag behind for every TTE. A prominent error peak is seen at 14 days for the ML method with all three feature types. This might relate to the words that are used in tweets posted in the weekend, when other matches are played. It seems that these words are not much different from the words that are used a weekend before a match, causing the classifier to confuse the corresponding TTE estimations.

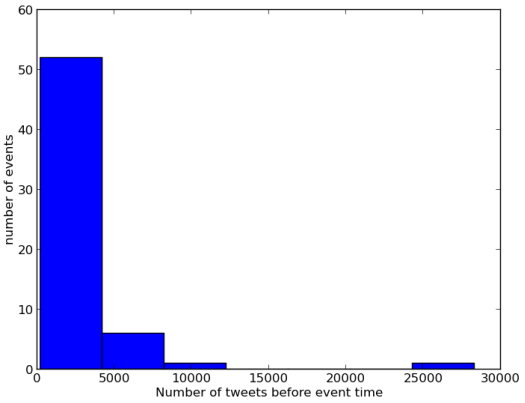
2.3.3 Error analyses

Quantitative Error Analysis

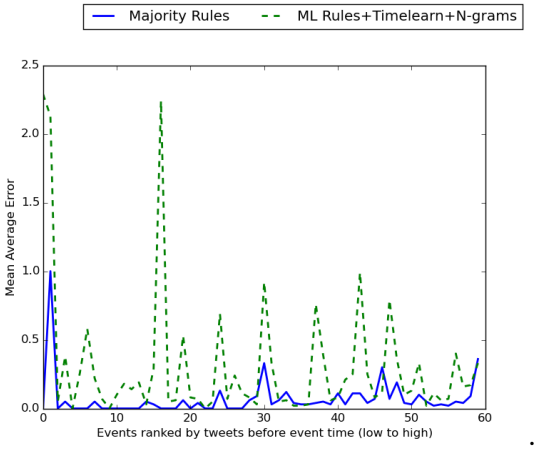
In section 2.3.2 we assessed the performance of our systems by averaging over the 60 events that were held out for testing. To acquire a sense of the performance at the event level, we visualise some characteristics of the events and the influence of these characteristics on performance in Figure 2.2: the number of tweets per event (Figure 2.2 (a) and Figure 2.2 (b)), and the proportion of correct matching rule-based time expressions per event (that match the actual date of the event, Figure 2.2 (c) and Figure 2.2 (d)). We focus on the performance of the best Majority method and the best ML method: Majority using Rules and ML using all three feature types, both using a window size of 50 and a step size of 10 tweets. The performance is scored by MAE. History information, that smoothes away some of the errors, is not included to highlight the actual errors.

Figure 2.2 (a) indicates that the bulk of the events in our set are referred to before event time in less than 5,000 tweets. One outlier event was tweeted about for over 25 thousand times. Figure 2.2 (b) displays the MAE by event size rank. The absolute number of tweets per event does not seem to influence the performance of both methods, that show peaks for both low and high ranked events. If anything, the Majority method seems to be suitable for smaller events, as it does not yield high peaks between rank 3 and 23.

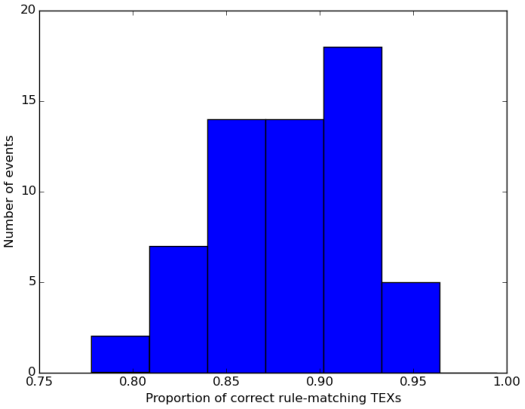
Another relevant characteristic is the proportion of correctly induced dates for all tweets per event, which ranges from 0.77 to 0.97 (Figure 2.2 (c)). The expectation that this proportion is directly correlated with the success of the Majority method is reflected in Figure 2.2 (d). When going higher up the ranking, most induced event dates are accurate and the error peaks get smaller. On the other



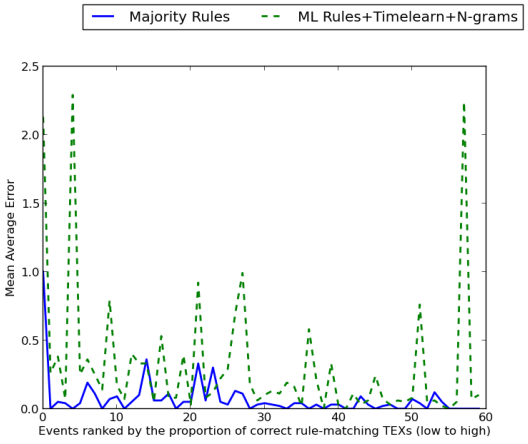
(a) Histogram of the number of tweets before event time, for all 60 football events



(b) performance by MAE on the events ranked by tweet count from low to high (ML = Machine Learning)



(c) Histogram of the proportion of correct rule-induced TIMEXs, for all 60 football events



(d) performance by MAE on the events ranked by the proportion of correct Rule-induced TIMEXs from low to high (ML = Machine Learning)

FIGURE 2.2: Overview of the characteristics of and performance on separate football events.

hand, the ML method, that makes use of Timelearn and word n-gram features in addition to the rules, shows error peaks throughout the ranking.

Qualitative Error Analysis

We inspected the contents of the events that showed the highest MAE for the most successful approach, the Majority method, to find out what caused the errors.

The Majority system most commonly made an inaccurate estimation when a number of tweets referred to a 'side event' rather than the football match itself.

Two side events were most prominent: first the action of buying a ticket for the match, and second the anticipation of another football match played by the same team(s). Below, we present some examples of references to these side events that were seen in our data. We translated the tweets, Dutch by origin, to English:

1. *Going to buy tickets for #psvfey Saturday*
2. *Going to feyenoord-kiev tuesday and to fc utrecht-feyenoord Sunday next week #feykie #utrfey #diehard*
3. *but first the Cup Final next week, still playing for the double :d #ajatwe*

The tweet in Example 1 states the intention to buy tickets on Saturday for a match that occurs on another day – this statement is ambiguous and could only be interpreted correctly with explicit knowledge of the match day. Example 2 mentions two matches by their hashtag that the user will visit, along with the two respective days at which they are played; disambiguation would require a segmentation of the two clauses in this tweet. Example 3 refers to the cup final, which is co-incidentally played between the same two teams that have an important league match one week later.

These errors mainly show the flaws that are related to clustering event tweets by a hashtag: the event hashtag is not necessarily the main topic of the tweet. In an open-domain setting, a more elaborate method is needed to cluster tweets that refer to the same event.

2.4 Other Types of Events

To test the generalisation of our method to other events, we tested all nine methods trained on the football events with their best window and step settings on five public events of different types. The best window and step setting per method is specified in Table 2.5.

2.4.1 Experimental Set-up

Data

We selected five recent events that took place in the Netherlands and could be identified based on a common hashtag or key phrase. Our test set contains two popular concerts in the Netherlands in 2013: a concert by Justin Bieber (identified by '#believetour') and a concert by Bruce Springsteen (identified by 'bruce springsteen'). Further, we included the national Queen's day celebration in 2013

event key (phrase)	# tweets	# tweets before event start time
#believetour	2,576	1,606
bruce springsteen	2,601	1,258
#koninginnedag	15,618	8,154
nationale iq test	982	51
project x haren	19,124	7,516

TABLE 2.4: The selected events and tweet counts after removing retweets.

('#koninginnedag'), the television broadcast of the national IQ test of 2013 ('nationale iq test') and a birthday celebration in 2012 of which the invitation was virally spread on social media, causing rioting and substantial damages ('project x haren'). For some of these events, the tweets were collected based on the most common hashtag, like for the football events, while for others the related tweets were collected based on string matching.

Like the football events, we collected the tweets that refer to the events by means of TwiNL. We specified a six-week search window around the known date of the events, to ensure a link with the TTE labels that ML trains for the football events.[9] We removed the retweets from the tweets that we obtained. The tweet counts of the resulting data sets are listed in Table 2.4.[10]

Approach

With the exception of the majority voting method using rule-based estimations (which does not require training), the methods were trained on the 60 football events and tested on the five test events. For each method we selected the optimal window and step size in terms of MAE as found during the experimentation with the football events, adopting the variant in which history estimations are included in the choice for each estimation, which proved to improve performance of all methods.

2.4.2 Results

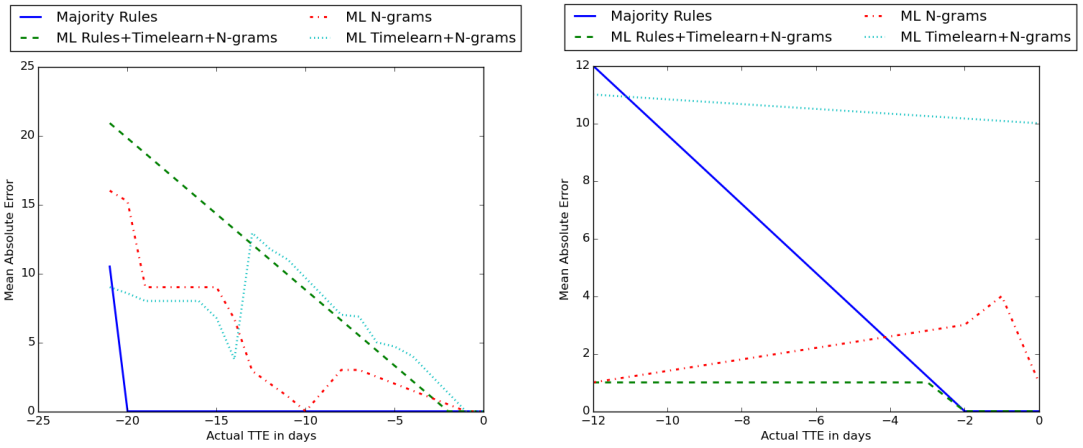
Table 2.5 lists the performance of the methods in terms of MAE. The results show that methods based on the rule-based features obtain the best performances, with a flawless performance on all events by the majority vote on rules. The ML methods are especially troubled by the 'nationale iq test'. Overall, most methods

[9] Realistically, future references to an event should be recognised any period of time ahead of the event, but three weeks before event time would capture most of the tweets for a lot of events.

[10] The tweet IDs for these events and their TTE label are available from `http://dx.doi.org/10.17026/dans-z8e-3uqb`.

	window and step	#believe-tour	bruce springsteen	#koningin-nedag	project x haren	nationale iq test	Mean
Rules	50 - 20	0.00	0.00	0.00	0.00	0.00	0.00
Rules+TL	50 - 10	0.00	0.16	0.08	0.00	0.00	0.05
ML n-grams	50 - 10	0.87	0.45	0.08	0.00	1.00	0.96
ML Rules	20 - 20	0.00	0.00	0.03	0.00	1.00	0.21
ML TL	50 - 10	1.26	1.64	1.04	0.13	4.00	1.61
ML n-grams+Rules	50 - 50	1.16	0.40	0.48	3.21	1.00	1.25
ML n-grams+TL	50 - 20	0.39	0.00	5.25	0.03	2.00	1.53
ML Rules+TL	20 - 20	0.40	0.00	0.19	0.00	1.50	0.42
ML n-grams+Rules+TL	50 - 10	0.11	0.04	0.52	0.06	4.00	0.95

TABLE 2.5: MAE for the TTE identification of open domain events after training on 60 football events.



(a) #koninginnedag

(b) nationale iq test

FIGURE 2.3: MAE per actual TTE in days, for a window size of 20 and the optimal step size per method (as scored during training and testing on the 60 football events).

are quite accurate in their estimations; also the ML based method trained on all features produces estimates with one day of error.

To obtain insights into the influence of event size, we plotted the MAE per actual TTE for the largest and smallest of the five events in terms of tweets posted before the event start time: '#koninginnedag' (Figure 2.3 (a)) and 'nationale iq test' (Figure 2.3 (b)). As a window of 50 would only allow one estimation for the latter event, we applied four of the methods with a window size of 20 and their best step size as estimated on the football events.

Interestingly, the plots show changing performances of the Majority method and the optimal ML method. Where the former generates flawless estimations from 20 days before '#koninginnedag' on to the beginning of the event, it is substantially off in its estimation 12 days before the 'nationale iq test'. In contrast, the optimal ML method generates poor estimations for almost all windows

when processing tweets referring to #koninginnedag, while its early estimation for the national IQ test is only one day off. The latter is an example where ML including word n-grams shows to improve TTE estimation in comparison to a more straightforward majority voting method based on TIMEXs.

A closer investigation into the 51 tweets that were posted before the National IQ test shows that only some of the earlier tweets contained a TIMEX. Furthermore, about half of them were not related to the event itself, confusing the Majority method in its estimation. We can conclude that our methods can cope with smaller events, but such events do not provide signals as stable as more popular events.

2.5 Conclusion and Discussion

In this chapter we experimented with different approaches to estimate the number of days until the start of an event mentioned on Twitter. We compared machine learning and majority voting approaches using different feature combinations and window and step sizes. The methods were first applied to 60 football events, and then to five different types of events with their optimal window and step size. The results demonstrate that a machine learning approach based on all types of features obtains an accurate performance equal to the majority voting approach based on knowledge-based time expression estimations, when trained and tested within a set of football events. However, the knowledge-based method obtains the most robust performance throughout different types of events. A qualitative error analysis indicated that time references in tweets sometimes refer to side events, so it is important to complement time features with other features in case of conflicting clues.

The results also show that a larger window generally leads to a better performance, which can be attributed to the larger amount of information in such windows. The question remains, however, whether such large windows are suitable when early TTE estimations are preferred. In future work, we aim to incorporate a variable window size that is obtained during training, in which both the MAE and early TTE estimation are optimised.

Although the approach based on majority voting over TIMEX estimations is often flawless in estimating the number of days to the event, for some points in time (between 13 and 7 days before football event start times), we found that machine learning based on all features was more accurate on average. This shows that different methods might complement each other in their estimation.

In sum, starting from a set of event-related tweets, with this study we have shown that the time until the start of an event can be deduced rather accurately from anticipating event tweets. However, an application that identifies the date at which events take place can not rely on the manual input of event terms, for which the start time is likely known. In Chapter 3 and 4, therefore, we describe two studies that aim to identify such events and their date automatically from an open set of tweets. Explicit references to a future point in time, which have proven useful in the current study, will form a central component in the study described in Chapter 4. Chapter 3 leverages a more direct signal: the time at which tweets are posted.

Chapter 3

Event detection in Twitter: A machine learning approach based on term pivoting

Based on: Kunneman, F. & van den Bosch, A. (2014). Event detection in Twitter: A machine-learning approach based on term pivoting. In F. Grootjen, M. Otworowska, & J. Kwisthout (Eds.), *Proceedings of the 26th Benelux Conference on Artificial Intelligence* (pp. 65–72). Nijmegen, The Netherlands

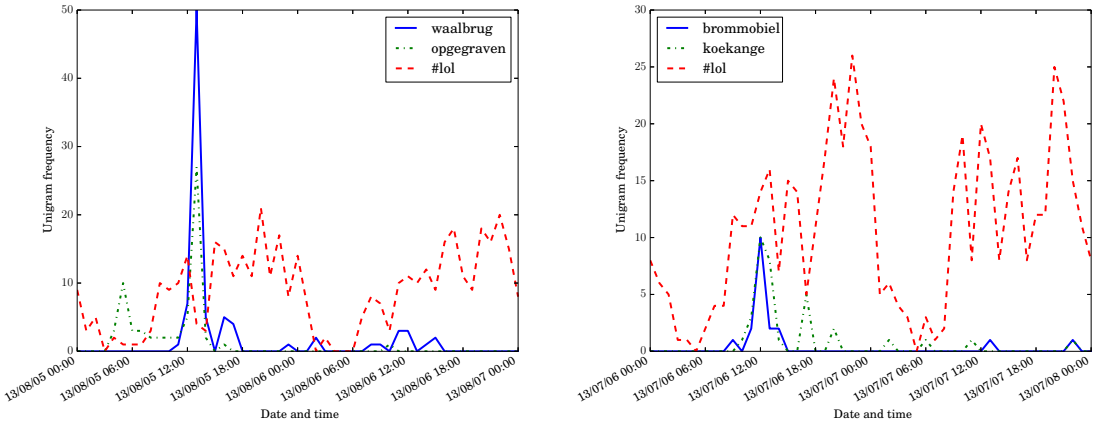
The large number of messages posted on Twitter each day provide rich insights into real-world events and public opinion. However, it is difficult to automatically distinguish tweets referring to such events from everyday chatter, and subsequently to distinguish significant events affecting many people from insignificant events. In this chapter, we apply a term-pivot approach to detect events from the Twitter stream as bursty terms. In order to filter out noisy and mundane events, we train a machine learning classifier on several rich features, and rank the events based on classifier confidence. After training and re-training the classifier using manually annotated data, we obtain an $F_{\beta=1}$ score of 0.79. However, a baseline that only takes into account the frequency of the tweets that refer to an event yields a better $F_{\beta=1}$ score of 0.86.

3.1 Introduction

Microblogging platforms such as Twitter give users a voice to share ideas, opinions, and experiences with friends and the general public. Owing to the large user base on Twitter, the platform provides real-time information about what happens in the world. Detecting events and harvesting references to them from Twitter is therefore a highly valuable goal. However, this task is hampered by the nature and dynamics of Twitter. While news media select newsworthy items to write about, there is no such top-down selection process in the Twitter ecosystem. Events of public interest and mundane, insignificant events may both be characterised by bursty peaks in the usage of a set of terms in Twitter.

As an illustration of term burstiness in Twitter, consider the two examples in Figure 3.1. Example (a) displays the event of an excavation near the bridge 'Waalbrug' in Nijmegen, represented by a single joint rise and fall in the usage of the words 'waalbrug' and 'opgegraven' (Dutch for 'excavated') in Twitter. As a comparison, we plot the frequency of the commonly used hashtag '#lol' in the same time window, which does not show any burstiness. It could be hypothesised that the first two terms both refer to an event, and possibly to the same event. Example (b) shows a similar pattern for the terms 'brommobiel' and 'koekange', peaking at about the same point in time, contrasted again with the non-bursty hashtag '#lol'. Without any additional knowledge, a system that leverages term burstiness might label the joint peaks in both examples as an event. However, further inspection shows that the peaks in example (b) denote a news report about a criminal act in the place of Koekange and an unrelated traffic accident with a scooter. A proper event detection system needs to filter out such insignificant events, possibly by taking into account additional features beyond burstiness.

The aim of this study is to expand existing work on detecting significant events on Twitter. As a definition of what makes an event significant, we use the definition that we described in Chapter 1: 'events that might be reported in the news media and that are hence of interest to a large group of people'. We build on the approach proposed by Qin et al. (2013). They implement the *Twevent* approach to event detection in Twitter (C. Li, Sun, & Datta, 2012), and expand it by training a classifier on several features of an event to recognise significant events in contrast to mundane, insignificant events. We reproduce their experimentation and apply it to two months of Dutch tweets.



(a) (b)

FIGURE 3.1: Illustration of bursty and non-bursty term occurrences. Left: 'waalbrug' and 'opgegraven' (bursty) and '#lol' (non-bursty); right: 'brommobiel' and 'koekange' (bursty) and '#lol' (non-bursty).

3.2 Related Work

The detection of events in Twitter has been the goal of many studies. It is mainly approached as a clustering problem, with burstiness as the most important characteristic to detect an event. The most salient dichotomy among approaches is what Fung, Yu, Yu, and Lu (2005) call *document-pivot clustering* and *term-pivot clustering*: burstiness is either measured at the level of documents that share common terms, or at the level of single terms that display a joint burstiness over time. We provide an overview of the most important event detection systems, and summarise the performance on retrieving significant events reported by these studies.

3.2.1 Document-pivot clustering

The clustering of documents for the detection of events originates from the Topic Detection and Tracking (TDT) area of research (Allan, Papka, & Lavrenko, 1998). Given a stream of news messages, any incoming message is linked to an existing event cluster or is the start of a new event cluster. Petrović, Osborne, and Lavrenko (2010) propose an adaptation of this approach to fast text streams such as Twitter. Incoming messages are either linked to an existing cluster, or grouped into a new one dependent on the distance to their nearest neighbour. Events are distinguished from other clusters based on the growth rate of a cluster. Petrović et al. (2010) obtain an average precision of 0.34 of retrieved event

tweets versus tweets that are not related to an event, or spam. McMinn et al. (2013) reproduce the approach of Petrović et al. (2010), resulting in the retrieval of 1,340 events in 28 days of tweets, of which 382 (28%) are found to be significant.

Instead of clustering incoming tweets based on their raw content, alternative approaches focus on specific aspects of tweets that refer to future events. Ritter et al. (2012) state that important events on Twitter, in comparison to mundane events, have a common point in time to which multiple tweets refer explicitly. They extract events by clustering tweets that refer to the same point in time and mention the same entity. When ranking events based on the strength of the association between their date and entity, Ritter et al. (2012) obtained a precision-at-100 (precision within the top 100 events) of 0.90 and a precision-at-1,000 of 0.52.

Yet another way to cluster tweets into events is to apply Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), by which individual words are linked to a topic based on their co-occurrence with other words. While LDA originally does not incorporate time as a feature, different extensions of the approach have been developed to obtain smoother topics over time (Blei & Lafferty, 2006; X. Wang & McCallum, 2006). The short nature of tweets poses a problem for the application of LDA to Twitter. This problem can be handled by aggregating tweets to a bigger document based on a common characteristic (Mehrotra, Sanner, Buntine, & Xie, 2013) or by assigning only one topic to each tweet and including a background model for noisy or general words (W. X. Zhao et al., 2011). To detect bursty topics in Twitter, Diao, Jiang, Zhu, and Lim (2012) build on the Twitter-tuned LDA implementation by W. X. Zhao et al. (2011), and expand it by adding topic distributions per time window and per user. Bursty topics are typically detected as a set of tweets from different users that contain similar words within a time window. A disadvantage of LDA is its dependence on a predefined number of topics. In addition, running LDA on a large volume of tweets is computationally expensive, due to the high number of sampling iterations that are required. Diao et al. (2012) set the number of topics to 30 in a period of 91 days, and obtained a precision of 0.76 for these topics (a precision-at-5 of 1.00).

3.2.2 Term-pivot clustering

Fung et al. (2005) propose term-pivot (or feature-pivot) clustering as an alternative to document-pivot clustering for event detection from a news stream. Its two main advantages are the independence from parameter settings, and the event

summary that is readily given by clustered terms. The first effective application of term-pivot clustering to event detection on Twitter is proposed by Weng and Lee (2011), who capture the burstiness of words by considering them as signals and applying wavelet analysis. Signals are clustered based on modularity-based graph partitioning on their cross-correlation scores. Clusters with higher cross-correlation scores between the representing signals are seen as more significant. As a cluster with a large number of signals is not likely to represent a coherent event, Weng and Lee (2011) penalised the significance of a cluster by the number of signals. They obtained a precision of 0.76 for 21 events retrieved in a month of tweets from a Singapore user base.

C. Li et al. (2012) argue that multi-word segments or word n-grams, rather than single words, are beneficial both for the interpretation of an event and the detection of significant events. At the core of their *Twevent* system is the extraction of meaningful n-grams from tweets. The most meaningful non-overlapping n-grams from a tweet are selected based on their prior probability as calculated from the Microsoft Web N-Gram service and their likelihood to be used as an anchor text in Wikipedia. N-grams are scored by their burstiness, and bursty n-grams are clustered into candidate events. The significance of a candidate event is dependent on the newsworthiness of the individual n-grams, formulated as the combined chance of any n-gram sub-phrase to occur as an anchor text in Wikipedia, and the mutual similarity scores between the n-grams. C. Li et al. (2012) obtained a precision of 0.86 for 101 detected events on the same dataset as Weng and Lee (2011).

For the works discussed above, event significance is scored by an intuitive measure, such as the number of cluster terms (Weng & Lee, 2011) or the growth rate of a cluster (Petrović et al., 2010). Aiming to improve over these simple estimations of event significance, Qin et al. (2013) apply *Twevent* to 15 days of English tweets and annotated the 4,249 resulting clusters as 'True news event' or 'False news event'. The clusters are linked to 15 rich features presumed to be indicative of their significance (a selection of these features is described in more details in Section 3.3.3). A classifier is trained and tested through 10-fold cross validation on all event clusters, resulting in a precision of 0.84 on 146 retrieved events, compared to 0.76 on 107 events by the original *Twevent* system.

In the study described here we adopt the approach by Qin et al. (2013). Where they build on the framework of *Twevent* to form clusters of segments, we base this clustering on unigrams rather than on segments. As a proxy to identify significant events, we borrow the idea of C. Li et al. (2012) to include

the presence of a certain name or concept as an article on Wikipedia as a weight in determining the significance of the candidate cluster of terms.

3.3 Experimental Set-up

3.3.1 Data

We collected the available tweets from June 22nd 2013 until August 22nd 2013 from TwiNL, and filtered out non-Dutch tweets according to the language identification that it offers, resulting in a set of 65.02 million tweets.

3.3.2 Event detection

Our event detection approach takes the following steps.

Unigram selection by burstiness

To select candidate unigrams we tokenised the tweets with ucto,[1], removed punctuation and user names, and lowercased the remaining words. Additionally, we removed stop words from each tweet based on the list of Dutch stopwords from Snowball.[2] For each unigram we generated a time sequence of the tweets that contain the unigram. Following C. Li et al. (2012) we set the window size for this sequence to 24 hours, focusing on events that occur within a day.[3]

Given a day-by-day sequence of counts for a unigram, we score its burstiness per day by applying the state automaton approach to burstiness detection (Kleinberg, 2003). Each day a unigram can take on a bursty or normal state. The most likely sequence of states for a unigram can be uncovered by applying a Hidden Markov Model on the observed probability at each stage and the transition probability from state to state. We base the modelling of these two probabilities on the implementation by Diao et al. (2012). The observed probability of a count is based on a Poisson distribution for each state, which is defined as follows:

$$p(f_{ut} \mid v_t = l) = \frac{e^{-\mu_l} \mu_l^{f_{ut}}}{f_{ut}!} \tag{3.1}$$

Where f_{ut} is the frequency f of unigram u for time window t, l is either 0 or 1, and the normal and bursty states are denoted by μ_0 and μ_1, respectively. Following Diao et al. (2012), we set μ_0 to the average count of a unigram over

[1] `https://languagemachines.github.io/ucto/`

[2] `http://snowballstem.org/algorithms/dutch/stop.txt`

[3] An overview of the sequence of counts by unigram can be accessed from `http://dx.doi.org/10.17026/dans-zk4-aq5x`.

time and we set $\mu_1 = 3\mu_0$, i.e. an observed frequency has a higher probability to represent a bursty state when it approximates three times the average count. Also following Diao et al. (2012) we set the transition probability σ_0 to 0.9 and σ_1 to 0.6, implying that a transition from a normal state to a bursty state is not very likely with a chance of 0.1. The chance that a bursty state reverts to a normal state is higher, with 0.4.

We use the Viterbi algorithm (Forney & David, 1973) to dynamically find the bursty states for each unigram, and discard the unigrams without a bursty state as candidates. In our data set of 62 days, the method identifies $253,472$ bursty unigrams, with an average of $4,088$ per day ($\sigma = 703$).

Unigram similarity

To cluster unigrams into event clusters, we adopt the approach by C. Li et al. (2012). For each day in our dataset, the similarity between all pairs of bursty unigrams is calculated and clusters are formed based on this similarity graph. To calculate the similarity, each time window t is divided into M sub-time-windows. Following C. Li et al. (2012) we set the size of M to 12 (i.e. two hours per sub-time-window). The similarity between any pair of unigrams u_a and u_b on a day is calculated as follows:

$$sim(u_a, u_b) = \sum_{m=1}^{M} w_t(u_a, m) w_t(u_b, m) sim(T_t(u_a, m), T_t(u_b, m)) \tag{3.2}$$

The sub-time-window similarity between unigrams is computed by collecting the tweets in which the unigrams are mentioned, and generating two pseudo-documents containing all concatenated tweets in which one or the other unigram occurs. Terms in these documents are weighted by $tf * idf$ (Day & Edelsbrunner, 1984), and the cosine similarity (Steinbach, Karypis, & Kumar, 2000) between the two pseudo documents is calculated. The unigram similarity calculation favours pairs of unigrams that are mentioned with comparable content and that are most bursty in the same sub-window. Furthermore, it considers the similarity between tweets rather than the co-occurrence of unigrams, which is reasonable given the shortness of tweets.

Term clustering

Given the similarity graphs of bursty unigrams per day that result from the previous step, unigrams are clustered into event clusters. Following C. Li et al. (2012), we apply Jarvis-Patrick clustering (Jarvis & Patrick, 1973). This algorithm

has two parameters, k and l. For any two unigrams to be clustered together, they have to occur in each others k-nearest neighbours and they have to share at least l common neighbours in their k-nearest neighbours. This algorithm does not require a specification of the number of clusters to make, and has a limited computational cost.

C. Li et al. (2012) found that the l parameter is too restrictive for this task. Following them, we only took into account the k parameter and set $k = 3$, linking unigrams if they occur in each others top 3 most similar unigrams. Unigrams that were not linked to any other unigram were discarded. As a result, we retrieved a total of $31,758$ event clusters from the 61 days of bursty unigrams (512 on average per day).

3.3.3 Event significance classification

Event significance classification can be seen as what Qin et al. (2013) call 'event filtering'. The events that result from clustering are sorted into significant and insignificant events. We apply the same approach to event filtering as Qin et al. (2013): describing event clusters by rich features and training a classifier to distinguish significant from insignificant events.

Features

In their research, Qin et al. (2013) include 15 rich features. Most of the features that we include are adopted from Qin et al. (2013). We describe the features below, and make a distinction between cluster features and tweet features: respectively the characteristics of the unigrams that describe a cluster and the characteristics of the tweets in which the unigrams of a cluster occur.

Cluster features

- Unigrams - The number of unigrams in the event cluster. Arguably, a cluster which is described by many unigrams is not likely to represent a coherent, significant event.
- Edges - The average number of clustering edges between the unigrams in the event cluster. This feature describes the density of a cluster.
- Similarity - The average similarity score, as described in section 3.3.2, between unigrams in the event cluster. A higher score might point to a more coherent event cluster.

- Burstiness - The average burstiness of unigrams in the event cluster. A higher burstiness might point to a higher importance. While we initially identified bursty unigrams by a binary metric, we scored this burstiness feature by means of a continuous scale to make a more precise distinction between events. As metric, we adopted the bursty probability calculation in C. Li et al. (2012). This probability is based on the expected frequency $E[u|t]$ of a unigram u in a time window t, given its Gaussian distribution:

$$E[s|t] = N_t P_s = N_t * \frac{1}{l} \sum_{t=1}^{L} \frac{f_{u,t}}{N_t} \tag{3.3}$$

  Here, N_t is the number of tweets during day t, L is the number of time windows containing u, and $f_{u,t}$ is the frequency of u in time window t. Given $E[s|t]$, the bursty probability $P_b(s,t)$ is calculated as follows:

$$P_b(s,t) = S(10 * \frac{f_{s,t} - (E[s|t] + \sigma[s|t])}{\sigma[s|t]}) \tag{3.4}$$

  S is the sigmoid function and $\sigma[s|t] = \sqrt{N_t P_s (1 - P_s}$, the standard deviation of the Gaussian distribution.

- Newsworthiness - The average newsworthiness of unigrams. Mundane events likely have a low degree of newsworthiness. This feature is operationalised by C. Li et al. (2012) as the ratio by which terms that are (in) the title of a Wikipedia page are referred to from other pages from anchored links. Terms that have a high probability to be used as anchor to their page are believed to be more newsworthy. To calculate the newsworthiness score for all bursty terms, we downloaded a dump of the Dutch Wikipedia pages from November 14th 2013 (the closest date after the latest tweet in our data set).[4]

Tweet features

- Document Frequency - The relative frequency of the event tweets, calculated as the number of event tweets on the day of the event divided by the total number of tweets on that day. This metric might highlight the popularity of an event.
- User Document Frequency - The relative number of different users that refer to the event, calculated as the number of users that posted one of the

[4] http://dumps.wikimedia.org/nlwiki/20131114/

event tweets, divided by the total number of event tweets. A high document frequency is arguably less significant of only a few users have posted them.

- Cohesiveness - The average number of bursty unigrams in tweets. If the event tweets contain two or more of these unigrams, they are more likely to refer to a cohesive event.
- Informativeness - The relative number of different words in the event tweets. Spam messages are often characterised by a narrow vocabulary, while events that arouse the attention of a lot of people might be referred to with a bigger variation of words.
- Hashtags - The average number of hashtags per event tweet. These final four feature types point to characteristics of Twitter posts.
- URLs - The percentage of event tweets that contain a URL (any token starting with ('http(s)://')
- Replies - The percentage of event tweets that start with a user name (tokens that start with a '@'), which is typical of a reply.
- Mentions - The percentage of event tweets that contain a mention of a user name, on any position other than the start of a tweet.

Classification

While Qin et al. (2013) annotate all 4,249 event clusters retrieved by the *Twevent* approach from their data set, this was not feasible for the 31,758 event clusters that we retrieved. Instead, we selected a subset of the data. To ensure enough significant events in this subset, we trained a classifier on 350 labeled event clusters on the first two days in our data set and applied it to the remaining days. The 1,000 events of which the classifier was most confident were used as data set for our experimentation.

As classifier we made use of the SNoW implementation of Winnow (Carlson, Cumby, Rosen, & Roth, 1999).[5] This algorithm is known to offer state-of-the-art results in text classification, and outputs a per-class confidence score by which instances could be ranked. To tune the different parameters of Winnow (α, β, $\theta+$, $\theta-$, the number of iterations and the thick separator), we applied a heuristic hyperparameter optimisation scheme that makes use of wrapped progressive sampling on training data (van den Bosch, 2004).

[5] `http://cogcomp.cs.illinois.edu/page/software_view/SNoW`

To obtain the initial event clusters to train the classifier, we selected 350 event clusters from the first two days. We focused on events with a likely high cluster quality, by ranking them based on the average similarity score of their unigrams and selecting the clusters with the most similar scores. One of the authors annotated the top 350 of these events as significant or not, resulting in 153 events labeled as significant and 197 events labeled as insignificant. The classifier was trained on all 13 features in these 350 labeled events and was applied to the total of events in the remaining days in the data set. The 1,000 events that were most confidently scored as significant by the classifier were used in our main experimentation.

To obtain trustworthy labels for the 1,000 events we asked eight annotators to each label 250 events as significant or not significant. The data was split in a way that each event was annotated by two annotators, with eight unique annotator pairs (125 events per pair). We presented them with a list of events represented by a date, the event unigrams, and a sample of ten of the event tweets. In our explanation of the task, we gave them the definition of a significant event that we specified in the introduction of this chapter, as well as a few examples of typically significant and insignificant event clusters.[6]. The task was to annotate each event as either significant, insignificant, or doubtful. We included the latter category to gain insight into the difficulty of the task. For events that were judged as significant or doubtful, we additionally asked the annotators to indicate if the event was a social event.[7]

354 of the 1,000 event clusters were indicated by both annotators as significant, 723 were annotated as significant by at least one of the two annotators and 277 events were annotated by both annotators as either insignificant or doubtful. Of those 277 events, 164 were annotated by both annotators as insignificant and 35 by both as doubtful. 156 events were deemed doubtful by at least one annotator. The average inter-annotator agreement was fair (Landis & Koch, 1977) at $\kappa = 0.25$ with a standard deviation of 0.11.

3.3.4 Evaluation

Based on the 1,000 annotated event clusters, we evaluated classification performance by 10-fold cross validation. We applied classification with a strict and lax labelling. For strict labeling, only events that were indicated as significant by two annotators were labeled as significant, while for the lax labeling, events that were annotated by one or two as significant sufficed to carry this label. To

[6]The task description, translated from Dutch, is included in Appendix A.

[7]When conducting this study, we planned to use these annotations for additional research.

	Strict			Lax		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
DF	0.80	0.95	**0.86**	0.73	0.99	0.84
SIM	0.54	0.93	0.68	0.84	0.95	0.89
BST	0.57	0.84	0.68	0.93	0.94	**0.94**
All	0.76	0.90	0.82	0.91	0.93	0.92

TABLE 3.1: Results for significance classification of events in a strict and lax setting after 10-fold cross validation, by performing classification based on a single feature (DF, SIM or BST) and based on all 13 features.

score the performance, we calculated the precision, recall and $F_{\beta=1}$ scores for the retrieval of significant events. As baselines we ran the classifier separately on the intuitively most effective features for significant event classification: burstiness (BST), the number of tweets mentioning the event (DF), and the similarity between unigrams (SIM).

3.4 Results

The results are given in Table 3.1. Both in the strict and the lax setting the classifier that bases its judgements on all feature values is outperformed by one of the classifiers based on a single feature. In the strict setting, the relative document frequency yields the optimal performance, while for the lax setting the term burstiness leads to the highest $F_{\beta=1}$ score of 0.94. All classifiers score a higher recall than precision, indicating a bias towards the significant event category. The performance in the lax setting is in most cases markedly higher than in the strict setting, helped by a positive balance of significant events (723 versus 277). Only the classifier based on the DF feature drops in performance, due to a considerable overprediction of significant events.

To illustrate the output of burstiness-based event detection, we show the ten events that were most confidently classified as significant in the strict setting by the classifier that uses all features in Table 3.2. The top ranked events predominantly follow the news media, with nine of them related to a news report. Four bursty clusters are seen as a significant event by both annotators, while three events are seen as significant by none. A clear example of such a nonsignificant is the output at rank 2, which represents an advertisement. The high rank seems to be caused by the high volume of tweets that mention the advertisement. Most of the events comprise a report of an event after it took place. Only the cluster of bursty tweets on rank 8 represents an event that is speculated to happen, namely a transfer of football player Didier Drogba from Galatasaray

Event rank	Event terms	Event tweet (translated from Dutch)	Judged as significant (by % of annotators)
1	bestseller, rosamund, lupton	Book R. Aslan is bestseller after outrageous interview FoxNews URL	0%
2	lingerie, boxer	RT USER Ladies Exclusive lingerie for only 3,90 URL #actionweek #18 URL	0%
3	ruta, meilutyte, lithuanian, favorite	Lithuanian Meilutyte lives up to expectations URL #swimming #swimmingpool #swimmingclass	100%
4	bogers, opening	Bogers: apart from opening satisfied with my team URL #zuidholland #region #headlines	50%
5	slotervaartziekenhuis, enrichment	#amsterdam Former CEO Slotervaartziekenhuis accused of enrichment. Aysel Erbudak, the former CEO of URL	50%
6	nieuwmarkt, pleinschout	Once again pleinschout opposed to annoyance Nieuwmarkt URL #haarlem #noordholland #region	50%
7	maryland, bitten off	Highway closed because of ear bitten off URL	100%
8	drogba, didier	RT USER Jos Mourinho wants to bring Didier Drogba back to Chelsea. The transfer seems only a matter of time URL	100%
9	canaldigitaal, transfer, muntendam	UPC and Fox make last-minute deal for transfer of channels URL #mediaan #fd	100%
10	beverages, alcoholic, curvers	Succesful Argos is getting used to alcoholic beverages. The alcoholic beverages during dinner time are starting to URL	0%

TABLE 3.2: Top 10 events most confidently classified as significant, based on all 13 features in the strict setting after 10-fold cross validation.

to Chelsea. A reason that many bursty clusters relate to an event in the future or the past is that part of the bursty terms reflect tweets that follow an original news message. Indeed, the average informativeness (e.g. type–token ratio) of all 31,758 bursty events is 0.39 ($\sigma = 0.20$), which indicates that many events are reflected by highly similar tweets. In addition, we found that for 38% of the events over half of the tweets carry a URL. Another factor of influence is the inclusion of retweets in the data set, based on which the fast replication of messages likely became a prominent characteristic of bursty terms.

3.5 Conclusion and Discussion

We reproduced the term-pivot approach to event detection proposed by C. Li et al. (2012) and applied it to two months of Dutch tweets. In line with Qin et al. (2013) we annotated the resulting events on their significance and trained a machine learning classifier based on 13 features. We found that the relative frequency by which an event is mentioned provides a sufficient cue to recognise significant events as opposed to feeding the classifier all 13 features, yielding $F_{\beta=1}$ scores of 0.86 and 0.82, respectively.

Our system obtains precision values that are similar to the ones reported by Qin et al. (2013) (around 0.80), while our recall values are considerably higher. An explanation is that Qin et al. (2013) train and test on a much larger set of 4,249 events with a smaller fraction of significant events, making the task more challenging. Furthermore, we train and test on the already ranked output of our classifier.

In our experiment, two annotators labeled each event. The fair agreement value ($\kappa = 0.25$) shows that it is difficult even for humans to decide whether an event is significant or not. We found that it is not trivial to provide the annotators with an unambiguous definition of what makes a significant event. Another factor that might have hampered the annotation is that we indicated in the annotation guidelines that a vague, ambiguous or in any case doubtful event could be annotated with a rest category. We found that this actually made the task more confusing for the annotators. Specifically, the terms or the tweets by which each event was represented could *both* be vague and ambiguous. Annotation would probably be easier if both cases were linked to a separate annotation category. Based on these findings, we altered the annotation task for evaluating the output of the event detection that is described in the next study, in Chapter 4.

Inspecting the type of events that our system retrieved, we noticed that the system seems more sensitive to news messages that are being reposted, than to people who report an event that they experienced themselves. The next study is aimed at the detection of social events, which are referenced in a more diverse manner. We set out to detect such events by consulting future references to the time at which they take place in tweets.

Chapter 4

Open-domain extraction of future events from Twitter

Based on: Kunneman, F. & van den Bosch, A. (2016). Open-domain extraction of future events from Twitter. *Natural Language Engineering*, doi:10.1017/S1351324916000036.

In this chapter we describe a system that detects future events from the Twitter stream based on references to a future date. Because its main components are the extraction of time references and entities, we will refer to this procedure as 'event extraction' throughout this chapter. The approach consists of extracting future time expressions and entity mentions from tweets, clustering tweets together that overlap in these mentions above certain thresholds, and summarising these clusters into event descriptions that can be presented to users of the system. Terms for the event description are selected in an unsupervised fashion. We evaluated the system on a month of Dutch tweets, by showing the top 250 ranked events found in this month to human annotators. 80% of the candidate events were indeed assessed as being an event by at least three out of four human annotators, while all four annotators regarded 63% as a real event. An added component to complement event descriptions with additional terms was not assessed better than the original system, due to the occasional addition of redundant terms. Comparing the found events to gold-standard events from maintained calendars on the Web mentioned in at least five tweets, the system yields a recall-at-250 of 0.20 and a recall based on all retrieved events of 0.40.

4.1 Introduction

A significant part of the messages posted on the social media platform of Twitter relate to future events. A system that can extract this information from Twitter and present an overview of upcoming popular events, such as sports matches, national holidays, and public demonstrations, is of potentially high value. This functionality may not only be relevant for people interested in attending an event or learning about an event; it may also be relevant in situations requiring decision support to activate others to handle upcoming events, possibly with a commercial, safety, or security goal. As an example of the latter category, *Project X Haren*,[1] a violent riot on September 21, 2012, in Haren, the Netherlands, organised through social media, was abundantly announced on social media, with specific mentions of the date and place. A national advisory committee, installed after the event, was asked to make recommendations to handle similar future events. The committee stressed that decision-support alerting systems on social media need to be developed, 'where the focus should be on the detection of collective patterns that are remarkable and may require action' (M. J. Cohen, van den Brink, Adang, van Dijk, & Boeschoten, 2013, p. 31, our translation). We describe a system that provides a real-time overview of open-domain future events of potential interest to any audience, by leveraging explicit references to the start time of upcoming events. As stated in Chapter 1, the type of event that we focus on is typically not of a personal nature.

Many Twitter users choose and like to share their anticipations, as can be inferred from the frequent occurrence of the Dutch hashtag '#zinin' ('#lookingforwardtoit') or of the term 'vanavond' ('tonight'). Queries for tweets with these terms yield, respectively, 677,156 tweets in 2011 and 2012 (see Chapter 7) and about seven million tweets between August 2010 and Spring 2012 (Weerkamp & de Rijke, 2012). Given an estimated average of four million Dutch tweets per day, the two terms comprise about 0.02% and 0.29% of all Dutch tweets in their respective periods. Thus, a system based on anticipatory references to future events starts with a wide selection. However, among this load of future event references are a lot of events that are not of public importance, such as a person's holiday leave or a family visit.

The challenge is then to distinguish events of public interest from personal events. We adopt the approach by Ritter et al. (2012), who look for the co-occurrence between the key descriptive entities of an event and an explicitly mentioned date of the event. Often, public events are referred to by different

[1] http://en.wikipedia.org/wiki/Project_X_Haren

persons in combination with the same time reference. Ritter et al. (2012) show that ranking events based on this evidence indeed results in a large majority of socially anticipated events in the top rankings. The tacit assumption here is that event significance is at least partly based on the number of people that post about the event. This assumption rules out the possibility of detecting significant events about which only few people post on Twitter.

An adoption of the approach by Ritter et al. (2012), the current work offers the following contributions:

- A downside of the approach by Ritter et al. (2012) is that it requires natural language engineering tools that can cope with the non-standard language use in tweets. Entities are extracted by means of the named entity tagger tailored to English tweets as described by Ritter, Clark, and Etzioni (2011), and event phrases are identified by training a classifier on annotated English tweets. Applying the approach to a different language would require substantial adaptation, such as the annotation of a sufficient amount of tweets. We propose an adaptation of the system that operates largely in an unsupervised fashion and can easily be adopted to different languages. Specifically, an approach that leverages Wikipedia is applied to select entities, and a $tf * idf$-based approach replaces the extraction of event phrases to enrich the event description.

- We extend the approach with a clustering stage to decrease duplicate output and a procedure to rank tweets that describe an event best by their informativeness.

- We conduct an extensive evaluation of the system, by presenting its output to a pool of human annotators who are unbiased towards the system. We also compare the system output to gold standard events from curated calendars on the Web, to assess the system's recall. Some components are evaluated in isolation.

4.2 Related Work

Our system is an adaptation of the system proposed by Ritter et al. (2012), referred to by them as TWICAL. Explicit displays of knowledge of events in tweets are detected by scanning for the joint and frequent occurrence of a reference to a point in time, a so-called event phrase, and a named entity. The number of tweets in which an entity is mentioned with the same date is used as a signal

to extract significant events as opposed to mundane or personal events. Events are ranked by the fit between the date and entity, leading to a precision at the ranked top 100 events of 90% and a precision at 500 of 66%. An advantage of TWICAL is that it does not pose any restrictions on the type of event that is extracted, making it an open-domain approach: any event that people refer to with a future date can be found.

To our knowledge, no follow-up research has been carried out to replicate or further develop the research by Ritter et al. (2012). A reason could be that the approach relies on supervised natural language processing tools that are not readily available. To put the approach in a wider perspective, we give an overview of approaches that aim to find real-world events from tweets. We make a distinction between event extraction and event detection, and between the detection of known and unknown event types.

4.2.1 Event extraction

A comparable approach to TWICAL is proposed by Weerkamp and de Rijke (2012). Rather than scanning tweets for a variety of temporal expressions, they focus on the Dutch word 'vanavond' ('tonight'). Tweets are compared to a background corpus to highlight distinctive activities, and co-occurrence patterns are relied on to find the most important activities of the upcoming evening and night.

Both TWICAL and the approach of Weerkamp and de Rijke (2012) rely on the explicit mentioning of the time of future events. This general clue leads to the extraction of open-domain events of unknown types. Approaches that aim to find events of a known type focus on other clues in the short messages on Twitter, such as marker words or hashtags. Sakaki, Okazaki, and Matsuo (2010) aim to find earthquakes by harvesting tweets that mention (a variant of) the word 'earthquake' and relate the location at which they were posted to geological faults, enabling them to forecast the progression of the earthquake along the faults. Benson, Haghighi, and Barzilay (2011) focus on the extraction of music events in the region of New York City, and scan tweets for mentions of an artist and venue.

In another strand of research, events and their properties are retrieved from an event database, and the task is to identify tweets that refer to the event and may provide additional information (Jackoway et al., 2011; Reuter & Cimiano, 2012; Becker et al., 2012). Becker et al. (2012) refer to this task as *event identification*.

4.2.2 Event detection

Event detection, as opposed to event extraction, is typically focused on discovering events that have happened already and are having an effect on social media. A valuable clue for such events is an unexpected rise in usage, or *burstiness*, of a set of terms. Research has shown that the collective of Twitter users functions as a real-time sensor of social and physical events (S. Zhao, Zhong, Wickramasuriya, & Vasudevan, 2011): posts about significant events can be found on Twitter right after they occur. A diversity of approaches make use of such information.

In several works, tweets are clustered by their similarity, and bursty clusters are selected as events. Petrović et al. (2010) were one of the first to apply online clustering, by Locality Sensitive Hashing, to a large amount of tweets. Incoming messages are either linked to an existing cluster or grouped into a new one, depending on the distance to their nearest neighbour. Events are distinguished from non-event clusters based on the growth rate of a cluster. Many variations of this approach have been applied since, leveraging user and network information in clusters to better identify events (Aggarwal & Subbian, 2012; Kumar, Liu, Mehta, & Venkata Subramaniam, 2014), clustering tweets based on their (semantically expanded) hashtags (Ozdikis et al., 2012) and applying tweetLDA (W. X. Zhao et al., 2011) to find bursty topic models (Diao et al., 2012). McMinn et al. (2013) describe a corpus of events to evaluate event detection from tweets, and compare the approaches by Petrović et al. (2010) and Aggarwal and Subbian (2012).

Apart from tweet clustering, single text units might form the starting point of event detection. Weng and Lee (2011) focus on the clustering of single terms, by approaching each term in a tweet as a signal and applying wavelet analysis to terms. Signals that correlated in time are clustered together as an event. C. Li et al. (2012) take an approach similar to that of Weng and Lee (2011), but focus on segments of words rather than single words. Cordeiro (2012) extracts hashtags as wavelet signals and selects bursty hashtags as events. Weiler et al. (2013) combine the detection of temporally co-occurring tweets with geographical co-occurrence information, as Twitter users close to the action might be the most reliable source.

A valuable clue for event detection other than bursty topics or terms is the influence of real-world events on emotions. Indirectly, emotion bursts (mood swings) could indicate events. Ou et al. (2014) monitor emotion throughout

Twitter communities, and look for bursty emotion states. Valkanas and Gunopulos (2013) aggregate emotions in tweets by location.

Although the discussed event detection methods mostly rely on clues after an event has occurred, these clues relate to a variety of unknown event types and will also be sensitive to picking up clues to events that have not occurred yet but are mentioned nonetheless. Many events will be preceded by a rise of anticipatory tweets, though more gradual and diffuse and over a longer period of time than the sudden burst that the actual occurrence of an event may cause. This is relevant for the detection of future events. Ritter et al. (2012) demonstrate that their future event extraction approach leads to a result with a higher precision than the baseline burstiness-based event detection approach. Explicit mentions by Twitter users gathered over a longer time seem to be a more reliable information source than the short-lived, sudden burstiness of terms.

4.3 System Outline

TWICAL represents events by four units of information: the calendar date, a named entity, an event phrase, and an event type. In comparison, our system represents events by two information units: their calendar date and one or more *event terms*: words or word n-grams that represent the event. These terms may implicitly include both the named entity and the event phrase that are part of TWICAL. In contrast to the named entity and event phrase, event terms emerge from an unsupervised procedure.

Processing within our system is divided in three stages. The first is tweet processing, during which potential key event information, date mentions and event terms are extracted from single tweets in the Twitter stream. The second stage is event extraction, during which the strongest pairs of dates and event terms are extracted as events. The final stage is event presentation, during which additional event terms are extracted, the final set of event terms is selected and ordered, and tweets that mention an event are ordered.

We describe and motivate the different components below. A separate evaluation of the most important components is presented in Section 4.6.2.

4.3.1 Tweet processing

The setting of our experiment is the Dutch Twitter verse. Taking a relatively lesser used language is illustrative of a situation in which we cannot rely on standard English tools. We used TwiNL to simulate operating on the live Twitter

stream (see Section 4.4.1 for more details). All tweets are tokenised[2] and turned to lower case. Each tweet is then scanned for TIMEXs. Tweets that contain a TIMEX are fed to the second component in this stage, concept extraction. All other tweets are discarded.

Extraction of time expressions

In view of our aim of future event extraction, we are only interested in TIMEXs that indicate a future date. It is important to extract a large amount of TIMEXs during this stage, as all tweets that are not found to have a TIMEX are discarded. Apart from extracting TIMEXs, an additional transformation step is needed that maps a TIMEX to a future date.

Dutch TIMEXs can be extracted by means of the Heideltime tagger (Strötgen & Gertz, 2010). Testing the Heideltime tagger, we observed that it misses many future TIMEXs. Another disadvantage is that it not always specifies the future date to which a TIMEX refers. We therefore manually formulated a more comprehensive set of rules. We distinguish three kinds of TIMEXs: 'Date', 'Weekday', and 'Exact'. When any of the rules are matched, it is translated into an explicit future date. Appendix B can be consulted for a complete overview of the rules. An empirical comparison between our approach and the Heideltime tagger is given in Section 6.2.1.

The 'Date' category of rules consists of the different variations of date mentions in Dutch. If a month is matched without a day, this is not considered specific enough and there is no match. When no year is included, we assume that the date refers to the next occurrence of the date. Any date that refers to a point in time before the tweet was posted is not taken into consideration.

The 'Exact' rules comprise a variety of phrase combinations that specify an exact number of days ahead. Most of them are Dutch variations of 'x days until', but also 'overmorgen' ('the day after tomorrow') is included. We did not include 'morgen' ('tomorrow' or 'morning') to avoid the large amount of ambiguous tweets that would be returned by this TIMEX, overwhelming the other output. 'Vanavond' ('tonight') was also excluded from these rules. For any tweet matching the exact rules, we calculated the date by adding the mentioned number of days ahead to the post date of the tweet.

The 'Weekday' rules match a mention of a weekday, optionally preceded by the phrase 'volgende week' ('next week') or followed by 'ochtend' ('morning'), 'middag' ('afternoon'), 'avond' ('evening') or 'nacht' ('night'). The weekday is

[2]Ucto was applied for tokenisation: `https://languagemachines.github.io/ucto/`.

translated into a date by computing the number of days to the forthcoming occurrence of the weekday after the time of the tweet post, and adding seven days if the weekday is preceded by 'volgende week'. To exclude tweets that refer to the previous occurrence of the weekday, and thus to a past event, we scanned the tweets that match a weekday for verbs in the past tense by applying automatic part-of-speech tagging with Frog (van den Bosch, Busser, Canisius, & Daelemans, 2007), a Dutch morpho-syntactic tagger and parser. Tweets containing a verb in the simple past or past perfect were discarded.

Our system gives preference to the most specific TIMEX if more than one TIMEX is seen in a tweet. A TIMEX matching an exact rule is preferred over a TIMEX matching a weekday, and an exact rule matching TIMEX is overruled by a TIMEX matching a specific date. If a tweet contains more than one future time reference from the same rule type, both future dates are related to the tweet.

Extraction of concepts

After having extracted tweets that contain a reference to a future date, these tweets are scanned for entities that the time reference might relate to. The goal here is to select n-grams that relate well to an event. The entities that are extracted during this stage are subsequently paired up with the dates with which they co-occur. To achieve the extraction of a wide range of event types, it is important to achieve a high recall of entities.

Off-the-shelf Natural Language Processing tools have shown poor performances for Named Entity Detection from Twitter data. This is mainly due to deviating spelling on Twitter and the large number of entities that are mentioned on this platform (Ritter et al., 2012). Clues that might assist Named Entity Detection, such as capitalisation and Part-of-speech tags, are less reliable on Twitter. Ritter et al. (2011) trained a Part-of-Speech tagger on annotated tweets, outperforming the Stanford tagger by a considerable margin. The tagger was used in TWICAL to detect entities in tweets.

Rather than developing a Part-of-Speech tagger for Dutch tweets ourselves, we chose to apply the *commonness* metric, as formulated by Meij, Weerkamp, and de Rijke (2012). They match the word n-grams in a tweet with equally named Wikipedia articles, and assign a score to such n-grams based on their commonness in Wikipedia. By leveraging the crowd-sourced platform of Wikipedia, on which many entities are described and added, we expected to extract a wide

range of event types. We compared this approach to the performance of an off-the-shelf system for Named Entity Detection in Dutch, and found that the former yields a significantly better performance. See section 6.2.2. for a description of this experiment.

Commonness is formulated as the prior probability of a concept c (the n-gram) to be used as an anchor text q in Wikipedia (Meij et al., 2012):

$$\text{Commonness}(c, q) = \frac{|L_{q,c}|}{\sum_{c'} |L_{q,c'}|} \tag{4.1}$$

Where $L_{q,c}$ denotes the set of all links with anchor text q pointing to the Wikipedia page titled c, and $\sum_{c'} |L_{q,c'}|$ is the total sum of occurrences of q as an anchor text linking to any concept (including c).

Meij et al. (2012) aim to identify the main concept that a tweet refers to automatically, based on whether the concept is mentioned on Wikipedia. Concepts are often named entities; they can be a product, brand, person, city, event, et cetera. Meij et al. (2012) compared several approaches to link a tweet to a concept, including supervised machine learning, and found that the relatively simple and unsupervised commonness metric already leads to a very good performance. Other advantages of this metric are that it can be applied to any language in which Wikipedia pages are available, it is adaptive to new concepts, and it does not rely on capitalisation or preceding words to extract concepts from a text.

We downloaded the Dutch Wikipedia dump of November 14, 2013,[3] and parsed it with the Annotated-WikiExtractor[4]. Then, we used Colibri Core[5] to calculate the commonness of any concept that has its own Wikpedia article, and is used as an anchor text on other Wikipedia pages at least once. These statistics are used to extract concepts from a tweet. Tweets that matched a future time reference in the first stage are stripped of this time reference, and n-grams with n up to 5 are extracted. Any n-gram that is found to have a commonness score which is above 0.05 is extracted as a concept.

In addition to explicit references to events in tweets, events might be referred to implicitly with hashtags. These can be seen as user-designated keywords, and are often employed as an event marker. To include this information we selected any hashtag in a tweet directly as event term. Although some hashtags will not relate to an event, we assumed these would be filtered in the subsequent event ranking stage.

[3] http://dumps.wikimedia.org/nlwiki/nlwiki-20131114-pages-articles.xml.bz2
[4] https://github.com/jodaiber/Annotated-WikiExtractor
[5] http://proycon.github.io/colibri-core/doc/

4.3.2 Event extraction

The goal of the event extraction phase is to rank date–term pairs co-occurring in the selected tweets by their fit. As multiple terms might all fit one event, an additional clustering step is performed to link these to each other.

At this point, the system has obtained a list of date–term pairs and the tweets in which they occur. The aim of the next stage is to select the pairs that represent an event.

Event ranking

The pairs of dates and event terms that result from the tweet processing stage represent events with a varying degree of significance. The current step serves to quantify this degree and rank the date–term pairs accordingly, and is central to the extraction of events.

A first criterion for event significance is the number of times an event is tweeted about. Ritter et al. (2012) employ a minimum of 20 tweets for a named entity to qualify as a potential event. We set the threshold to 5, which is more in line with the lower density of Dutch tweets.

As a second criterion, named entities more frequently mentioned with the same date are seen as the more significant events. This follows the intuition that many significant events are attended, viewed or celebrated by many different persons on the same date. On the other hand, the less significant, personal events take place on different dates for different persons. Following Ritter et al. (2012), we calculate the fit between any frequent event term and the date with which it is mentioned, by means of the G_2 log likelihood ratio statistic:

$$G_2 = \sum_{z \in \{e, \neg e\}, y \in \{d, \neg d\}} O_{z,y} \times \ln \left(\frac{O_{z,y}}{E_{z,y}} \right) \tag{4.2}$$

The fit between any event term e and date d is calculated by the observed (O) and expected (E) frequency of the four pairs $\{e, d\}$,$\{e, \neg d\}$,$\{\neg e, d\}$ and $\{\neg e, \neg d\}$. The expected frequency is calculated by multiplying the observed frequencies of z and y and dividing them by the total number of tweets in the set.

Arguably, events that are tweeted about by many different users are of a higher significance than events that are referred to by only one or two Twitter users who repeatedly post messages about the same events. We implemented this intuition by multiplying the G_2 log likelihood ratio statistic with the fraction of different users that mention the event. The events are ranked by the resulting G_2u score:

$$G_2 u = \left(\frac{u}{t}\right) \times G_2 \tag{4.3}$$

Here, u is the number of unique users that mention the date and entity in the same tweet, while t is the number of tweets in which the date and entity are both mentioned.

The calculation of $G_2 u$ for each pair results in a ranked list of date–term pairs. To reduce subsequent computational costs, we discarded all pairs with a rank number below 2,500. In other words, at any point, we are computing the top 2,500 most significant date–term pairs.

Event clustering

As an event might be described by multiple event terms, it is likely that the ranked list of date–term pairs contains several event terms that describe the same event. Ritter et al. (2012) report on such duplicate output from their system. This is unfavourable in view of the redundant information that a user of the system would have to process. In addition, a single entity might be a poor representation of an event that comprises multiple entities, such as the two opposing teams of a football match. We believe that clustering is an effective way to decrease duplicate output and enhance event representations.

Arguably, if two event terms refer to the same event, this analogy is reflected in the words other than these event terms, in the tweets that mention them. Hence, we compare the tweets from which two date–term pairs were extracted to decide if they should be combined. Clustering is performed by means of Agglomerative Hierarchical Clustering (Day & Edelsbrunner, 1984). The advantage of this algorithm is that it does not require a fixed number of clusters as parameters, but rather allows to cluster up to a specified similarity threshold. This is precisely what we want, as there is no indication of the number of clusters beforehand.

As a preparation for clustering, each set of tweets in which the same date–term pair occurs is aggregated into one big document. Subsequently, the documents are converted into a feature vector with $tf * idf$ weighting (Day & Edelsbrunner, 1984). The idf value is based on all aggregated documents in the set of 2,500 date–term pairs.

A useful constraint for date–term pairs to be clustered is that the dates of the two pairs be equal. Instead of generating a similarity matrix of all 2,500 date–term pairs, a similarity matrix is generated for each set of date–term pairs

labeled with the same date. The cosine similarity (Steinbach et al., 2000) is calculated between each date–term pair in such a set, based on their feature vector, and the similarity pairs are ranked from most similar to least similar. Each date–term pair forms an initial cluster with only one event term. Starting from the two clusters that are most similar, they are merged if their similarity is above threshold x. This process was repeated until the highest ranked similarity was below x. We chose to apply single-link clustering rather than calculating a centroid after each merge, so as to reduce computational costs. Hence, only the initial similarity table is used, and one combination of event terms with an above-threshold similarity suffices to merge for example two large clusters.

Whenever two clusters were merged into a new cluster, the metadata of the two former clusters were merged in the following way:

- The event terms were combined. Any duplicate event terms (typically occurring when clusters with multiple event terms are clustered together) are removed;
- The event tweets were combined. Again, duplicate tweets are discarded;
- The cluster is assigned the highest G_2u score of the two former clusters.

The threshold x was empirically set to 0.7, by testing on the first two days in the tweet set that is described in Section 4.4.1. Arguably, event clusters that comprise multiple actual events are more harmful than duplicate events. We therefore preferred a precision oriented clustering, with a minimum amount of false positives. An evaluation of clustering performance is presented in Section 4.6.2.

Event filtering

Although we try to discard references to a past weekday falsely identified as a coming weekday, by scanning the tweets for verbs in the past tense, some references might still surpass this filter. For example, the (translated) tweet 'State police takes over Ferguson safety - Thursday, Missouri's state police has ... URL' clearly refers to a past event, while it does not contain a verb in the past tense. As such news reports are often repetitively forwarded in an unaltered form, we add another filter by discarding any event with a type–token ratio below 0.40.

The type–token ratio is calculated from the tweets of an event by dividing the number of different words in the tweets by the total number of word tokens. A low type–token ratio indicates repetition; a high type–token ratio indicates a high variance of words, and may represent an event that is referred to from

different angles. With a threshold of 0.4 we aim to filter the events that were described with the most repetitive tweets, while minimising the chance to discard any event with a more diverse vocabulary.

As an example, the tweets listed below typically represent tweeted news headlines. They refer to an event that took place the past Thursday, which is falsely identified as the upcoming Thursday. Apart from the short URLs these tweets are identical and would not pass the type–token filter (type–token ratio = 0.36).

- repeated trouble after Thursday Meppel day ' #police arrests couple after violence URL
- repeated trouble after Thursday Meppel day ' #police arrests couple after violence URL
- repeated trouble after Thursday Meppel day ' #police arrests couple after violence URL

In comparison, the tweets below are typical anticipations of a social event, and do pass the filter (type–token ratio = 0.76).

- guys, all world trouble aside, in two weeks something more important will start: the new season of Doctor Who!
- omg 23 August the new Doctor Who?! will start
- only 6 days until the new Doctor Who!!!! #excited

4.3.3 Event presentation

Resolving overlap of concepts

An extracted event is potentially represented by several event terms, as a result of the clustering stage. These event terms might have overlapping semantic units. Consider for example the event terms 'mario kart 8', 'mario kart' and 'kart'. The latter two terms are redundant with respect to the first, and including them would result in a superfluous representation. We describe the procedure that was undertaken to remove such redundancy.

First, we rank the event terms by their commonness score. A list of 'clean' event terms is initiated, which at first consists only of the event term with the highest rank. Starting from the second ranked event term, the term is compared to the list of clean event terms and added to this new list when no overlap exists with any of the terms in this list. An overlap occurs if two event terms have

overlapping word tokens. Thus, event terms are only added to the clean list if they contain completely new information. Hashtags are seen as a unigram for this comparison, and are stripped from their hashtag symbol ('#'). This way, a redundant presentation that would concatenate for example 'pukkelpop' and '#pukkelpop' or 'mario kart' and '#mario', is avoided. As hashtags are not linked to a commonness score, they are at the bottom of the list, so that only non-overlapping hashtags are added to the resulting list.

Enriching the event description

Terms that present an event should ideally provide a sufficient summarisation of the event, in the same way as a news headline. We added a method to our framework to enrich the existing event terms with additional terms. The method is unsupervised and bases the addition of terms on the set of tweets that announce the event. The procedure is described below:

- The event tweets are aggregated into one document, and the word tokens are sorted by their importance to the event based on their $tf * idf$ weight. $tf * idf$ is calculated in relation to the other event documents in the set;
- The five types with the highest $tf * idf$ are extracted, and any of them is added to the list of existing event terms, if:
  - it does not resemble or overlap with one of the existing event terms;
  - it is identified either as a verb, noun, adjective or adverb by a generic part-of-speech tagger (van den Bosch et al., 2007).

The part-of-speech tag is checked in order to exclude user names, URLs, and numeric word types, which we consider insufficient event descriptors. In addition to nouns, which might describe entities that relate to the event, we focused on verbs, which might describe an action associated with the event (such as 'confirmed' if a music artist is announced for a festival) as well as adjectives and adverbs that might describe properties of the event (such as 'free' if an event can be attended for free).

Ordering of event terms

For the event terms to provide a sufficient summary of the event, they should be presented in a proper order. For example, for the terms 'outdoor', '#db14' and 'decibel' that describe the Decibel Outdoor Festival, the proper order would arguably be 'decibel', 'outdoor', and '#db14'. We set the order for event terms by

calculating their average position in the event tweets and sorting them accordingly.

Ranking of tweets

In relation to the event terms that provide a summary of an event, tweets can be consulted for a more detailed description. The informativeness of these tweets, however, might be low if only near-duplicates are shown at the top (Tao, Abel, Hauff, Houben, & Gadiraju, 2013). To make sure that the top tweets are diverse and yet descriptive of the event, they are automatically re-ordered:

- The tweets that describe an event are sorted by their importance to the event. The importance of a tweet is scored by the summed $tf * idf$ values of the words in the tweet. These values are in line with the ones that were generated in the term addition procedure (Section 4.3.3). The intuition is that words with a high $tf * idf$ are more specific than words with a low $tf * idf$, and are likely to describe key aspects of the event. By summing up the $tf * idf$ values of a tweet, its descriptiveness can be scored heuristically.
- The re-sorting of tweets is iterated, and any tweet that overlaps for over 80% with a higher ranked tweet is transferred to the bottom of the list. This procedure runs until every tweet has been processed. The result is a re-ordered list of tweets.

4.4 Experimental Set-up

4.4.1 Data

We collected the Dutch tweets posted in August 2014 from TwiNL to test our system on. The sample of tweets in August totalled 27,682,311 individual posts.

4.4.2 Precision evaluation

To test the different components we apply three versions of our system: Ngram, Commonness, and Commonness+. The names of the versions refer to the way in which event terms are generated.

The Ngram system acts as a baseline. It has a different approach to concept extraction from tweets: rather than basing the extraction on an above-threshold commonness score, any n-gram with $n \leq 5$ qualifies as a concept. Accordingly, any n-gram that has a good fit with a date might be clustered with n-grams with similar tweets to form an event. The Commonness system does not include the

addition of event terms (described in Section 4.3.3). This is to evaluate the value of this component to the description of events. Commonness+ comprises the full system as described in Section 4.3.

By incorporating the variants Commonness and Ngram we test two objectives of Commonness+: the accurate extraction of events and a proper presentation of events. We evaluate their output on these two aspects.

Of the 27.7 million tweets, 367,232 were found by our systems to have a TIMEX. 270,440 of these contain at least one concept or hashtag; 1.99 on average per tweet, and 731,497 in total.[6] We evaluated the top 250 events of the three systems, as ranked by the G_2u score. We asked thirty Dutch annotators who had no background knowledge of the systems to assess fifty events from the output. We made sure that these fifty events represented a balanced set of events from all three systems. Additionally, we shuffled the event rankings to make sure that the annotator would encounter higher and lower ranked events from each system. The annotators did not know that the presented output originated from one of three systems or were related to a ranking.

For the layout and distribution of the evaluation, we made use of survey tool Qualtrics[7]. Each annotator was sent a unique survey with fifty specifically assigned events. For each event, the annotator was presented with the five top-ranked tweets and was asked whether these tweets refer to the same event. At the start of the survey, the annotator was given a definition of what is an event: 'An event takes place at a specific point in time and has value for a larger group of people'. The annotator was also told that the five tweets might refer to different sub-events that relate to one overarching theme. In these cases, they should assess the overarching theme as event or no event. If the theme is an event itself, such as a football match, the tweets can be assessed as referring to the same event, whereas the name of a city (occurring in tweets referring to different events in that city) as overarching theme does not qualify as an event. The complete instructions that were shown to the annotators (translated from Dutch) is included in Appendix C.

Whenever the annotator assessed the five top-ranked tweets as referring to an event, he was subsequently presented with the event terms. The task then was to assess, on a scale from 1 to 3 (corresponding to 'poor', 'moderate' and 'good'), how well the terms relate to the event that was identified. If an annotator

[6] All tweetids along with the extracted TIMEXs and entities per tweet are accessible from `http://dx.doi.org/10.17026/dans-227-36wn`.

[7] `http://www.qualtrics.com`

did not identify an event in the five tweets, he would directly move on to the next output.

Each output was assessed by two annotators, to obtain a sense of agreement. As the Commonness and Commonness+ variants only differ by their term output while their event output is the same, the latter output is by implication assessed by four annotators.

4.4.3 Recall evaluation

Targeting open-domain events, we cannot perform a complete recall evaluation. As an approximation, we made a selection of six event types that should arguably be extracted by our system, and collected date–event pairs from manually curated event calendars on the Web. These gold standard events give an impression of the quality of event extraction by event type.

We selected six common types of social events, and looked for websites that provide an overview of events of these types in the Netherlands. We collected the source code of the calendar overviews from each of these Web pages, and parsed the HTML code to extract gold standard event names along with their date. We chose to focus on events in August and September 2014, the months closest in time to our tweet set. An overview of the event types and the websites from which we collected event calendars is given below:

- Matches in the top-level Dutch national football league, the *Eredivisie*. We extracted an overview of the matches in the 2014–2015 season, starting August 8, from `sport.infonu.nl`,[8] and selected all matches in August and September.

- Public events; local or national events that take place at a single location, such as expositions, carnivals and parties. We scraped the overviews of August and September as listed on `www.evenementenkalender.nl`,[9] a calendar website to which anyone can submit events. Submitted events are checked by administrators before being placed on the calendar.

- Music Festivals. We extracted an overview from `http://www.festivalinfo.nl/`,[10] a popular festival website maintained by volunteers, that aims to provide an exhaustive overview of bigger and smaller music festivals in The Netherlands and Belgium.

[8] `http://sport.infonu.nl/voetbal/128666-speelschema-eredivisie-2014-2015-programma-en-uitslagen.html`

[9] `http://www.evenementkalender.nl/2014-08` and `http://www.evenementkalender.nl/2014-09`

[10] `http://www.festivalinfo.nl/festivals/?type_select=maand`

	#Curated	#Mentioned		#Mentioned ≥ 5	
Football matches	63	51	(81%)	40	(63%)
Public events	2361	63	(3%)	30	(1%)
Music festivals	518	195	(38%)	98	(19%)
Movie premieres	50	29	(58%)	20	(40%)
Game releases	79	19	(24%)	14	(10%)
Stage performances	1,066	85	(8%)	29	(3%)
Total	4,137	442	(11%)	231	(6%)

TABLE 4.1: Overview of the number of events that were collected as gold standard for recall evaluation. The percentages between brackets give the share of the curated total.

- Releases of computer games. We extracted a list of game release dates in August and September 2014 in The Netherlands on any gaming platform, from `www.gamersnet.nl`,[11][12] a website maintained by professional editors.
- Movie Premieres. A list of Dutch movie premiere dates in August and September 2014 was extracted from `www.filmvandaag.nl`,[13][14], a website maintained by professional editors.
- Stage performances: music concerts and theater plays. We extracted an overview of performances in August and September 2014 from `www.podiuminfo.nl`,[15][16] a website that is linked to `www.festivalinfo.nl`.

We performed a subsequent filtering by removing gold standard events that are not mentioned in our tweet set. We compared the name of each event to each of the 27.7 million tweets and listed all tweets that refer to an event name. We subsequently inspected the list of matching tweets to see if they actually mention the event, which is not self-evident for event types such as movies. Any falsely selected tweet was discarded from the list. We performed a second filtering by imposing a minimum threshold of five tweets per event, which is equivalent to the threshold for event significance during the system component of event ranking (Section 4.3.2).

The numbers of gold standard events by type, before and after filtering, are given in Table 4.1. Viewing the total, only a small part of the gold standard events are actually mentioned on Twitter (11%). Furthermore, only about half of these are mentioned five times or more (6%). The bulk of the gold standard

[11] `http://www.gamersnet.nl/gamereleases/201408/`
[12] `http://www.gamersnet.nl/gamereleases/201409/`
[13] `http://www.filmvandaag.nl/bioscoop/08-2014`
[14] `http://www.filmvandaag.nl/bioscoop/09-2014`
[15] `http://www.podiuminfo.nl/concertagenda/?input_zoek=&Date_Day=01&Date_Month=08&Date_Year=2014`
[16] `http://www.podiuminfo.nl/concertagenda/?input_zoek=&Date_Day=01&Date_Month=09&Date_Year=2014`

events are stage performances or public events. However, a long tail of events in these sets is either never or hardly ever mentioned in the tweets. The set of football matches are referred to for the largest part (81%), followed by movie premieres (58%). The type of events that is mentioned the most is music festival, with 195 events mentioned at least once and 98 events mentioned five times or more.

For recall evaluation, the events extracted by the Ngram and Commonness system are compared to the gold standard events that are mentioned in at least five tweets.

4.5 Results

4.5.1 Output

We display the top ranked output of the test on August 2014 tweets in Table 4.2.[17] Nine of the ten output units are judged as a significant event by at least three of four annotators. The event described by the term 'werkstress', referring to personal insights into the cause of sleepless Sunday nights, is judged as such by two annotators. As the event is not characterised by one specific date, it is arguably incorrectly extracted as event. Music festivals are most dominant in this ranking (rank 1, 2, 4, 6, 8 and 9). This relates to the summer period during which the tweets were posted. Other event types are the release of a device (#iphone6), a music concert (Ben Howard), and a football match (#azaja, AZ Alkmaar vs. Ajax). The 'Decibel' festival is represented twice, at rank 4 (decibel) and rank 8 (#db14). While the two output units should have been clustered together, it appears that the dissimilar language in the tweet sets has prevented this. The tweets that mention 'decibel' focus more on specific performances during the festival as well as the forecasted bad weather, while the users that mention '#db14' are mostly looking forward to the event.

Inspecting the event terms for Commonness and Commonness+, the former often only provides one term, while the latter is more informative about the event. For example, for the 'Appelsap' festival, the additional terms provide information on the type of event and the venue at which it takes place.

To obtain insights into the range of dates at which the events take place, we plotted the number of extracted events per week within rank 250 in Figure 4.1. The events are more concentrated close to the tweet postings in August (week 31–35). The number drops below ten events from week 39 (September

[17] An overview of all events that were extracted, including their tweet ids, is accessible from `http://dx.doi.org/10.17026/dans-227-36wn`.

Event rank	Event terms: Commonness	Event terms: Commonness+	Event tweet (translated from Dutch)	Judged as significant (by % of annotators)
1	appelsap	appelsap, festival, oosterpark	I want to go to Appelsap Saturday but none of my friends wants to join. Can I join anyone? #dta #appelsap	100%
2	dutch valley	radio, dutch valley, spaarnwoude	After the success of Dance Valley, this Saturday it is time for Dutch Valley. Will you go and who would you like to see? Watch URL	100%
3	#iphone6	aangekondigd, apple, #iphone6	Add to your calendar, on September 9th Apple will reveal the iphone 6 URL #iphone #iphone6 #apple	100%
4	decibel	decibel, zin, outdoor	Celebrating my birthday at Decibel on Saturday #db14 — at Decibel Outdoor Festival URL	100%
5	ben howard, hmh	ben howard, heineken, hall, hmh	Life goal 'attending a Ben Howard concert' is almost achieved. tickets in tha pocket! 18 dec @hmh #soexcited #benhoward #hmh He is genius.	100%
6	mysteryland	zin, mysteryland	Only 4 nights and then... Mysteryland!! Hope the sun will brightly shine that day so we can make a party under the sun #mysteryland	100%
7	werkstress	werkstress, zorgt, slapeloze	labour stress leads to sleapless Sunday nights. URL do you recognize this?	50%
8	encore	encore, festival, ndsm, werf	Encore Festival, NDSM-werf: on August 31 Encore Festival will take place at the NDSM-werf in Amsterdam. This... URL #news	100%
9	#db14	decibel, outdoor, #db14	I have only 1 ticket for sale for the Decibel Outdoor Festival this Saturday: URL #db14	100%
10	#azaja	blom, ajax, #azaja	Blom is the designated referee for AZ-Ajax Sunday #ajax #az #azaja	75%

TABLE 4.2: Top 10 ranked events from the Commonness systems.

22nd) onward, but never touches zero in any of the subsequent weeks. Hence, although the bulk of anticipations concerns events within a couple of weeks, our system captures tweets that refer to events taking place months ahead.

4.5.2 Precision

The precision-at-250 of the Ngram baseline and the Commonness approach (which is the same for Commonness and Commonness+) is displayed in Table 4.3. As the output of the Commonness approach was rated by four annotators, the precision can be scored with different degrees of strictness: labeling output as event only when all four annotators identify them as event, when at least three of the four see them as representing an event, and when half of the annotators do so. The results in the table show that almost two-thirds (63%) of the output of the Commonness approach is seen as event by all four annotators, while only 42% is scored as such for the Ngram approach. When taking a majority vote of three annotators, the percentage increases to 80%, while a lax setting in which two or more of the annotators identify an event yields a precision of 87%.

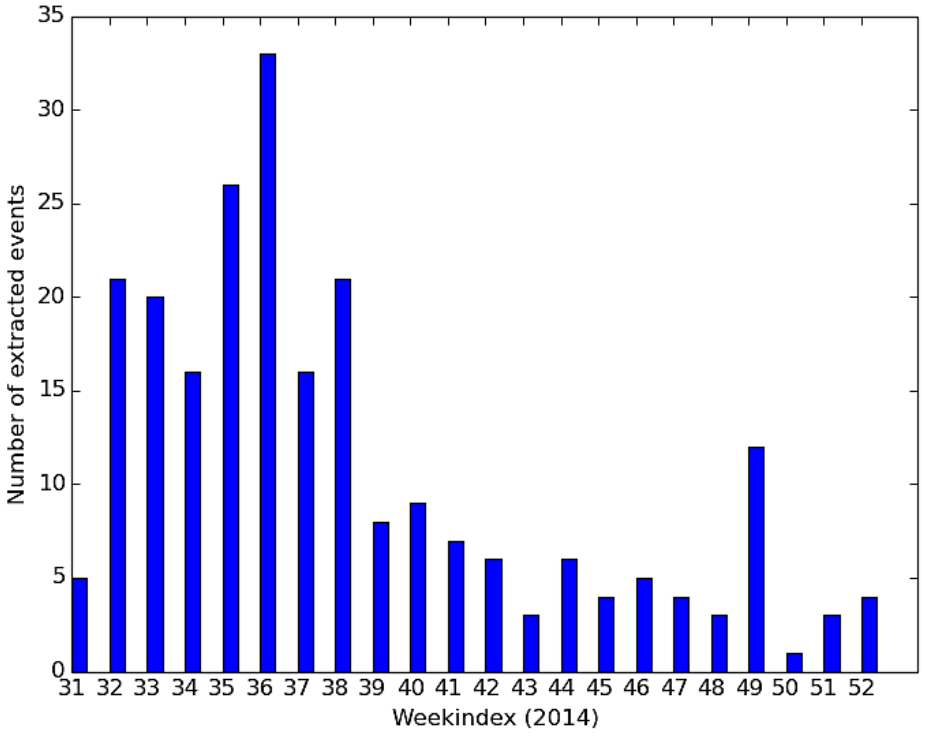


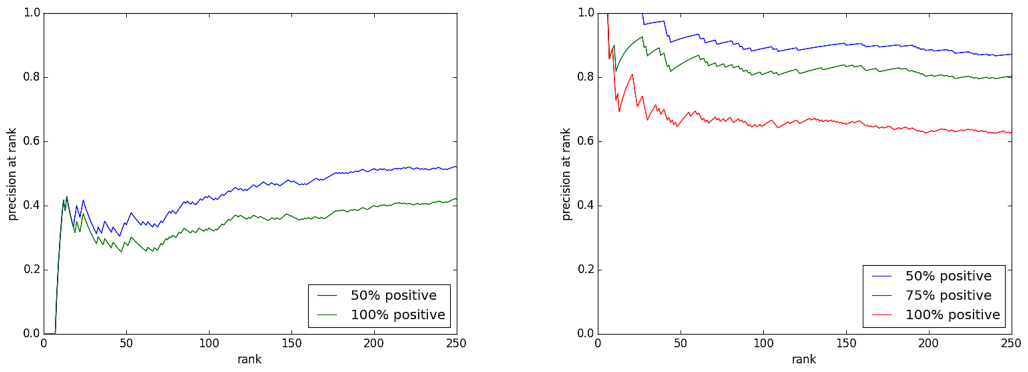
FIGURE 4.1: Counts of the number of extracted events by week number in 2014, from the top 250 events extracted from the Twitter stream in August 2014 (weeks 31–35).

	Precision-at-250				
	100%	75%	50%	Cohen's Kappa	Mutual F-score
Ngram	0.42	-	0.52	0.80	0.89
Commonness	0.63	0.80	0.87	0.48	0.90

TABLE 4.3: Precision-at-250 of output identified as event by human annotators at 100%, 75%, and 50% agreement, and Cohen's Kappa and Mutual F-score between the annotators.

We scored the inter-annotator agreement by Cohen's Kappa and Mutual F-score. The latter provides insight into the agreement for the positive (event) class. The Kappa score for the Ngram approach is substantial (Landis & Koch, 1977) with 0.80, while the agreement for the Commonness events is only moderate with 0.48. However, the mutual F-score shows that the agreement for the positive event class in both approaches is quite accurate with 0.89 and 0.90, respectively.

We plot the precision-at from rank 1 to 250 for the two approaches in Figure 4.2. Surprisingly, the curves for the Ngram approach show an increasing performance lower down the ranking. It seems that the G_2 log likelihood ratio statistic by which the Ngrams are ranked does not relate well to the likelihood



(a) Ngram approach

(b) Commonness approach

FIGURE 4.2: Precision-at-curves for the Ngram and Commonness approach with different degrees of strictness.

	Avg. term assessment	Weighted Cohen's Kappa
Ngram	2.57	0.16
Commonness	2.69	0.10
Commonness+	2.63	0.21

TABLE 4.4: Average assessment of terms for all three approaches. Assessment is on a scale of 1 (bad) to 3 (good).

that the n-grams signify an event. In contrast, higher rankings for the Commonness approach do relate to event probability. For all three degrees of strictness, a plateau is reached after rank 60. Any output up to a ranking of about 50 is seen by at least two annotators as an event.

Upon identifying output as event, annotators are asked to assess the quality of the event terms in relation to the event. The outcome of these assessments is presented in Table 4.4. The assessment was given on a scale from 1 to 3, as a poor, moderate, or good representation. As the event terms were only presented if an annotator rated the tweets as representing an event, for each event these terms could be either assessed by 0, 1 or both annotators. When only one of two annotators gave an assessment, this single value represented the event assessment. When both annotators gave an assessment, their average was taken as the event assessment. For each system, the quality of event terms was calculated as the average of all event assessments. The agreement was scored with the Weighted Cohen's Kappa metric (Gwet, 2001), in line with the ordinal annotation. Missing fields were taken into account in this metric.

The average assessment of terms does not show a large difference between the three approaches. Surprisingly, the Commonness+ approach, for which

	Ngram		Commonness	
	Recall@250	Recall all	Recall@250	Recall all
Football matches	0.00	0.00	0.35	0.53
Public events	0.00	0.00	0.20	0.37
Music festivals	0.07	0.08	0.17	0.38
Movie premieres	0.00	0.00	0.10	0.25
Game releases	0.36	0.43	0.50	0.57
Stage performances	0.03	0.07	0.03	0.31
Total	0.06	0.07	0.21	0.40

TABLE 4.5: Recall performance by event type, based on a gold standard set of events that are mentioned in at least 5 tweets (see Table 4.1 for the exact numbers).

terms were added in a post-processing step, are generally assessed as a slightly worse representation than the terms for the Commonness approach. The agreement is only slight or poor. This is in line with post-hoc remarks made by several annotators that it was hard to assess the quality of the terms.

4.5.3 Recall

To assess recall, we collected gold standard events that took place in August or September 2014, for six event types from curated web sites (see Section 4.4.3). We compared the events that were extracted by the Ngram and Commonness system to the gold standard events that were tweeted about at least five times (the last column in Table 4.1). For both systems, we report a recall-at-250, which relates to the precision oriented evaluation in the previous Section, as well as a recall of all events (318 for Ngram and 966 for Commonness).

The results are given in Table 4.5. The Commonness approach outperforms the Ngram approach for each of the six event types. It yields the best performance in retrieving football matches and game releases. The Ngram approach fails to retrieve any event for some of the types. Overall, the Commonness approach scores a recall of 0.20 at rank 250, and a recall of 0.40 for all 967 events on these event types.

We chose to apply the G_2u formula (Section 4.3.2) to rank events, which is an extension of the G_2 formula. As a comparison of the two formulas, we implemented the Commonness system with both and scored the recall at each rank by comparing the extracted events to the accumulated gold standard events of all event types (242 events in total). The recall at each rank is plotted in Figure 4.3. The shorter line of the G_2 rank is due to a larger number of discarded events. The G_2u rank has a comparable recall to G_2 up to rank 100, but retrieves increasingly more events lower down the rankings. Although this evaluation

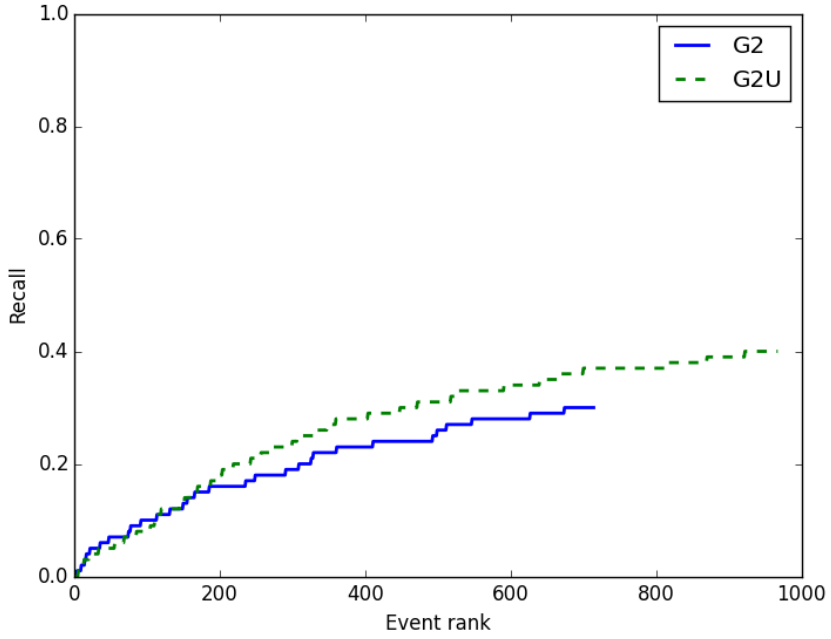


FIGURE 4.3: Recall-at-curves of the Commonness system, ranked by either the G_2 formula or the G_2U formula.

was performed on specific types of events, this outcome shows that favouring events that are mentioned by a higher diversity of users (as is done in the G_2u formula) may help to outrank insignificant output.

4.6 Analysis

4.6.1 Event output

To obtain insight in the causes of non-event output, annotator disagreement, and the assessment of event terms, as well as the impact of the event term clustering component, we analysed the top 250 events from the Commonness approach in relation to the annotator assessments. We analysed all 250 events on their five event tweets and the event terms of both the Commonness and Commonness+ approach.

Event annotation

Of the 250 annotated events, 157 are annotated by all four annotators as event, leaving 93 events that are deemed doubtful by at least one annotator. Of these 93 events, 44 are still annotated by three of four annotators as event, 17 by half of them, 14 by only one annotator, and for 18 entries all four annotators agree that they are not an event. We analysed the five tweets that were shown to the

annotator for these 93 events, and distinguished six causes for an annotator to doubt if all tweets represent the same event:

1. Side event - One or more tweets refer to an event that is related to or is a sub-event of the event that the other tweets refer to, and only loosely mention the event. Example: *rufus wainwright will perform at the 32nd Night of Poetry in Tivolivredenburg on Saturday 20 September URL*. The tweet mentions a performance as sub-event of the Night of Poetry, while other tweets only mention the Night of Poetry itself.

2. Too general event term(s) - The event tweets represent different events that are related to one or more general keywords. The general keyword does not refer to any single event. Example: *The first cup match is on Tuesday at 6:30 PM: GSVV A1 - V.V. Niekerk A1. #away #cupmatch*. This tweet mentions an event that is linked to the general keyword 'cup match', as do the other tweets in the set of five.

3. Outlier tweet(s) - most of the tweets represent the same event, but one of them clearly refers to something else. Example: *On Sunday September 7 the opening of Power of Water as part of the Uitfeest takes place! For education on the power of water ... URL*. While all other tweets refer to the 'Hiswa te water' event, this tweet points to another event that takes place on the same day, and also contains the word 'water' in the name.

4. Mundane event(s) - All tweets represent one or more events that are considered too mundane or personal. Example: *Looking for a ride on august 16 16:15 from Den Bosch to Amsterdam #ridealong #carpool #toogethr*. This tweet links to the event terms 'ride' and 'Amsterdam', and refers to the personal event of carpooling.

5. Discussion - The event tweets do not describe the event, but contribute to a discussion on the event. Hence, one can argue that the tweets refer to the discussion rather than to the social event itself. Example: *If Black Pete is prohibited I will still walk around dressed as Black Pete on the 5th of December, you know*. This tweet contributes to the discussion of the format of the 'Sinterklaas' celebration in the Netherlands.

6. Contest - The event tweets advertise about a product or participate in a contest. Example: *@afcajax because my friend is only free on Sunday and we would really like to go to the match together #weareajax*. This tweet joins a contest to win free tickets to a football match by stating a motivation.

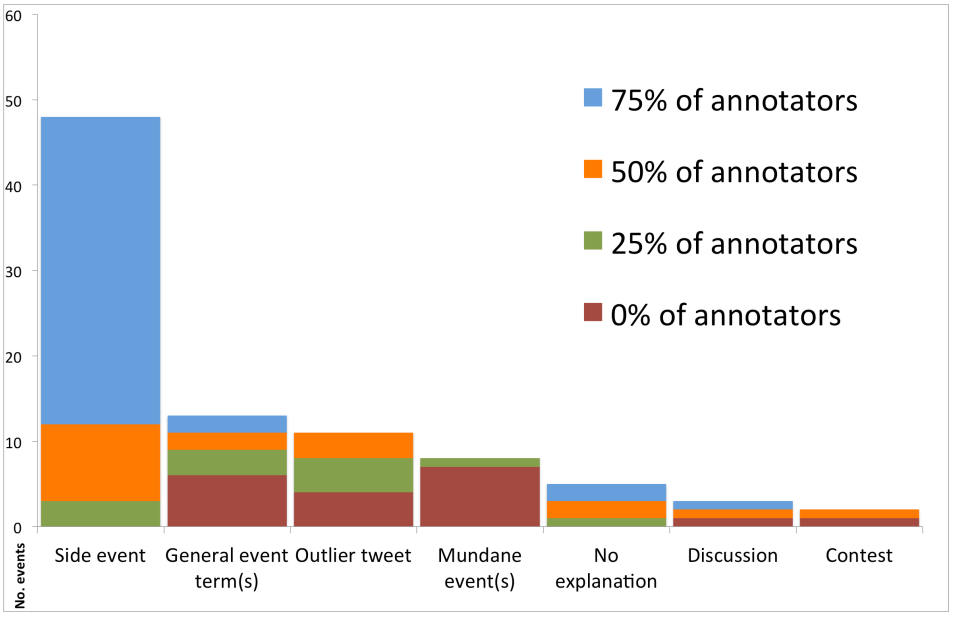


FIGURE 4.4: Overview of the properties of output units that are not rated as event by at least one annotator, divided into the percentages of annotators who rated the output as event.

We tallied the occurrences of each category and made a division by the percentage of annotators that nonetheless deemed the occurrence an event. The outcome is displayed in Figure 4.4. The bar chart shows that about half of the entries with negative annotations are due to side events being mentioned. In most cases, a majority (75%) of the annotators still judged the event cluster as a proper event. On the other hand, general event terms and mundane events are decisively not seen as event. Event tweets that include an outlier tweet (the third bar in Figure 4.4) might still be seen as event by some.

The side event as cause of not annotating an entry as event embodies the larger part of errors, but can be valued as proper events in view of the link between a side event and main event. Extra evidence for this is seen in the bulk of such entries that are coded as event by three of four annotators. On the other hand, general event terms, mundane events and to a lesser extent outlier tweets, can be seen as genuinely wrong output.

Assessment of event terms

We implemented a component in Commonness+ to add additional event terms and improve the event description. However, as is shown in Table 4.4, the annotators on average assess an event description better if no event terms were added. To analyse the cause of this outcome, we observed the terms and the

Category	Description of category	Number of occurrences	Percentage of total
Benefit	The addition of event terms leads to a better assessment	51	24%
Redundant	The addition of event terms leads to a worse assessment	65	30%
More	The addition of event terms leads to the same assessment	84	39%
Equal	No extra terms are added	14	7%

TABLE 4.6: Overview of the effect of enriching event descriptions with additional event terms as part of Commonness+, based on their assessment by the annotators. Only the 214 events of which the event terms were assessed for both Commonness and Commonness+ are included in the counts.

assessment of the Commonness and Commonness+ approaches. The four combinations that we found are displayed, along with the number of times they occurred, in Table 4.6.

The Commonness+ approach does not always result in the addition of terms, but for 93% of the events it does. In these cases, the addition of terms most frequently yields an assessment similar to the standard Commonness terms. Most striking, however, is that for 30% of the events the event terms outputted by Commonness are valued better, because the added terms include redundant information. This percentage outweighs the number of times that the addition of terms is actually beneficial for the event description (24%).

An explanation of this outcome is the way in which terms were assessed. The annotator was asked how the event terms relate to the identified event, with the options 'good', 'moderate' and 'bad'. Any redundant information might be penalised harder than a sparse set of event terms that nonetheless relate well to the event. Consider for example the terms 'Ed Sheeran' and 'gwn, Ed Sheeran, concert'. While the latter provides a richer description of the event, by including the word 'concert', the inclusion of the seemingly unrelated term 'gwn' likely leads the coders to assess them only as a moderately good representation of the event. The single event term 'Ed Sheeran', on the other hand, is assessed as a good representation.

This analysis shows that the approach to add event terms should be improved to minimise the output of redundant event terms. Apart from this, the design of the evaluation might have been of influence. We asked the annotators to assess event terms after having judged the tweet cluster to be an event. The quality of the event terms to describe the nature of the event without any prior

knowledge is not assessed. In the example given above, 'gwn, Ed Sheeran, concert' might be valued better than 'Ed Sheeran' as pointers to the type of event.

Characteristics of extracted events

In addition to analysing the tweets and event terms in relation to annotator assessments, we analysed the characteristics of the events: the number of duplicate events, the extent to which they took place in the future and whether anticipations of demonstrations were part of the output.

We kept a record of the number of duplicates and the number of clustered event terms in the top 250 output. This relates to the event clustering component (Section 3.2.2), which is aimed at diminishing the number of duplicates. In the output we found a total of 17 duplicates (6.80% of the total), while 69 event terms were clustered with other event terms. This shows that the clustering module does combine many event terms. However, the part of the top-ranked output that is redundant suggests room for improvement. A detailed evaluation of the clustering module is described in Section 4.6.2.

During the evaluation, the annotators were asked to assess whether the output was an event. It was not specified in the annotator guidelines that the tweets should refer to a future event, although this is an explicit goal of our system. For example, in Section 4.3.1 and Section 4.3.2, we describe approaches to filter tweets and events that take place in the past. As a check, we analysed the 'futureness' of the top 250 events and found that all output that was assessed as event by the annotator was actually a future event when the last tweet in the set was posted.

In Section 1 we mentioned that the functionality of our system can assist security by extracting and displaying upcoming demonstrations that might form a security risk. In the top 250 output, we found two events of this type: a demonstration against the Islamic group of ISIS in The Hague, organised by Pro Patria (a group on the right side of the political spectrum), and a demonstration during the opening of the academic year in Maastricht. As such demonstrations are sensitive to annulments, we inspected whether these two events actually took place. We found that the demonstration in The Hague, that was planned for September 20th, was canceled one day before. Although tweets in September were not fed to the system, this shows that it is important for these types of events to be able to link such an annulment to a formerly planned event.

	Number of tweets with a TIMEX found (% of total)	Tweets in common (% of found tweets)	Exclusive tweets	∼Recall
Rule-based	367,206 (1.32%)	135,565 (37%)	231,641	0.78
Heideltime	239,082 (0.86%)	135,565 (57%)	103,517	0.51

TABLE 4.7: Comparison between Heideltime and our rule-based approach (described in Section 4.3.1) to finding tweets with a TIMEX that points to a future date, from the data set described in Section 4.4.1.

4.6.2 Assessment of components

Rule based extraction of future referring time expressions

In Section 4.3.1 we stated that the Heideltime tagger (Strötgen & Gertz, 2010) fails to detect part of the informative TIMEXs in Dutch, and that it makes sense to work with manually formulated rules. To substantiate this statement, we applied the Heideltime tagger to the tweets that we used in our experiment, as described in Section 4.4.1, and compared the output of the two approaches.

We used Heideltime version 1.8.[18] We set the language to Dutch and the document type to 'news' (other options were 'narrative', 'colloquial' and 'scientific'). In line with our rule-based system, we removed retweets and focused on TIMEXs that point to a future date. Hence, any time tag in the output of Heideltime that points to a duration or a date in the past was not taken into account. Also, 'tomorrow' was excluded as was deliberately done in our rule-based system.

The performance of the two approaches is displayed in Table 4.7. The rule-based approach outperforms Heideltime in terms of the number of extracted tweets with a future referring TIMEX. Of these 367,206 tweets, 37% is also extracted by Heideltime, leaving 231,641 additional tweets only found by our rules. 103,517 tweets are only found by the Heideltime tagger. If we regard the combined output of both approaches as gold standard and take the union of the tweets that are extracted by them (totalling 470,723), the rule-based approach has a recall of 0.78 and Heideltime has a recall of 0.51. This recall should be seen as an approximation, as we do not know which of the TIMEXs the two approaches failed to retrieve, or which of the extracted TIMEXs are correct.

An analysis of the exclusive tweets that are extracted by both approaches shows that the rule-based approach succeeds in extracting TIMEXs that explicitly mention a specified amount of days in the future, like 'nog 12 nachtjes slapen'

[18] `https://code.google.com/p/heideltime/wiki/Downloads`

	Precision	Recall	$F_{\beta=1}$
commonness	0.50	0.87	0.63 ± 0.02
Frog NED	0.37	0.82	0.51 ± 0.01

TABLE 4.8: Significance estimates of commonness and Frog NED on retrieving entities from an annotated sample of 1,000 tweets. Scores are obtained after bootstrap resampling with 250 samples.

('12 nights of sleep to go'). On the other hand, most tweets that were extracted exclusively by the Heideltime tagger contain TIMEXs like 'volgende week' ('next week') and 'dit weekend' ('this weekend'). A disadvantage of such phrases is that it is hard to link them to a specific date.

Extraction of entities

For the extraction of entities from tweets, we have applied the commonness approach as described by Meij et al. (2012), which does not rely on common Named Entity Detection (NED) markers such as part-of-speech tags or capitalisation. To obtain an impression of its performance, we annotated 1,000 tweets, sampled from the data set described in Section 4.4.1, by their named entities and compared the performance of commonness to a competing NED system for Dutch, the NED component in Frog (van den Bosch et al., 2007).[19] We converted the output of both approaches for these sentences, as well as the annotated sentences, into the IOB-tagging format, and evaluated them with the CoNLL-2000 shared task evaluation script.[20] For commonness, possible overlap between output was resolved (Section 4.3.3).

We estimated significance in the differences between the commonness method and Frog's NED by using bootstrap resampling (Noreen, 1989). Per system we selected 250 random samples of sentences. We assume that performance A is significantly different from performance B if A is not within the center 90% of the distribution of B. Results are presented in Table 4.8.

The Frog NED system is outperformed by the commonness approach both in recall and precision, with a resulting F1 score of 0.63 for commonness against 0.51 for Frog NED. The difference is significant with small standard deviations of 0.02 and 0.01, respectively. This shows that off-the-shelf tools are lacking generalisation power when applied to non-standard language. Applying the commonness approach in such a setting has proven an effective replacement.

[19] https://languagemachines.github.io/frog/

[20] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

	Precision	Recall	$F_{\beta=1}$	RI	#Total	#Merged	#Correct
After clustering	0.84	0.35	0.49	0.97	2370	978	822

TABLE 4.9: Performance of the clustering component (described in Section 4.3.2). RI = Rand Index. #Total = the term pairs that should be merged according to a manual gold standard clustering. #Merged = the merges made by the clustering component. #Correct = the correctly merged pairs.

Event clustering

Event clustering is an important component in our system, aimed at reducing duplicate event output and enhancing event descriptions. In Section 4.6.1 we report on 69 clusterings of event terms and 17 duplicate events in the top 250 generated events. As these numbers do not give a complete overview of the clustering performance, we also evaluated all clusterings that were made on the initial set of 2,500 date-term pairs.

We manually made clusters of the 2,500 date–term pairs, by inspecting the tweets in which each event term is mentioned, and compared the automatic clusterings to this reference set. We evaluated performance by inspecting the *pairs* of event terms that were clustered (Halkidi, Batistakis, & Vazirgiannis, 2001). Any pair that was clustered in the reference set but was not clustered by the clustering component was added to the false negatives, while any pair that was clustered by the clustering component but should not have been clustered was added to the false positives. We used these numbers to assess precision, recall and $F_{\beta=1}$. We also calculated the Rand Index (Rand, 1971), an accuracy metric that not only takes into account objects that are classified in the same cluster, but also rewards objects that are rightfully not clustered together (the true negatives).

The cluster performance is presented in Table 4.9. The optimal clustering would result in 2,370 merges of event term pairs. The clustering component actually makes 822 correct merges, and incorrectly merges 156 pairs. This results in a precision of 0.84 and a recall of 0.35. The high value of the Rand Index, 0.97, is due to the large number of true negatives. These results show that the clustering component manages to merge part of the duplicates, at a fairly low rate of false positives. Nonetheless, this component leaves room for improvement, especially with regard to the recall performance.

4.7 Conclusion and Discussion

We proposed a system for open-domain event extraction. An adaptation of the work by Ritter et al. (2012), it operates in a more unsupervised way and can be implemented relatively easily for any language, provided that a rule set is written for detecting future time references. Central to the system is the extraction of *event terms*, for which we apply the Commonness approach (Meij et al., 2012). Additional event terms are added based on the $tf * idf$ of words in the event tweets.

Where Ritter et al. (2012) assessed the outcomes of the system themselves, we asked human annotators to evaluate our system in two variants, and a third baseline system. Of the top 250 output of our system, 87% was assessed by at least two of four human annotators as representing an event. All four annotators assessed 63% as event, markedly outperforming a baseline based on word n-grams, which yielded a precision of 42%. This performance appears comparable to Ritter et al. (2012), who report a precision at 100 of 0.90, and precision values of 0.66 at rank 500 and 0.52 at rank 1,000.

The addition of event terms does not appear to improve the event description, as seen from the average score, between 'moderate' and 'good', of 2.63 in comparison to 2.69 when no event terms are added. Our analysis confirms that the addition of event terms often produces redundant terms, which the annotators penalise more than missing terms.

A recall evaluation based on six types of gold standard events reveals that the system is able to capture any of the event types at an overall recall of 0.40 on events that are tweeted about more than four times. While the system relies on tweet mentions of an event, the recall could be improved by identifying more TIMEXs and entities.

This study shows that many events can be discovered by using time references in tweets, yielding a useful precision and recall. Apart from their date and most important descriptors, however, little is known about the events themselves. The following chapter describes a study to automatically enrich the event description with an additional characteristic: periodicity.

Chapter 5

Automatically Identifying Periodic Social Events from Twitter

Based on: Kunneman, F. & van den Bosch, A. (2015). Automatically Identifying Periodic Social Events from Twitter. In G. Angelova, K. Bontcheva & R. Mitkov (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing 2015* (pp. 320–328), Hissar, Bulgaria

Many events referred to on Twitter are of a periodic nature, characterised by roughly constant time intervals in between occurrences. Examples are annual music festivals, weekly television programs, and the full moon cycle. Our study on event detection in Chapter 4 enables us to study the detection of such periodic events from a longitudinal window of Twitter posts. In this chapter, we describe our study on applying a timeline-based and a calendar-based approach to perform this task.

5.1 Introduction

As a popular communication channel for sharing news, experiences, and intentions, Twitter has been found to provide an accurate reflection of many aspects of the real world (Bollen, Mao, & Pepe, 2011; S. Zhao et al., 2011). For example, the periodicity of daily life can be exposed by visualising the frequency of hashtags such as '#breakfast' and '#goodmorning' (Preoţiuc-Pietro & Cohn, 2013). In addition, real-world events can be automatically detected by signalling a sudden rise and fall of word occurrences in tweets (Petrović et al., 2010; McMinn et al., 2013). We propose a system that can identify *periodic* events from Twitter: provided with a continuous stream of raw tweets, it returns an overview of periodic social events.

Surprisingly, this topic of periodicity has not yet been studied in the context of events mentioned on Twitter. The identification of periodically recurring events has obvious gains for a system that detects events in the Twitter stream. For example, events that are found to recur periodically can be automatically linked together and their tweets can be analysed for similarities, differences and trends, such as participant appreciation by edition. Furthermore, detected periodic patterns can be used to predict future events before they are referred to on Twitter. Finally, the periodicity of an event can be included as feature for several tasks, such event type classification.

The rich set of references to the real world made on Twitter make it a suitable platform to mine for periodic patterns in relation to events of any type. At the same time, the non-standard language and high volume of streaming messages make it a challenging task. We facilitate this task by applying an event detection approach that identifies terms that might represent a social event, and that relates them to a frequently and explicitly mentioned date of the event. After this first event detection stage, periodicity detection can be applied to the clean date sequences that are linked to event terms.

In this Chapter, we propose two approaches to detect periodicity from a longitudinal set of detected events. We first describe related work on the detection of periodicity and periodicity in social media. Next, we describe our approach to detect events from an extended period of tweets and introduce the two approaches to search for periodic patterns in the resulting set. We subsequently present our evaluation of these approaches, as well as the results. We then perform an analysis of the output and the utility of periodicity detection in an event prediction task, and end with the conclusions.

5.2 Related Work

Finding periodic patterns is a valuable task in many contexts of sequential data, such as DNA or protein sequences (Zhang, Kao, Cheung, & Yip, 2007), market basket data (Mahanta, Mazarbhuiya, & Baruah, 2008) and complex signals such as sound (Sethares & Staley, 1999). Elfeky, Aref, and Elmagarmid (2005) distinguish between 'segment periodicity' and 'symbol periodicity'. The first refers to the repetition of a specific sequence, while the second refers to single symbols in a sequence that recur at roughly constant time intervals. The latter is what we aim to detect.

Several patterns of periodicity have been analysed in social media. Chu, Gianvecchio, Wang, and Jajodia (2012) aim to distinguish bots from human user accounts on Twitter, and find that the periodicity of tweet postings is a strong indicator to recognise bots. They estimate periodicity by the entropy rate of post intervals, where a low entropy points to a non-random, periodic pattern. Chu, Widjaja, and Wang (2012) adopt this entropy-based periodicity feature to help distinguish spam campaigns from proper campaigns on Twitter. Fan, Zhao, Feng, and Xu (2014) analyse temporal patterns in topics discussed on Weibo, and find that the topic 'business' displays a highly periodic pattern. Yang, Lee, and Yan (2013) aim to classify Twitter users in predefined categories. They find that the periodicity pattern of words linked to a category is a strong indicator, as users tend to mention their topic of interest at similar times of the day and week. At the word level, Preoţiuc-Pietro and Cohn (2013) apply Gaussian processes to model the periodicity of hashtag mentions. They use this information to predict hashtag frequencies at any hour.

Other than the periodicity of words, topics and tweet postings, which is mainly characterised by regular time intervals, social events typically follow a calendar pattern. In this sense, our aims mostly compare to the work of Mahanta et al. (2008), who apply periodicity detection on market basket data with the aim to find co-occurrence patterns with calendar periodicity. To this end, they link basket data to hierarchical time stamps (year, month and day) and look for patterns that re-occur on a yearly or monthly basis. We build on this work in one of our approaches.

5.3 Approach

We apply periodicity detection in two stages. In the first stage, event detection is applied on all tweets that are available in TwiNL, based on the approach described in Chapter 4. In the second stage, we experiment with two approaches to find periodicity from date sequences of equal event terms that result from the first stage.

5.3.1 Event detection

For a description of our approach to event detection, we refer the reader to Chapter 4. Crucially, this approach is based on a *fixed* set of tweets that are posted within a month. When a longer duration of tweets is used, references to recurring events might overlap and performance will drop. To apply this approach to a set of tweets that spans multiple years, we employ a daily sliding window that spans a month of tweets from which events are detected. After each new window, the top 2,500 ranked events are compared to the existing output. New and existing events of which 10% of tweets overlap are merged, while the remaining new events are added to the existing output. This results in a large set of events, while avoiding a surplus of duplicate output. Periodicity detection is applied on this set.

5.3.2 Periodicity detection

The event detection procedure results in a set of events represented as one or more event terms linked to a date. Next, periodic events can be found by scanning for events that are linked to at least three dates, the minimum indication of periodicity.

We compare two approaches to finding periodic patterns from the date sequence related to an event term: a timeline-based approach and a calendar-based approach. We refer to them as 'PerTime' and 'PerCal'.

PerTime

PerTime leverages the intervals between sequences of at least three dates. Any date sequence with roughly similar intervals is seen as periodic, where the intervals are measured at the level of days. We estimate the similarity by computing the relative standard deviation over the intervals, RSD:

$$\text{RSD} = \frac{s}{\bar{x}} \times 100\% \tag{5.1}$$

The RSD relates the average interval in days $\bar{x}$ to the standard deviation s, returning the standard deviation as percentage of the average values in a set. The RSD is a sensible approach to scoring the periodicity of date intervals. In contrast to regular standard deviation, a deviation of lengthy intervals, such as 365 days, is less penalised than deviation of smaller intervals, such as seven days. We set the minimum interval length to six days, ensuring weekly events as the minimal periodicity. A week is a likely minimum of periodic social events, and imposing this minimum reduces computational costs considerably.

An alternative approach to calculating the RSD is to calculate the entropy, as was done by Chu, Gianvecchio, et al. (2012) to score the periodicity of Twitter posts and thereby distinguishing bot accounts from human Twitter users. A disadvantage of this approach is that it requires a binning of the sequence data into categories, where it is not clear beforehand which is the optimal binning.

PerCal

Rather than looking for regular intervals between dates, PerCal searches for *similarities* between the dates in a sequence. For example, if the event term 'Christmas Day' is mostly linked to the date of '25 December', it is likely a periodic event. Likewise, an event term might recur with 'the third Saturday of May' or 'the Wednesday of week 44'. The calendar-based approach scans a date sequence for such repetitions.

The detection of calendar periodicity has mainly been the focus in studies that aim to find periodic transactional patterns (Y. Li, Wang, & Jajodia, 2001; Mahanta et al., 2008). Y. Li et al. (2001) propose an intuitive calendar scheme to describe a periodic pattern. The pattern has the form of ⟨year,month,day⟩. Any of these fields can be filled with a specific value, while the '*'-character is used to denote 'every'. For example, the pattern ⟨*,2,1⟩ represents 'every year on the 1st of February', while ⟨2011,*,12⟩ denotes 'every twelfth day of the month in 2011'. We adopt this pattern scheme, and extend it with the additional fields week, weekday, and #weekday (the index of a given weekday within a month, such as 'the second Thursday'). We add the '-' character as a possible value, to account for fields that are irrelevant to a pattern. As an additional extension, we allow the model to describe patterns like 'every six months' or 'every two years', by specifying a step size that relates to the field that is described by 'every'. For example, ⟨*2,1,-,-,Sunday,2⟩ denotes 'every two years on the second Sunday of January', and ⟨2011,*,-,1,-,-⟩ denotes 'every first day of the month in 2011'.

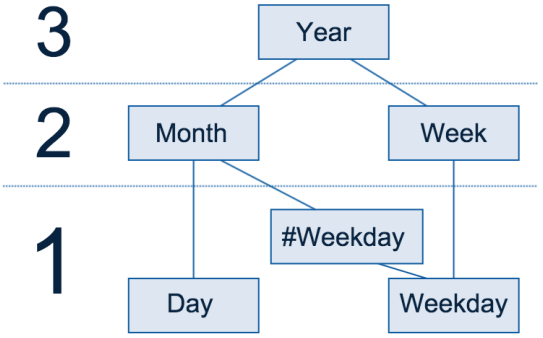


FIGURE 5.1: Diagram of included calendar fields for the PerCal approach and their relation on three levels.

The relationship between the included calendar fields is illustrated in Figure 5.1. The scheme has three levels of granularity. On the first level are 'day' (1–31), 'weekday' (Monday–Sunday) and '#weekday' (1–5). The 'day' field relates to 'month' (1–12) at the second level; any combination between the two values can be made. '#weekday' has a connection to both 'weekday' and 'month', and represents the index of a weekday in a month (for example: the *third* Wednesday of October). Finally, 'weekday' connects directly to 'week' (1–53), which enables relations like 'every Wednesday' or 'Monday on week 40'. At the top level is the 'year' field, so as to describe yearly patterns or patterns during a specific year.

A periodic calendar pattern can be detected by ascending the hierarchy of calendar fields and scanning the date sequence of an event term for regularities. Similar to PerTime, weekly periodicity is the smallest pattern that we search for. Starting from the lower-level fields ('day', 'weekday' and the 'weekday-#weekday' combination), the PerCal algorithm scans whether any of the values of these fields occurs three times or more (the minimum requirement for a periodic pattern). If this requirement is met, the dates that include this value are selected and passed on to the higher level: 'month' (if the day or the weekday-#weekday combination is periodic) or 'week' (if the weekday is periodic). Because the patterns we look for can describe either a sequence on this second level (like 'every two months' or 'every week') or a sequence of years on the third level, we scan both for a sequence and a repetition of the month or the week values on this second level. If a sequence is found, the pattern is finalised. If a repetition is found, the algorithm proceeds to find a yearly pattern.

A sequence of weeks, months or years might have steps of unequal size. In such a case, we describe the pattern with the smallest step size found. Any date between larger steps is denoted as a missing date. The sequence '04/03/14 – 04/04/14 – 04/06/14' has monthly periodicity with step size '1' and a missing

date '04/05/14'.

Some patterns show stronger periodicity than others. As mentioned above, a sequence might contain missing dates, decreasing the evidence for periodicity. In addition, not all dates that link to an event term may combine into a pattern. Following Y. Li et al. (2001), we quantify these two inconsistencies as *confidence* and *support* estimates. Confidence is estimated by dividing the dates that *could* fill in a pattern (from the first date to the last) by the number of dates that are *actually* seen. Support is the percentage of all dates that satisfy the pattern. To obtain an overall score of the quality of a pattern, we additionally calculate the average of these two metrics.

PerCal searches for periodic patterns at different levels. As a result, it may find multiple patterns in the same date sequence. As many social events are likely to only display one periodic pattern,[1] we choose to disallow overlapping patterns of the same event term. If two patterns overlap, the one with the highest overall score is selected.

Periodic term clustering

By applying periodicity detection on separate event terms, the output of both PerTime and PerCal might contain duplicate periodic events that are described by different terms. To de-duplicate the output, we cluster event terms with a periodic sequence together. For both approaches, we aggregate all tweets linked to the periodic pattern of an event term, in order to form big documents. Any pair of terms with 90% overlapping dates for PerTime and any pair with a similar pattern for PerCal were tested as clusters. Clustering was applied in the same fashion as described in Chapter 4, Section 4.3.2. The threshold for clustering was set to a cosine similarity above 0.5.

5.4 Experimental Set-up

5.4.1 Data

We tested our system on all Dutch tweets that were collected from the Tweet IDs in TwiNL, from December 16th 2010 up to February 16th 2015, amounting to 2.73 billion tweets in total. After processing these tweets, 14,896,146 were found to have a matching time expression. 9,937,596 of these contained one or more entities or hashtags, 2.15 on average and 21,347,777 in total.

[1] An exception are events that show bounded periods of periodicity, such as seasons of a television show. We discuss such occurrences in Section 5.6.

5.4.2 Procedure

We applied event detection on the span of tweets as specified in Section 5.3.1, with a sliding window of a month and a daily sliding frequency. Events were merged if they were extracted from (partly) the same tweet IDs. Based on this procedure, a calendar was filled with 94,526 events.[2]

Periodicity detection was applied to single event terms; we kept a log of the dates linked to each term. For both PerTime and PerCal, we searched for periodic patterns in each date sequence that linked to an event term, provided that the sequence comprised a minimum of three dates. We clustered terms with a similar periodic pattern after all events were processed.

5.4.3 Evaluation

We ranked the periodic event patterns returned by the two approaches by their respective metrics to score periodicity: RSD for PerTime and the average value of support and coverage for PerCal. One of the authors manually assessed the top 500 patterns from both rankings, deciding for each output whether it represented a regularly recurring sequence of events, rather than events or event terms that share a coincidental temporal regularity. The terms, dates, and tweets linked to each output, and if needed the Google search engine, were consulted to guide this decision.

In order to acquire a sense of agreement for the annotations, the second author annotated the top 200 events of the two systems. The mutual F-score of positive annotations was 0.92 for the PerTime output and 0.93 for the PerCal output.

5.5 Results

PerTime assigned a periodicity score to 5,301 events out of the total of 94,526 events. PerCal found 7,018 periodic patterns.[3] The precision and recall of their top 500 output are presented in Table 5.1. 315 correct periodic events were confirmed from the output of PerTime, and 379 from the output of PerCal, resulting in precision-at-500 scores of 0.63 and 0.76, respectively. We approximated a recall score by comparing the periodic event terms that were found by both approaches (637 in total), and calculating which percentage of these was returned

[2] A dataset with the date, event terms, event score(s) and tweet ID's of all 94,526 events is accessible from `http://dx.doi.org/10.17026/dans-xn5-wq59`.

[3] The periodic event patterns that were found by both systems, along with the annotations of both raters, is made publicly available from `http://dx.doi.org/10.17026/dans-xn5-wq59`.

	Precision	∼Recall
PerTime	0.63	0.52
PerCal	0.76	0.69

TABLE 5.1: Periodicity detection quality based on a manual evaluation of the top 500 detected periodic events by PerTime and PerCal.

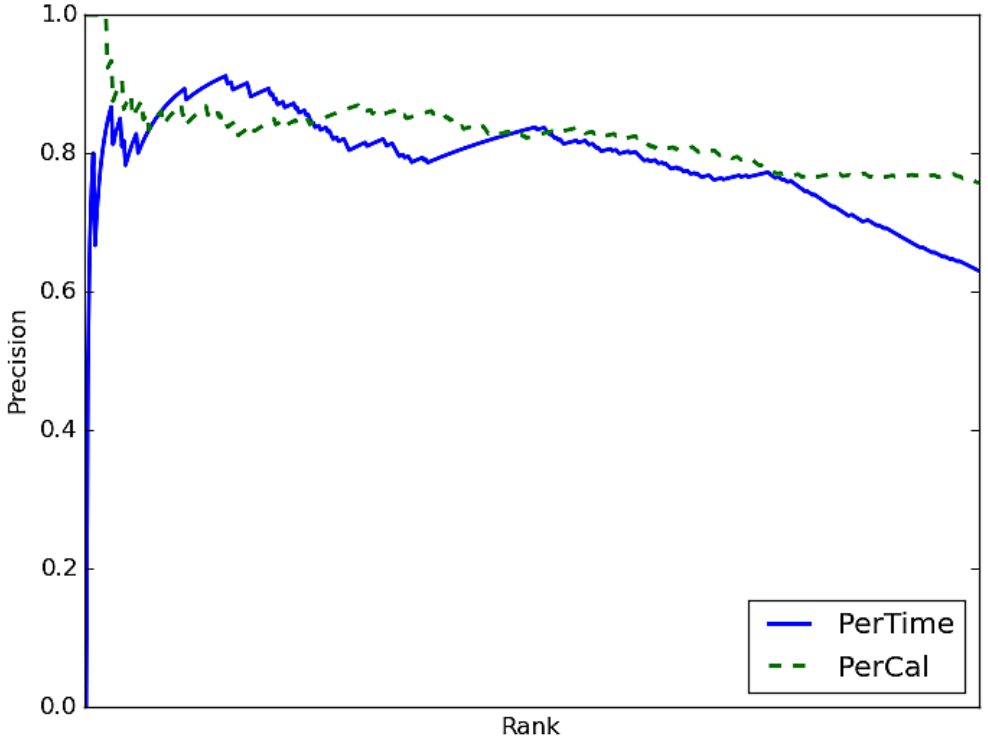


FIGURE 5.2: Precision-at-curves of the top 500 detected periodic events by PerTime and PerCal.

by either of them. The recall scores are lower than the precision scores, due to an overlap of only 116 events (18%) between PerTime and PerCal.

Precision-at-curves of the top 500 rankings are given in Figure 5.2. For PerTime, the RSD at rank 500 is 10.2 days. A perfect RSD score of 0.0 is maintained up to rank 81. We randomly shuffled the ranks of events with equal scores. The curve shows a progressive decay towards the end. The temporally increasing precision at rank 200 is due to the detection of a number of periodic events that are characterised by changing intervals, such as Easter and Pentecost. As these events are mostly related to the lunisolar calendar, they share the same non-perfect RSD score and cluster around the same rankings. For PerCal, the perfect score of 1.00 decreases from rank 10. In contrast to PerTime, precision is decreasing at a slower rate, with little decay up to rank 500. The pattern score of 0.76 at this rank is thus still fairly reliable.

5.6 Analysis

5.6.1 Error analysis

The errors that were found in the top 500 rankings concerned a flaw in the event detection procedure or a coincidental periodic sequence of events that are described by the same terms. An example of the former errors are terms like 'emo' for valentines day, or 'firecrackers' for new years eve. They re-occur as event term along with these celebrations, but do not actually describe a periodic event. A common example of a coincidental periodic sequence is a legislative event, such as an increase in excise taxes, that happened to re-occur by repeating periods of six months.

The clustering procedure, meant to reduce the amount of duplicates, led to 16 successful merges in the top 500 PerTime output, leaving 23 duplicates. A lot more merges were made of the PerCal output, at 71 successful merges, albeit against 4 incorrect ones, and 22 duplicates. A possible reason for the difference between the number of merges of the output of both approaches, is that PerCal identifies a considerably higher number of periodic patterns, which might result in more cluster candidates.

5.6.2 Output of PerTime and PerCal

PerTime and PerCal mainly return periodic events that recur on a yearly basis. The positively evaluated periodic patterns in the top 500 output of PerTime have an average interval length of 360.9 days, with a standard deviation of 45.7. The mode and median are both 364 days, in 119 out of 319 events. Likewise, in the output of PerCal only five weekly sequences and three monthly sequences are seen. The yearly sequences are divided into 164 periodic patterns of type '#weekday', 111 date repetitions, and 96 patterns with a recurring weeknumber.

A reason for this bias might be that more people tweet about yearly events, which are more significant than weekly or monthly events. As a consequence of its ranking strategy, the event detection approach does not pick up on part of the weekly events. Another reason for the bias is that PerTime and PerCal both focus on the total date sequence of an event term. This makes it hard to find for example one season of a television show as a periodic sequence. A way to enable the detection of more weekly sequences is hence to apply a segmentation step before periodicity detection, indicating the start and the end of the periodicity. For example, a television programme might re-occur on a weekly basis within a period of three months, followed by a long period without this event until a

	Event term(s)	Dates (dd/mm/yy)	Timeline pattern	Calendar pattern
Periodic events approaches	#kicd, #keepitcleanday	21/09/12, 20/09/13, 19/09/14	364 - 364	⟨*,9,-,-,Friday,3⟩
	#trendrede	13/09/11, 11/09/12, 10/09/13, 09/09/14	364 -364 - 364	⟨*,9,-,-,Tuesday,2⟩
	#valentinesday	14/02/13, 14/02/14, 14/02/15	365 - 365	⟨*,2,-,14,-,-⟩
Periodic events only found by PerTime	romantische muziek	14/08/11, 12/08/12, 25/08/13, 24/08/14	364 - 378 - 364	-
	paaszondag	24/04/11, 08/04/12, 31/03/13, 20/04/14, 05/04/15	350 - 357 - 385 - 350	-
	musical sing-a-long	28/08/11, 26/08/12, 01/09/13, 31/08/14	364 - 371 - 364	-
Periodic events only found by PerCal	#7hloop	20/11/11, 18/11/12, 16/11/14	364 - 728	⟨*,11,-,-,Sunday,3⟩
	fortarock	02/07/11, 02/06/12, 09/11/12 01/06/13, 31/05/14, 06/06/15	336 - 160 - 204 - 364 - 371	⟨*,-,22,-,Saturday,-⟩
	#tuinvogeltelling	23/01/11, 19/01/13, 20/01/13, 19/01/14, 18/01/15	727 - 1 - 364 - 364	⟨*,1,-,-,Sunday,3⟩

Table 5.2: Examples of periodic events in the top 500 output of PerTime and PerCal.

new season starts. If such dense periods could be singled out, the periodicity in this period can be disclosed.

Examples of detected periodic events are given in Table 5.2. To illustrate the complementary strength of both approaches, we made a distinction between events that are only found by one of them and by both. Examples of periodic events found by both approaches are '#trendrede' (an annual speech by the regent of the Netherlands), '#valentinesday' and the event described by the hashtags '#kicd' and '#keepitcleanday'. Events like this, linked to a fixed date, are characterised by equal yearly intervals (only allowing for a minor deviation of 366 instead of 365 days in leap years).

The events 'romantische muziek' (referring to the 'Day of Romantic Music') and 'musical sing-a-long' are not found by PerCal, which is due to an inconsistent pattern of dates. PerTime can typically deal with such small inconsistencies. The event described by 'paaszondag' ('Easter Sunday') follows the lunisolar calendar, while the calendar approach follows a Gregorian calendar scheme[4]. Again, PerTime only penalises the inconsistencies in day intervals, without discarding the event altogether.

While PerTime can deal with inconsistencies in the intervals between dates, PerCal displays a higher tolerance towards missing dates. An example is '#7hloop' (a running event in The Netherlands), which was not found by the event extraction module in 2013. The resulting interval of 728 days (two years) at this point

[4]To find events like Easter, the framework of PerCal could be extended by including a lunisolar scheme or other existing schemes.

	50	100	200	400	800
PerTime	0.42	0.34	0.19	0.17	0.09
PerCal	0.64	0.65	0.63	0.56	0.44

TABLE 5.3: Accuracy of event prediction based on periodic patterns, by rank of pattern certainty.

results in a poor periodicity score for PerTime. PerCal, having detected the overall pattern, gives a smaller penalty for the missing entry in 2013. The support for these days is 1.0, while the confidence is 0.75, leading to an overall score of 0.88. The same applies to the '#tuinvogeltelling' event (a yearly count of garden birds), which is absent in 2012. Similarly, noisy date sequences in which only part of the dates form a periodic pattern can only be dealt with by PerCal. PerTime assigns a low overall periodicity score to the date sequence associated with 'fortarock' (a music festival in The Netherlands), due to the irregular intervals.

5.6.3 Event prediction

Strong periodic patterns, once detected, can be extrapolated to put future events on the calendar. We put this to the test by processing all the events that were detected until March 2014 and starting to detect periodic patterns from January 2013. Based on the found patterns, we made predictions up to March 2015 (the end of our event set). We could evaluate the quality of these predictions by comparing them against the events that were actually detected by our system in this period. For PerTime, a future event date was predicted by adding the number of days of the most frequently seen interval to the last seen date. For PerCal, the calendar pattern was simply continued. For example, ⟨*,1,-,-,Wednesday,3⟩ was continued with 'the third Wednesday of January 2015'.

Predictions were evaluated by checking whether the event terms in the predicted pattern were actually seen at the date of prediction in the set of detected events (which, as evaluated in Chapter 4, is the result of an automatic procedure that is about 87% correct according to the most relaxed majority vote). One can expect a better periodicity score to lead to a better prediction. Hence, we sorted the patterns seen on March 2014 by their score, and checked the extent to which the prediction was correct.

The results are given in Table 5.3. The precision values are given at increasingly higher ranks. 64% of the predictions of the top 50 PerCal patterns are confirmed by the later detected events, and a roughly similar score is maintained up to rank 200. In comparison, the precision of PerTime predictions is poor at

0.42 in the top 50 events, and decreases drastically further down the rankings. The PerCal approach is hence the most suitable to put future events on the calendar. It can be expected that its performance increases as sequences of event terms span a longer period of time, providing more recurrances to rely on.

5.7 Conclusion and Discussion

As far as we know, this is the first work that deals with the task of periodic event detection from Twitter data, which serves to extract long-range patterns from Twitter, detect periodic events among those patterns, and predict events before they are mentioned on Twitter. Applying the procedure to over four years of Dutch tweets, a timeline-based and calendar-based approach to periodicity detection yield a precision-at-500 of 0.63 and 0.76, respectively. The calendar-based approach maintains a consistent performance throughout the rankings, and yields an estimated precision of up to 0.65 when the periodic patterns are applied to predict future events.

In addition to periodicity, a valuable characterisation of Twitter events is the emotion in tweets that refer to it. The next chapters describe studies to model the language of figurative speech and emotions based on hashtags, which ultimately culminates into the detection of the most anticipointing events in Chapter 8.

Part II

Hashtag-based patterns

> Considering that the brain itself stays connected only by constant conversation, it's hard to argue that our connections to others belong strictly on a lower tier. What makes the transmissions passing through the corpus callosum all that different from the transmissions passing through the air, from mouth to mouth?
> Brian Christian - The most human human

Chapter 6

Signalling sarcasm: From hyperbole to hashtag

Based on: Kunneman, F., Liebrecht, C., Van Mulken, M. & van den Bosch, A. (2015), Signalling sarcasm: From hyperbole to hashtag, *Information Processing & Management, 51(4)*, 500–509

To avoid a sarcastic message being understood in its unintended literal meaning, in Twitter messages sarcasm is often explicitly marked with a hashtag such as '#sarcasm'. We collected a training corpus of about 406 thousand tweets with hashtag synonyms denoting sarcasm from TwiNL. Assuming that the human labeling is correct (annotation of a sample indicates that about 90% of these tweets are indeed sarcastic), we train a machine learning classifier on the harvested examples, and apply it to a sample of a day's stream of 2.25 million Dutch tweets. Of the 353 explicitly marked tweets on this day, we detect 309 (87%) with the hashtag removed. We annotate the top of the ranked list of tweets most likely to be sarcastic that do not have the explicit hashtag. 35% of the top 250 ranked tweets are indeed sarcastic. Analysis indicates that the use of hashtags reduces the further use of linguistic markers for signalling sarcasm, such as exclamations and intensifiers. We hypothesise that explicit markers such as hashtags are the digital extralinguistic equivalent of non-verbal expressions that people employ in live interaction when conveying sarcasm. Checking the consistency of our finding in a language from another language family, we observe that in French the hashtag '#sarcasme' has a similar polarity switching function, be it to a lesser extent.

6.1 Introduction

In the general area of sentiment analysis, sarcasm is a disruptive factor that causes the polarity of a message to flip. Unlike a simple negation, a sarcastic message often conveys a negative opinion using only positive words – or even intensified, hyperbolic positive words. Likewise, but less frequently, sarcasm can flip the polarity of an opinion with negative words to the intended positive meaning. The detection of sarcasm is therefore important, if not crucial, for the development and refinement of sentiment analysis systems, but is at the same time a serious conceptual and technical challenge.

In this chapter we introduce a sarcasm detection system for tweets. In doing this we are helped by the fact that sarcasm appears to be a commonly recognised concept by many Twitter users, who explicitly mark their sarcastic messages by using hashtags such as '#sarcasm' or '#not'. Hashtags in tweets are explicitly marked keywords, and often act as categorical labels or metadata in addition to the body text of the tweet (Chang, 2010). By using the explicit hashtag, any remaining doubt a reader may have is taken away: the message is not to be taken literally; it is sarcastic.

While such hashtags primarily function as conversational markers of sarcasm, they can be leveraged as annotation labels in order to generate a model of sarcastic tweets from the text that co-occurs with these hashtags. A clear advantage of this approach is the easy acquisition of a vast amount of training data. On the other hand, its performance is dependent on the correctness of two assumptions: first that users who include one of the selected hashtags in their tweet actually intended to convey sarcasm and indeed intended to flip the polarity of the message, and second that the pattern of sarcasm in a tweet still holds when the hashtag is excluded from it as a training label. We set out to test these assumptions along with the quality of the resulting sarcasm detection system by applying it on a realistically large and unbiased sample of tweets (of which the vast majority is non-sarcastic) posted on the same day.

The hashtag as a marker of sarcasm has been leveraged in previous research to detect sarcasm in tweets (Reyes, Rosso, & Veale, 2013; González-Ibáñez, Muresan, & Wacholder, 2011). One contribution of this paper to the existing body of work is that a sarcasm classifier is trained on several markers of sarcasm in tandem, the most frequent being '#not', and performance is assessed on a realistically large and unbiased sample of tweets. Furthermore, we provide insight into the role of hyperbole in sarcastic tweets, and we perform a cross-lingual comparison of the use of sarcasm in Twitter by annotating French tweets ending with

'#sarcasme'.

6.1.1 Definitions

Twitter members mark their sarcastic messages with different hashtags. As described in more detail in Section 6.3.1, we find that four words tend to be used as hashmarks in sarcastic posts: '#sarcasm', '#irony', '#cynicism' and '#not'. Although sarcasm, irony and cynicism are not synonymous, they have much in common. This is especially true for sarcasm and irony; many researchers treat those phenomena as strongly related (Attardo, 2007; Brown, 1980; Gibbs & O'Brien, 1991; R. J. Kreuz & Roberts, 1993; Muecke, 1969; Mizzau, 1984), and sometimes even equate the terms in their studies in order to work with a usable definition (Grice, 1978; Tsur, Davidov, & Rappoport, 2010). Cynicism is more mocking and tells us more about human beliefs than irony and sarcasm (Eisinger, 2000), but there is a close correlation between these concepts (Yoos, 1985). The hashtag '#not' is not the name of a rhetorical device or trope such as sarcasm, irony and cynicism, but it is a conventionally used meta-communication marker to indicate that the message contains a shift in evaluative valence.

In psycholinguistics and cognitive linguistics sarcasm has been widely studied, often in relation with concepts such as cynicism, and with verbal irony as a broader category term. A brief overview of definitions, hypotheses and findings from communication studies regarding sarcasm and related concepts may help clarify what the hashtags convey.

In this study, we are interested in sarcasm as a linguistic phenomenon, and how we can detect it in social media messages. Yet, Brown (1980) warns that sarcasm 'is not a discrete logical or linguistic phenomenon' (p. 111), while verbal irony is. Indeed, Reyes et al. (2013) see sarcasm 'as specific extension[s] of a general concept of irony' (p. 755). In line with the extensive use of #sarcasm in tweets to mark verbal irony, we take the liberty of using the term sarcasm while verbal irony would be the more appropriate term. Even then, according to Gibbs and Colston (2007) the definition of verbal irony is still a 'problem that surfaces in the irony literature' (p. 584).

There are many different theoretical approaches to verbal irony. It should (a) be evaluative, (b) be based on incongruence of the ironic utterance with the co-text or context, (c) be based on a reversal of valence between the literal and intended meaning, (d) be aimed at some target, and (e) be relevant to the communicative situation in some way (Burgers et al., 2011). Although it is known that irony is always directed at someone or something (the sender himself, the

addressee, a third party, or a combination of the three, see Burgers et al. (2011); Livnat (2004)) and irony is used relatively often in dialogic interaction (Gibbs, 2007), these two elements of irony are hardly examinable in the case of Twitter: the context of the Twitter messages is missing and it is inconvenient to investigate interaction. Therefore, it is hard to interpret the communicative situation and the target of the message. However, it is possible to analyse texts, such as tweets, on their evaluative meaning and a potential valence shift in the same way as Burgers et al. (2011) did. They define verbal irony as 'an utterance with a literal evaluation that is implicitly contrary to its intended evaluation.' (p. 190).

Thus, a sarcastic utterance involves a shift in evaluative valence, which can go two ways: it could be a shift from a literally positive to an intended negative meaning, or a shift from a literally negative to an intended positive evaluation. Since Reyes et al. (2013) argue that users of social media often use irony in utterances that involve a shift in evaluative valence, we use the definition of verbal irony of Burgers et al. (2011) in this study on sarcasm, and we use both terms synonymously. The definition of irony as saying the opposite of what is meant is commonly used in previous corpus-analytic studies, and is reported to be reliable (R. Kreuz, Roberts, Johnson, & Bertus, 1996; Leigh, 1994; Srinarawat, 2005).

In order to ensure that the addressees detect the sarcasm in the utterance, senders use markers in their utterances. Attardo (2000) states that those markers are clues a writer can give that 'alert a reader to the fact that a sentence is ironical' (p. 7). The use of markers in written and spoken interaction may be different (Jahandarie, 1999). In spoken interaction, sarcasm is often marked with a special intonation (Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant & Tree, 2005; Rockwell, 2007), air quotes (Attardo, 2000) or an incongruent facial expression (Muecke, 1978; Rockwell, 2003; Attardo et al., 2003). In written communication, authors do not have such clues at their disposal. Since sarcasm is more difficult to comprehend than a literal utterance (Gibbs, 1986; Giora, 2003; Burgers, van Mulken, & Schellens, 2012a), it is likely that addressees do not pick up on the sarcasm and interpret the utterances literally. To avoid misunderstandings, writers use linguistic markers for irony (Burgers, van Mulken, & Schellens, 2012b): tropes (a metaphor, hyperbole, understatement or rhetorical question), schematic irony markers (repetition, echo, or change of register), morphosyntactic irony markers (exclamations, interjections, or diminutives), or typographic irony markers (such as capitalisation, quotation marks and emoticons). Thus, besides hashtags to mark the opposite valence of a tweet, Twitter members may also use linguistic markers. A machine learning classifier that learns to detect sarcasm should in theory be able to discover at least some of the features that

Burgers et al. (2012b) list, if given sufficient examples of all of them in a training phase. While metaphor and understatement may be too complex to discover, exclamations and typographical markers should be easy to learn. Hyperbole, or the 'speaker overstating the magnitude of something' (Colston, 2007, p. 194), may be discovered by the classifier by the fact that it is often linked to words that signal intensity, as we now analyse in more detail.

Especially in the absence of visual markers, sarcastic utterances need strong linguistic markers to be perceived as sarcastic (Attardo et al., 2003), and hyperbole is often mentioned as a particularly strong marker (R. J. Kreuz & Roberts, 1995). It may be that a sarcastic utterance with a hyperbole ('fantastic weather') is identified as sarcastic with more ease than a sarcastic utterance without a hyperbole ('the weather is good'). While both utterances convey a literally positive attitude towards the weather, the utterance with the hyperbolic 'fantastic' may be easier to interpret as sarcastic than the utterance with the non-hyperbolic 'good'. Hyperbolic words carry intensity. Bowers (1964) defines language intensity as 'the quality of language which indicates the degree to which the speaker's attitude toward a concept deviates from neutrality' (p. 416). According to van Mulken and Schellens (2012), an intensifier is a linguistic element that can be removed or replaced while respecting the linguistic correctness of the sentence and context, but resulting in a weaker evaluation. Intensifiers, thus, strengthen an evaluative utterance and could make an utterance hyperbolic. Typical word classes of intensifiers used for hyperbolic expressions are adverbs ('very', 'absolutely') and adjectives ('fantastic' instead of 'good') , in contrast to words which leave an evaluation unintensified, like 'pretty', 'good' and 'nice'. According to Liebrecht (2015), typographical elements such as capitals and exclamation marks are also intensifying elements which can create hyperbolic utterances. So, there is an overlap between linguistic elements to intensify and linguistic elements to overstate utterances. It may be that senders use such elements in their tweets to make the utterance hyperbolic, in order to signal sarcasm.

6.2 Related Work

The automatic classification of communicative constructs in short texts has become a widely researched subject in recent years. Large amounts of opinions, status updates, and personal expressions are posted on social media platforms such as Twitter. The automatic labeling of their polarity (to what extent a text is positive or negative) can reveal, when aggregated or tracked over time, how the public in general thinks about certain things. See Montoyo, Martínez-Barco,

and Balahur (2012) for an overview of recent research in sentiment analysis and opinion mining.

A major obstacle for automatically determining the polarity of a (short) text are constructs in which the literal meaning of the text is not the intended meaning of the sender, as many systems for the detection of polarity primarily lean on positive and negative words as markers. The task to identify such constructs can improve polarity classification, and provide new insights into the relatively new genre of short messages and microposts on social media. Previous works describe the classification of emotions (Mohammad, 2012; Davidov, Tsur, & Rappoport, 2010a), irony (Reyes et al., 2013), sarcasm (Tsur et al., 2010; Davidov, Tsur, & Rappoport, 2010b; González-Ibáñez et al., 2011) and satire (Burfoot & Baldwin, 2009).

Most common to our research are the works by Reyes et al. (2013), Tsur et al. (2010), Davidov et al. (2010a), and González-Ibáñez et al. (2011). Reyes et al. (2013) collect a training corpus of ironic tweets labeled with the hashtag '#irony', and train classifiers on different feature sets representing higher-level concepts such as unexpectedness, style, and emotions. The classifiers are trained to distinguish '#irony'-tweets from tweets containing the hashtags '#education', '#humour', or '#politics', achieving $F_{\beta=1}$ scores of around 0.70. Tsur et al. (2010) focus on online product reviews, and try to identify sarcastic sentences from these in a semi-supervised fashion. They collect training data by manually annotating sarcastic sentences, and retrieving additional training data based on the annotated sentences as queries. Sarcasm is annotated on a scale from 1 to 5. As features, Tsur et al. (2010) infer patterns from these sentences that consist of high-frequency words and content words. Their system achieves an $F_{\beta=1}$ score of 0.79. Davidov et al. (2010a) apply a comparable system on a small set of tweets that were manually annotated as sarcastic or not, and achieve an $F_{\beta=1}$ score of 0.83. When testing the system on tweets marked with '#sarcasm', the $F_{\beta=1}$ score drops to 0.55. They state that apart from indicating the tone of a tweet, '#sarcasm' might be used as a search anchor and as a reference to a former sarcastic tweet, adding a fair amount of noise to the data. González-Ibáñez et al. (2011) aim to distinguish sarcasm from literally positive and negative sentiments in tweets. They collected tweets belonging to all three categories based on hashtags describing them ('#sarcasm' and '#sarcastic' for sarcastic tweets) and tested classifier performance on the discrimination of the three categories through 5-fold cross validation on a set comprising 900 tweets. As features they make use of word unigrams and higher-level word categories. While the classifier achieves an accuracy of only 57%, it outperforms human judgement. González-Ibáñez

et al. (2011) conclude that the lack of context makes the detection of sarcasm in tweets difficult, both for humans and for machines.

In the works described above, a system is tested in a controlled setting: Reyes et al. (2013) compare irony to a restricted set of other topics, Tsur et al. (2010) take from the unlabelled test set a sample of product reviews with 50% of the sentences classified as sarcastic, and González-Ibáñez et al. (2011) train and test in the context of a small set of positive, negative and sarcastic tweets. In contrast, we apply a trained sarcasm detector to a real-world test set representing a realistically large sample of tweets posted on a random day, the vast majority of which is not sarcastic. Detecting sarcasm in social media is, arguably, a needle-in-a-haystack problem: of the 2.25 million tweets we gathered on a single day, 353 are explicitly marked with #sarcasm or its pseudo-synonyms. It is therefore only reasonable to test a system in the context of a typical distribution of sarcasm in tweets.

6.3 Experimental Set-up

6.3.1 Data

Hashtag selection

As argued in Section 6.1.1, while '#sarcasm' ('#sarcasme' in Dutch — we use the English translations of our hashtags throughout this chapter) is the most obvious hashtag for sarcastic tweets, we base our training set on an expanded set of pseudo-synonymous hashtags. We also found empirical evidence that we need to expand the set of hashtags. Liebrecht, Kunneman, and van den Bosch (2013) reported that training a classifier solely on '#sarcasm' as a training label resulted in high weights for hashtags that have the same function as '#sarcasm': to switch the evaluative valence or give a description of the type of tweet. While Qadir and Riloff (2013) expand sets of hashtags denoting the emotion of a tweet by bootstrapped learning, this approach does not seem appropriate for the more subtle rhetorical instrument of sarcasm. We decided to extract all hashtags from the ranked list of features from the (Liebrecht et al., 2013) study and manually examine the tweets accompanying them by means of `twitter.com`. From this examination, we selected the hashtags that almost unambiguously denoted sarcasm in a tweet in addition to '#sarcasm': '#irony', '#cynicism', and '#not'. The former two denote tropes comparable to sarcasm, while the latter is also typically used to switch the evaluative valence of a message. Hashtags that only partly overlap in function such as '#joke' or '#haha' were not included due to

	# tweets after filtering
#not	353,758
#sarcasm	48,992
#irony	3,285
#cynicism	404
total	406,439

TABLE 6.1: Overview of the dataset with sarcastic tweets used for training.

their ambiguous usage (either shifting the evaluative valence of a message or simply denoting a funny tweet).

Data collection

For the collection of tweets we made use of the TwiNL database. We collected all tweets that contained the Dutch versions of the selected hashtags '#sarcasm', '#irony', '#cynicism', and '#not' until January 31st 2013. This resulted in a set of 644,057 tweets in total. Following Mohammad (2012) and González-Ibáñez et al. (2011), we cleaned up the dataset by only including tweets in which the given hashtag was placed at the end or exclusively followed by other hashtags or a url. Hashtags placed somewhere in the middle of a tweet are more likely to be a grammatical part of the sentence than a label (Davidov et al., 2010a), and may refer to only a part of the tweet. Additionally, we discarded retweets (repostings of an earlier tweet by someone else). Applying these filtering steps resulted in 406,439 tweets in total as training data. Table 6.1 offers more details on the individual hashtags; '#not' occurs a factor more frequently than '#sarcasm', which in turn occurs a factor more frequently than '#irony';'#cynicism' again occurs a factor less frequently.

We trained a classifier on sarcastic tweets by contrasting them against a background corpus. For this, we took a sample of tweets in the period from October 2011 until September 2012 (not containing tweets with any of the sarcastic hashtags). To provide the classifier with an equal number of cases for the sarcasm and background categories and thus produce a training set without class skew, 406,439 tweets were selected randomly, equal to the amount of sarcastic tweets. Again, we did not include retweets in the sample.

To test our classifier in a realistic setting, we collected a large sample of tweets posted on a single day outside the time frame from which the training set is collected, namely February 1, 2013. After removal of retweets, this set of tweets contains approximately 2.25 million tweets, of which 353 and with one of the sarcasm hashtags.

While the distribution of sarcastic versus other tweets on the test day is highly imbalanced, we chose not to copy this distribution to the training stage. As pointed out by Chawla, Japkowicz, and Kotcz (2004), in an imbalanced learning context classifiers tend to be overwhelmed by the large classes and ignore the small ones. To avoid the influence of class size, we decided to completely balance the tweets with and without '#sarcasm' in the training set. This is likely to drive the classifier to overshoot its classification of tweets as sarcastic in the testset, but we consider only the top of its confidence-based ranking; in our evaluation of the ability of the classifier to detect sarcasm in tweets that lack an explicit hashtag, we evaluate the ranking of the classifier with precision at $n < 250$.

6.3.2 Classification

All collected tweets were tokenised.[1] Punctuation, emoticons, and capitalisation information were kept, as these may be used to signal sarcasm (Burgers et al., 2012b). We made use of word unigrams, bigrams and trigrams as features (including punctuation and emoticons as separate words). User names and URLs were normalised to 'USER' and 'URL' respectively. We removed features containing one of the hashtags from the training set. Finally, we removed terms that occurred three times or less or in two tweets or less.

As classification algorithm we made use of Balanced Winnow (Littlestone, 1988) as implemented in the Linguistic Classification System.[2] This algorithm is known to offer state-of-the-art results in text classification, and produces interpretable per-class weights that can be used to, for example, inspect the highest-ranking features for one class label. The α and β parameters were set to 1.05 and 0.95 respectively. The major threshold ($\theta+$) and the minor threshold ($\theta-$) were set to 2.5 and 0.5. The number of iterations was bounded to a maximum of three.

6.3.3 Evaluation

To evaluate the outcome of our machine learning experiment we ran two evaluations. The first evaluation focuses on the 353 tweets in the test set ending with one of the selected sarcasm-hashtags, among 2.25 million other non-sarcastic tweets. We measured how well these tweets were identified using the true positive rate (TPR, also known as recall), false positive rate (FPR) and their joint score, the area under the curve (AUC). AUC is a common evaluation metric that

[1] Tokenisation was carried out with Ucto, `https://languagemachines.github.io/ucto/`.
[2] `http://www.phasar.cs.ru.nl/LCS/`

Class	# trainingdata	TPR	FPR	AUC	Samples	Classifications	Correct
background	406,439	0.83	0.13	0.85	2,246,551	1,870,760	1,870,714
sarcasm	406,439	0.87	0.17	0.85	353	376,144	307

TABLE 6.2: Retrieval of sarcastic tweets from the testset of 01/02/13 (TPR = True Positive Rate, FPR = False Positive Rate, AUC = Area Under the Curve).

is argued to be more resistant to skew than the $F_{\beta=1}$ score, by relying on FPR rather than precision (Fawcett, 2004).

For the second evaluation we manually inspected the test tweets that were identified by the classifier as sarcastic, but do not carry any of the sarcastic hashtags. While they would be labeled as false positives in the first evaluation, the absence of one of these hashtags does not necessarily imply the tweet is non-sarcastic. In fact, the proper detection of sarcastic tweets not explicitly marked as such with a hashtag would be the ideal functionality of our classifier. For this evaluation we make use of the classifier's characteristic to assign per-instance scores to each label, which can be seen as its confidence in that label. We rank its predictions by the classifier's confidence on the 'sarcasm' label and inspect manually which of the top-ranking tweets is indeed sarcastic. Based on this manual annotation we can compute the precision at different rank numbers, which may reveal whether the top-ranked false positives are in fact sarcastic tweets.

6.4 Results

Results for the first evaluation are displayed in Table 6.2. Of the 353 tweets explicitly marked with '#sarcasm' or its pseudo-synonyms on the test day, 307 (87%) are identified as sarcastic, in addition to $376,144$ tweets that do not contain such a hashtag. Because this latter amount is not a big part of the 2.25 million tweets in the test set, the FPR is fairly low and a good AUC of 0.85 is achieved.

Besides generating an absolute winner-take-all classification, our Balanced Winnow classifier assigns scores to each label that can be seen as its confidence in that label. We can rank its predictions by the classifier's confidence on the 'sarcasm' label and inspect manually which of the top-ranking tweets that do not contain any of the four target hashtags is indeed sarcastic. We generated a list of the 250 most confident 'sarcasm'-labeled tweets. Three annotators (three of the authors) judged these tweets as being either sarcastic or not. The instructions beforehand were to positively annotate tweets that were clearly expressing a positive or negative valence that is shifted by the language use. In case of doubt, for example due to the lack of context, a tweet should be annotated as

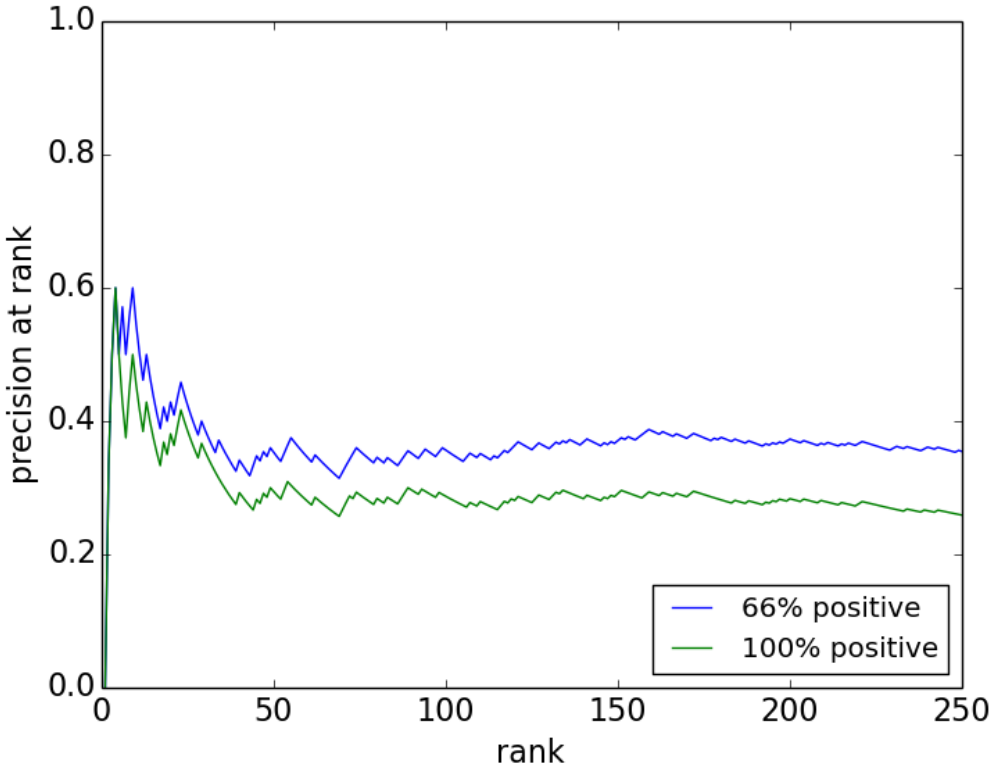


FIGURE 6.1: Precision at $\{1 \ldots 250\}$ on the sarcasm class.

non-sarcastic. The sarcasm should be clear from the text in a tweet; the annotator was not allowed to enquire into the conversational context when a tweet was addressed to one or more twitter users.

When taking the majority vote of the three annotators as the golden label, a curve of the precision at all points in the ranking can be plotted. This curve is displayed in Figure 6.1. The overall performance at the end of the plotted curve is about 0.35. After peaking at a precision of 0.60 after ten tweets, precision decreases rapidly before stabilising after rank 50. Precision scores are lower if sarcasm is only labeled with unanimous agreement between the annotators, ending below 0.30.

In order to test for inter-coder reliability, Cohen's Kappa was used. In line with Siegel and Castellan (1988), we calculated a mean Kappa based on pairwise comparisons of all possible coder pairs. The mean inter-coder reliability between the three possible coder pairs is moderate (Landis & Koch, 1977) at $\kappa = 0.53$. The average mutual F-score over all annotator pairs is 0.72, indicating that annotators disagree in about a quarter of all cases.

6.5 Analysis

6.5.1 Reliability of the training set

An important additional check on our results concerns the reliability of the user-generated sarcastic hashtags as golden labels, as Twitter users cannot all be assumed to understand what sarcasm is, or be versed in using tropes. The three annotators who annotated the ranked classifier output also coded a random sample of 250 tweets with sarcastic hashtags from the training set. The tweets were sampled proportional to the percentage of the four hashtags in the training set (e.g.: 162 '#not'-tweets, 86 '#sarcasm'-tweets and 2 '#irony'-tweets). The instructions beforehand were to decide whether a tweet contains a positive or negative valence, which is shifted by means of the hashtag at the end of the tweet.

The average score of agreement between the three possible coder pairs turned out to be moderate ($\kappa = .44$), but due to the majority of the tweets being genuinely sarcastic, the mutual F-score between the annotators is 0.94, indicating a disagreement on a fairly random 6% of cases. Taking the majority vote over the three annotations as the reference labeling, 212 of the 250 annotated sarcastic tweets, about 90%, were found to be sarcastic. Using hashtags as golden labels thus introduces about 10% noise into the labeled training data. The separate outcomes of the annotated tweets for '#not' and '#sarcasm' are in balance, with respective scores of 90% and 91%. These outcomes are in line with a similar annotation that was conducted in Liebrecht et al. (2013), sampling 250 tweets that were all labeled with the hashtag '#sarcasm'. Of these tweets, 85% were judged as being actually sarcastic.

6.5.2 Predictors of a sarcastic tweet

While the classifier performance gives an impression of its ability to detect sarcastic tweets, the strong indicators of sarcasm as discovered by the classifier may provide additional insights into the usage of sarcasm on Twitter. Including only word unigrams, bigrams, and trigrams as features brings about an unbiased classifier model to be analysed. We set out to analyse the feature weights assigned by the Balanced Winnow classifier, by taking into account the 500 tokens and n-grams with the highest positive weight towards the sarcasm class. Even though Liebrecht et al. (2013) reported on topical words appearing as strong predictors, relating to school, the weather, holidays, public transport, soccer, and television programs, in our current study, which is based on a significantly

larger training set, such topics are hardly present in the top 500. The 500 words and n-grams are mostly adverbs and adjectives that realise a positive evaluation (including intensifiers), exclamations, and non-sarcastic hashtags for meta-communication.

We expected that the sarcastic utterances contained many intensifiers to make the tweets hyperbolic. The list of strongest predictors shows that some intensifiers are indeed strong predictors of sarcasm, such as (with and without capitals) *geweldig* ('awesome'), *heerlijk* ('lovely'), *prachtig* ('wonderful'), *boeiend* ('fascinating'), *allerleukste* ('most fun'), *perfect*, and *super*. However, many unintensified positive adverbs and adjectives occur in the list of strongest predictors as well, such as *interessant* ('interesting'), *gezellig* ('cozy'), *leuk* ('fun'), *handig* ('handy'), *slim* ('smart'), *charmant* ('charming') and *nuttig* ('useful'). Considerably less negative words occur as strong predictors, supporting our hypothesis that the utterances are mostly positive, while the opposite meaning is meant. This finding corresponds with the results of Burgers et al. (2012b), who show that 77% of the ironic utterances in Dutch communication are literally positive. It also concurs with the observation that a sarcastic utterance always implies an evaluation: these (positive) adverbs and adjectives explicitly indicate (and thus mark) that the sender intentionally conveys an attitude towards his or her message.

A substantial set of positive exclamations are found by the classifier as strong predictors. Exclamations are another means to make an utterance hyperbolic and thereby sarcastic. Examples of Dutch exclamations within the top 500 of most predictive features are (with and without # or capitals): *jippie, yes, goh, joepie, jeej, jeuj, yay, woehoe*, and *wow*.

The fourth group of features in the top 500 are non-sarcastic hashtags that signal meta-communication, such as '#humor', '#lml' ('love my life'), '#wehebbenerzinin' ('we are looking forward to it'), '#gaatgoed' ('all is well'), '#bedankt' ('thanks'), and '#grapje' ('joke').

To inspect in more detail the actual occurrence of the four types of words that constitute the top 500 of most predictive features, we further analyse the sarcastic tweets without sarcastic hashtags that our classifier correctly identifies in the top 250 ranked tweets of our test day, and contrast this with the tweets in our training set that do have a sarcastic hashtag. Figure 6.2 displays two proportional Venn diagrams of occurrences in these two sets of tweet of the four aforementioned categories of markers: intensified and unintensified adverbs and adjectives, exclamations, and non-sarcastic hashtags. The Venn diagram on the right in Figure 6.2 visualises the proportions of tweets without a hashtag

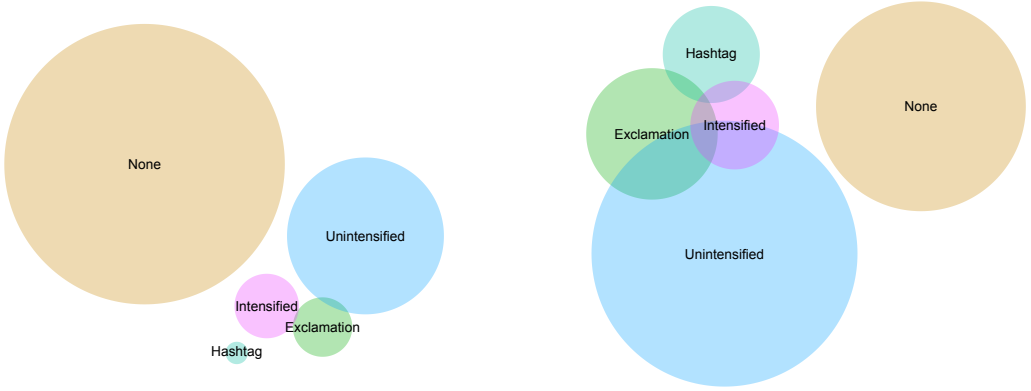


FIGURE 6.2: Proportional Venn diagrams of the co-occurrences of linguistic markers and hashtags in all sarcastic training tweets (left) and detected sarcastic test tweets without a hashtag #sarcasm (right).

correctly identified as being sarcastic that have one or more of these four categories, or none of them (the circle labeled 'None'). The overlap between circles in the Venn diagram visualises which proportion of tweets have a combination of two or more of the four marker categories. The left diagram represents all tweets in the training set with a sarcastic hashtag, while the right diagram represents sarcastic tweets without a sarcastic hashtag. As can be seen in the figure, most sarcastic tweets without a hashtag have unintensified evaluative words. The other three categories occur less frequently, and some of these occur in combination, the most frequent combination being between unintensified evaluative words and exclamations.

The left diagram differs in three aspects from the right diagram: first, the number of tweets containing none of the major linguistic sarcasm markers is the largest category; second, the four categories almost never co-occur. The overall third observation is that the presence of a sarcastic tag in the training tags appears to mute the occurrence of non-sarcastic hashtags. These differences suggest that the presence of an explicit sarcasm hashtag requires fewer other clues to signify sarcasm.

6.5.3 #sarcasm in French tweets

We have shown that a polarity shift between the actual and intended valence of a message can to a certain extent be recognised automatically in the case of Dutch tweets, by means of hashtag labels. This complements previous findings for English tweets. Thus, such hashtags are predominantly applied in the same way in both languages. Future research would be needed to chart the prediction

of sarcasm in languages that are more distant to Dutch. As the findings in this analysis suggests, sarcasm may be signalled rather differently in other cultures (Goddard, 2006). Languages may use the same type of marker in different ways, like a different intonation in spoken sarcasm by English and Cantonese speakers (Cheang & Pell, 2009). Such a difference between languages in the use of the same marker may also apply to written sarcastic utterances, such as tweets.

To investigate the potential success of leveraging hashtag marked tweets in other language regions, we set out to annotate a sample of 500 French tweets ending with #sarcasme (French for 'sarcasm'). The tweets were harvested from `topsy.com`, by means of the otter API[3]. We queried tweets containing #sarcasme, setting the language to French and including all days in 2012 and 2013. This resulted in 8,301 tweets. From this sample, we removed retweets and tweets that did not end with #sarcasme, and took a random sample of 500 tweets. The tweets were annotated by one of the authors and a second person. Both annotators were L1 speakers of Dutch with a French L2 near-native proficiency. The instructions beforehand were the same as for the annotation of the Dutch #sarcasm tweets sampled from the training data.

The annotators marked 63% of the tweets both as sarcastic, with a moderate κ of 0.43 and a mutual F-score of 0.85. The percentage of sarcastically marked tweets by both annotators is smaller than the 90% attained with the Dutch tweets. When we split the three annotators of the Dutch tweets in pairs of two, allowing a better comparison with the two annotators of the French tweets, the percentages are also higher (0.85, 0.82, and 0.84).

Speakers of French seem to be more lenient with the hashtag #sarcasme than speakers of Dutch (or English for that matter), because they also use it to signal other rhetorical figures, such as paradoxes, rhetorical questions and other types of humor. Since the instruction explicitly asked annotators to look for a shift in evaluative valence for a tweet being labeled as sarcastic, the percentage of polarity shifting tweets is accordingly lower. Moreover, the number of tweets without an explicit evaluation was also considerable. Apparently, users of French in tweets more heavily rely on context (and on the receiver being able to interpret the tweet correctly) than Dutch or English users do. The difference between Dutch and French sarcastic tweets suggests that culture influences the use and reception of sarcasm (and especially the use of '#sarcasme'). This is in line with Holtgraves (2005) who argues that the use and interpretation of non-literal meanings can be culture-specific.

[3] `https://code.google.com/p/otterapi/wiki/Resources`

6.6 Conclusion and Discussion

In this study we developed and tested a system that detects sarcastic tweets in a realistic sample of 2.25 million Dutch tweets posted on a single day, trained on a set of 406 thousand tweets, harvested over time, marked by the hashtags '#sarcasm', '#irony', '#cynicism', or '#not' by the senders, plus 406 thousand tweets without these tags. The classifier attains an AUC score of 0.84 and is able to correctly spot 309 of the 353 tweets among the 2.25 million that were explicitly marked with the hashtag, with the hashtag removed. Testing the classifier on the top 250 of the tweets that it ranked as most likely to be sarcastic, but did not have a sarcastic hashtag, it attains only a 35% average precision. We can conclude that it is fairly hard to distinguish sarcastic tweets from literally intended tweets in an open setting, though the top of the classifier's ranking does identify many sarcastic tweets not explicitly marked with a hashtag.

An additional linguistic analysis provides some insights into the characteristics of sarcasm on Twitter. We found that most tweets contain a literally positive message, and contain four types of markers for sarcasm: intensified as well as unintensified evaluative words, exclamations, and non-sarcastic hashtags. Intensified evaluative words and exclamations induce hyperbole, but they occur less frequently in sarcastic tweets than unintensified evaluative words. Note that we based our selection of marker categories on the top 500 most predictive features; other linguistic markers from Burgers et al. (2012b) did not occur in this set and were not included in this study. The differences between the occurrence or absence of markers displayed in the two Venn diagrams of Figure 6.2 indicate that the inclusion of a sarcastic hashtag reduces the use of linguistic markers that otherwise would be needed to mark sarcasm. Arguably, extralinguistic elements such as hashtags can be seen as the social media equivalent of non-verbal expressions that people employ in live interaction when conveying sarcasm. As Burgers et al. (2012a) show, the more explicit markers an ironic utterance contains, the better the utterance is understood, the less its perceived complexity is, and the better it is rated. Many Twitter users already seem to apply this knowledge.

To investigate the usefulness of a sarcastic hashtag to train sarcasm detection in other language regions, we annotated 500 French tweets containing #sarcasme, finding that the majority of French tweets could indeed be labeled as sarcastic, but to a lesser extent than Dutch tweets. In other words: also in French, the hashtag signals a polarity switch in most cases. Apart from hashtag usage,

markers of sarcasm can be language-specific. In Dutch, for example, diminutives can mark irony (Burgers et al., 2012a), while the neighbour language English does not have this device. Dedaić (2005) shows other language-specific markers that have been associated with irony in the Croatian language; as did Bennett-Kastor (1992) for the Ghanese language Sissala. Knowledge of specific sarcasm markers in a language could be used as explicitly added features to our system.

This study focused on the detection of one specific target. However, the procedure that we applied lends itself well for automatisation, and hence can be applied on many more hashtags. In the next study, we subject 24 hashtags to this procedure of collecting tweets with the hashtag, collecting an equal number of contrasting tweets, training a classifier on the words surrounding the hashtag and testing it on a large set of tweets posted on a single day. Focusing on hashtags that convey emotion, we test the predictability of these hashtags and essentially their suitability to train a general model of the emotion that they refer to.

Chapter 7

The (un)predictability of emotional hashtags in Twitter

Based on: Kunneman, F., Liebrecht, C. & van den Bosch, A. (2014). The (un)predictability of emotional hashtags in Twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)* (pp. 26–34), Stroudsburg, PA, USA: ACL.

Hashtags in Twitter posts may carry different semantic payloads. Their dual form (word and label) may serve to categorise the tweet, but may also add content to the message, or strengthen it. Some hashtags are related to emotions. In this study we employ machine learning classifiers to test to what extent tweets that are stripped from their hashtag could be re-assigned to this hashtag. About half of the 24 tested hashtags can be predicted with AUC scores of 0.80 or higher. However, when we apply the three best-performing classifiers to unseen tweets that do not carry the hashtag but might have carried it according to human annotators, the classifiers manage to attain a precision-at-250 of 0.70 for only two of the hashtags. We observe that some hashtags are predictable from their tweets, and strengthen the emotion already expressed in the tweets. Other hashtags are added to messages that do not predict them, presumably to provide emotional information that was not yet present in the tweet.

7.1 Introduction

Since the launch of Twitter in 2006 the microblogging service has proven to be a valuable source of research on the linguistic expression of sentiment and affect. Sentiments and emotions are important aspects of status updates and conversations in tweets (Ritter, Cherry, & Dolan, 2010; Dann, 2010). Many of them express an emotion of the sender: according to Roberts, Roach, Johnson, Guthrie, and Harabagiu (2012), 43% of the 7,000 tweets they collected are an emotional expression. Automatically detecting the emotion in tweets is key to understand the sentiment underlying real world events and topics.

Potentially, Twitter offers a vast amount of data to exploit for the construction of computational models able to detect certain sentiments or emotions in unseen tweets. Yet, in the typical scenario of applying supervised machine learning classifiers, some annotation effort will be required to label sentiments and emotions reliably. Currently there are two main approaches to labeling tweets. The first is the annotation of data by human experts (Alm, Roth, & Sproat, 2005; Aman & Szpakowicz, 2007). This approach is known to result in high-precision annotated data, but is labour-intensive and time-consuming.

The second approach is to use the annotations that Twitter users themselves add to a tweet: hashtags. A hashtag is an explicitly marked keyword that may also serve as a word in the context of the other non-tagged words of the post. The usage of a hashtag in Twitter serves many purposes beyond mere categorisation, most of which are conversational (Huang, Thornton, & Efthimiadis, 2010). Hashtags that express emotions are often used in tweets and are therefore potentially useful annotations for training data. W. Wang, Chen, Thirunarayan, and Sheth (2012) state that annotating interpretative labels by humans other than the author is not as reliable as having the data annotated by the author himself. As far as emotions can be self-observed and self-reported, authors arguably have the best information about their own emotions. Following González-Ibáñez et al. (2011), Mohammad (2012) presents several experiments to validate that the emotional labels in tweets are consistent and match intuitions of trained judges.

Using hashtags as annotated training data may therefore be useful for generating emotion detectors. Yet, not all hashtags are equally suitable for this task. Even a high level of consistency and predictability in hashtag usage might not be sufficient. Mohammad (2012) argues that emotion hashtags are included in tweets by users in two different ways. First, the hashtag can *strengthen* the emotion already present in the tweet. By adding the hashtag in for example 'I hate

making homework #fml' (#fml is an acronym for 'fuck my life'), the sender reflects on his own negative message and strengthens it with an abbreviated expletive. Second, the hashtag can *add* emotion to the message in order to avoid miscommunication. Lacking the richness of non-verbal cues in face-to-face communication, as well as the space to elaborate, attenuate, or add nuance, users of Twitter might deploy hashtags to signify the intention or emotion of their message. In the expression 'Making homework #fml', for example, a Twitter user adds sentiment to the message to clarify his negative attitude towards the described activity. Mohammad (2012, p. 248) formulates the second function of a hashtag as follows: 'reading just the message before the hashtag does not convey the emotions of the tweeter. Here, the hashtag provides information not present (implicitly or explicitly) in the rest of the message.'

Arguably, hashtags that are most often used to add emotion to an otherwise emotionally neutral message (the second function) will not provide proper training data for the detection of the emotion linked to the hashtag; only examples of the first function may serve that purpose. As this information is not explicit, the suitability of a hashtag as an emotion label needs to be revealed in another way. We propose an automatic method that uses machine learning-based text classification. We put this method into practice for a number of hashtags expressing emotion in Dutch tweets. The novel contribution of this study is the proposal of an objective, empirical handle of the two usages of emotion hashtags as formulated by Mohammad (2012). Furthermore, we exemplify a new type of study that tests our hypothesis in the realistic scenario of testing on a full day of streaming tweets with no filtering.

7.2 Related Work

Leveraging uncontrolled labeling to obtain large amounts of training data is referred to as *distant supervision* (Snow, Jurafsky, & Ng, 2005). With its conventions for hashtags and emoticons as extra-linguistic markers, Twitter is a potentially suitable platform for implementing classification based on distant supervision. In the field of sentiment analysis, Pak and Paroubek (2010) and Go, Bhayani, and Huang (2009) select emoticons representing positive and negative sentiment to collect tweets with either of the polarities.

Several studies focusing on the task of emotion detection in Twitter also apply distant supervision. The studies in which it is applied vary in a number of ways. First, the type of markers by which data is collected differs. Most often only hashtags are used, occasionally combined with emoticons. Davidov et al.

(2010a) use hashtags and emoticons as distinct prediction labels and find that they are equally useful. Suttles and Ide (2013) compare the usage of hashtags, emoticons, and emoji[1], and find that emoji form a valuable addition.

Second, the selection of emotions and markers differs. In many of the studies a predefined set of emotions form the starting point for the selection of markers and collection of data. Emotions can be classified according to a set of basic emotions, such as Ekman's (Ekman, 1971) six basic emotions (happiness, sadness, anger, fear, surprise, and disgust), or the bipolar emotions defined by Plutchik's wheel of emotions (Plutchik, 1980) which are based on the basic emotions anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. The majority of the studies rely on such categorizations (Mohammad, 2012; Suttles & Ide, 2013; W. Wang et al., 2012). In spite of the interesting findings in such studies, basic emotions do not tell the whole story; tweets may contain multiple basic emotions combining into more complex emotions (Roberts et al., 2012; Kamvar & Harris, 2011). Furthermore, by selecting a set of hashtags that are assumed to match the same emotion, the potential variation in the usage of specific hashtags by users is ignored. A different approach is to select single hashtags expressing emotion as starting points, regardless of their theoretical status. Davidov et al. (2010a) select frequent hashtags from a large twitter corpus and let annotators judge the strength of their sentiment. The fifty hashtags with the strongest sentiment are used as label. In our research, we also single out hashtags, focusing on a set of hashtags that are linked to emotions, some of which are complex.

Third, the way in which a classifier is trained and tested differs. In some studies multi-class classification is performed, distinguishing the different target emotions and optionally an emotionally neutral class (Purver & Battersby, 2012; W. Wang et al., 2012). The multitude of classes, class imbalance, and the possibility of single tweets conveying multiple emotions make this a challenging task. The alternative is to train a binary classifier for each emotion (Mohammad, 2012; Qadir & Riloff, 2013; Suttles & Ide, 2013), deciding for each unseen tweet whether it conveys the trained emotion. We apply the latter type of classification.

The fourth and final variation is the way in which classification is evaluated. In the discussed papers, evaluation is either performed in a 10-fold cross-validation setting or by testing the trained classifier on a small, manually annotated set of tweets. We deviate from these approaches by testing our classifiers

[1] http://en.wikipedia.org/wiki/Emoji

on a large set of uncontrolled tweets gathered in a single day, thereby approximating the real world scenario in which emotion detection is applied to the stream of incoming tweets.

7.3 Approach

Our approach is to train a machine learning classifier on tweets that contain an emotion-bearing hashtag and an equal amount of random tweets as counter-examples, which results in a balanced binary classifier for the hashtag. The hashtag itself is stripped from the tweet and purely considered as a label. The classifier is then run on a large sample of tweets, deciding which of the tweets might fit the target hashtag. As some of these test tweets actually contain the hashtag, a first evaluation is to score the amount of tweets of which the hashtag is correctly predicted by the classifier when this hashtag is concealed. Second, the tweets that do not contain the hashtag can be ranked by classifier confidence for the hashtag class, after which the 250 highest ranked tweets are scored by human annotators, who judge whether these tweets convey the emotion that is targeted.

This approach is based on the assumption that a hashtag as a label for emotion detection requires two relations between the hashtag and the text with which it co-occurs in tweets:

1. The context in which users include the hashtag is to a certain extent consistent with the hashtag. In other words, the context (the tweet) would predict the hashtag. If this is the case, our classifier should score well on the retrieval of unseen tweets that contain the hashtag (the first evaluation). Consistency can arise from many different types of features, ranging from topical words to emotion-bearing words.

2. The emotion that is denoted by the hashtag should be reflected in the words surrounding it. Hashtags that add emotion to an otherwise neutral message are inappropriate as annotation label for emotion detection. By evaluating retrieved tweets that do not contain the hashtag on the conveyed *emotion* (instead of their possible fit with the hashtag) we can score the extent to which the classifier trained a model of the emotion in tweets successfully.

Note that hashtags that add a specific emotion to otherwise unemotional tweets are good indicators themselves for detecting emotion in Twitter. Our goal, however, is to create generalisable models of emotion in Twitter that are not restricted to the occurrence of a hashtag.

7.4 Experimental Set-up

7.4.1 Data collection

As a starting point of our experiments we selected 24 hashtags used in Dutch tweets. The selection was inspired on a list of the 2,500 most frequent hashtags in 2011 and 2012, generated from TwiNL. Typically, emotion hashtags are not linked to any specific point in time, and therefore surface in such a list generated from an extended period of tweets. To create the training data, tweets that contain any of the hashtags were collected from TwiNL from the time frame of December 2010 up to and until January 2013. We queried a large sample of Dutch tweets (3,144,781) posted on February 1st 2013, a small portion of which was used as negative examples for our training data. The rest was used as test data.

7.4.2 Classification

For each of the hashtags, training data was generated by balancing the amount of collected tweets containing the hashtag with an equal amount of randomly selected tweets (not containing the hashtag) drawn from the set of tweets collected on February 1st, 2013. The resulting binary classifier was tested on the remainder of tweets in this set. The tweets were preprocessed by extracting word unigrams, bigrams, and trigrams as features. We maintained capitalisation and included punctuation and emoticons as tokens in the n-grams, as we expected such tokens to have predictive power in the context of emotions. Both user names and URLs were normalised to dummy values. All features containing a target hashtag were removed.

Classification was performed by the Balanced Winnow algorithm (Littlestone, 1988), as implemented in the Linguistic Classification System[2]. This algorithm is known to offer state-of-the-art results in text classification, and produces interpretable per-class weights that can be used to, for example, inspect the highest-ranking features for one class label. The α and β parameters were set to 1,05 and 0,95 respectively. The major threshold ($\theta+$) and the minor threshold ($\theta-$) were set to 2,5 and 0,5. The number of iterations was bounded to a maximum of three.

[2] `http://www.phasar.cs.ru.nl/LCS/`

7.4.3 Evaluation

Performance was evaluated by classifying all test tweets and counting the number of tweets with the target hashtag that were positively classified as such, deriving a true positive rate (recall), false positive rate, and area under the curve (AUC) score (Fawcett, 2004).

While this first evaluation gives an indication of the predictability of any hashtag, the ultimate value of a hashtag for emotion detection can be scored by assessing the emotion in positively classified tweets that do not contain the hashtag. This is done by manually annotating the fraction of these tweets that are most confidently positively ranked by the hashtag classifier, as containing the emotion signalled by the hashtag. Three annotators inspected the top 250 of these rankings.

7.5 Results

7.5.1 Hashtag predictability

The results of the 24 classifiers on labelling a large sample of tweets posted on February 1, 2013 are listed in Table 7.1. Each line with a target hashtag represents a separate experiment. The number of training tweets ranges from 19 thousand to 677 thousand for the target hashtag (balanced by an equal amount of random tweets as negative category). The results are sorted by the AUC score.

In this first evaluation our attention focuses on the tweets that include one of the target hashtags. The hashtags themselves are removed at classification time, as our goal is to measure how well our classifiers are able to detect these 'hidden' tags. In this particular stream of tweets, only a limited number of tweets occur that are labeled with our hashtags; the most frequent tag #zinin ('looking forward to it') occurs 1,328 times. Taking #zinin as example, the #zinin classifier labels 158,429 of the test tweets as likely candidates for the hashtag #zinin. Although this is a substantial over prediction, partly caused by the 50%-50% ratio between positive and negative cases in the training set, this still amounts to a false positive rate of only 6%. More importantly, of the 1,328 cases for which it should have predicted #zinin, the classifier labels 1,186 cases correctly, attaining a true positive rate of 89%. The area under the curve (AUC) in true positive rate–false positive rate space is 91%.

Inspecting the performance for all 24 hashtags we observe that about half of the hashtags obtain an AUC of 0.80 or more. The influence of the amount of training data on the AUC score is peripheral ($r = 0.27, p > 0.05$). Furthermore,

Target hashtag	Gloss	# Training tweets	Target instances on test day	Instances classified	Instances correct	TPR	FPR	AUC
#zinin	looking forward to it	677,156	1,328	158,429	1,186	0.89	0.06	0.91
#geenzin	not looking forward to it	427,602	653	231,463	583	0.89	0.08	0.91
#fml	fuck my life	139,044	308	126,045	265	0.86	0.05	0.90
#lml	love my life	41,031	197	343,936	167	0.85	0.11	0.87
#balen	bummer	219,342	134	271,308	108	0.81	0.09	0.86
#jeej	yay	107,667	31	353,807	25	0.81	0.12	0.85
#nietleuk	not nice	85,825	43	359,709	33	0.77	0.12	0.83
#yeah	yeah	290,288	328	349,598	247	0.75	0.12	0.82
#loveit	love it	259,935	336	290,822	247	0.74	0.10	0.82
#jippie	yippie	66,992	27	396,805	21	0.78	0.13	0.82
#joepie	yippie	53,217	39	422,348	29	0.74	0.14	0.80
#yes	yes	115,707	151	373,874	104	0.69	0.12	0.78
#yay	yay	50,737	45	421,660	31	0.69	0.14	0.78
#hmm	hmm	110,171	95	341,936	63	0.66	0.11	0.78
#grr	argh	70,659	145	397,201	97	0.67	0.13	0.77
#like	like	68,499	284	412,714	178	0.63	0.13	0.75
#woehoe	woohoo	19,236	32	584,552	22	0.69	0.19	0.75
#leuk	nice	391,626	971	307,277	592	0.61	0.11	0.75
#bah	grose	298,842	228	273,454	127	0.56	0.10	0.73
#stom	lame	72,957	99	355,731	57	0.58	0.12	0.73
#omg	oh my god	590,560	145	394,447	79	0.54	0.13	0.71
#wauw	wow	146,145	103	467,503	58	0.56	0.15	0.70
#wow	wow	52,488	50	587,662	29	0.58	0.19	0.70
#huh	huh	48,456	25	352,396	12	0.48	0.11	0.68

TABLE 7.1: Results for the prediction of a target hashtag for about 3.1 million Dutch tweets posted on February 1st 2013 (TPR = True Positive Rate, FPR = False Positive Rate, AUC = Area Under the ROC Curve).

there is no clear difference between the predictability of hashtags denoting a positive or negative emotion. The predominantly negative hashtags #geenzin, #fml, #balen and #nietleuk obtain a high AUC, while the other negative hashtags #grr, #bah and #stom are not as predictable. There does not seem to be an a priori property that makes a hashtag more or less predictable, indicating the need for experimentation to confirm the usefulness of a hashtag for emotion detection.

Interestingly, some pairs of synonymous hashtags (#jippie-#joepie, #wauw–#wow, #yes–#yeah, homophonous variants of the same exclamation) and antonymous hashtags (#zinin–#geenzin, #fml–#lml) achieve similar AUC scores. This outcome supports the validity of our approach. Synonymous and antonymous hashtags are employed in similar contexts and are therefore likely to have a similar predictability. This is indeed confirmed by our results. There are counterexamples, however. The pair #yay–#jeej exhibits dissimilar scores. In the case of #leuk there are two antonyms: #nietleuk and #stom. #leuk and #nietleuk have a

	Precision (67% majority)	Precision (100% majority)	Cohen's Kappa	Mutual F-score
#zinin	0.75	0.35	0.09	0.67
#geenzin	0.31	0.21	0.60	0.73
#fml	0.69	0.46	0.48	0.81
#omg	0.49	0.25	0.29	0.67

TABLE 7.2: Precision of correct hashtag predictions of the top 250 'false positives' based on human annotations.

dissimilar score, while #leuk and #stom are rather similar.

7.5.2 Emotion detection

The second evaluation is based on the manual annotation of the 250 tweets that are most positively ranked by a hashtag classifier, on the emotion linked to the target hashtag. Due to the labour-intensive nature of this evaluation, it was not possible to analyse all 24 hashtags. We focused on the output for #zinin, #geenzin, #fml and #omg. The first three achieved the highest true positive rates ranging between 0.86 and 0.89, and AUC scores of 0.90 to 0.91. The latter was included as a comparison, where we expected a poor emotion detection in view of its bad predictability.

For these four hashtags the 250 'false positives' of which the classifier was most certain were annotated by the three authors by taking the binary decision whether a tweet conveys the emotion presumed in tweets containing the hashtag. The emotions most strongly linked to the four hashtags were the following:

- #zinin: conveying anticipatory excitement;
- #geenzin: conveying disinterest
- #fml: conveying self pity
- #omg: conveying an aroused level of indignation, fear, or excitement

Note that #omg is not linked to a single emotion, but rather strengthens several sorts of emotions. This might have been a hampering factor for its predictability. In the annotation for #omg we focused on all three emotions.

Table 7.2 displays the precision scores when taking a simple majority decision over the three annotators (67% majority) and when only counting the cases in which all three annotators agreed (100% majority). The outcomes show reasonably high precision levels for #zinin (75%) and #fml (69%) along with equally

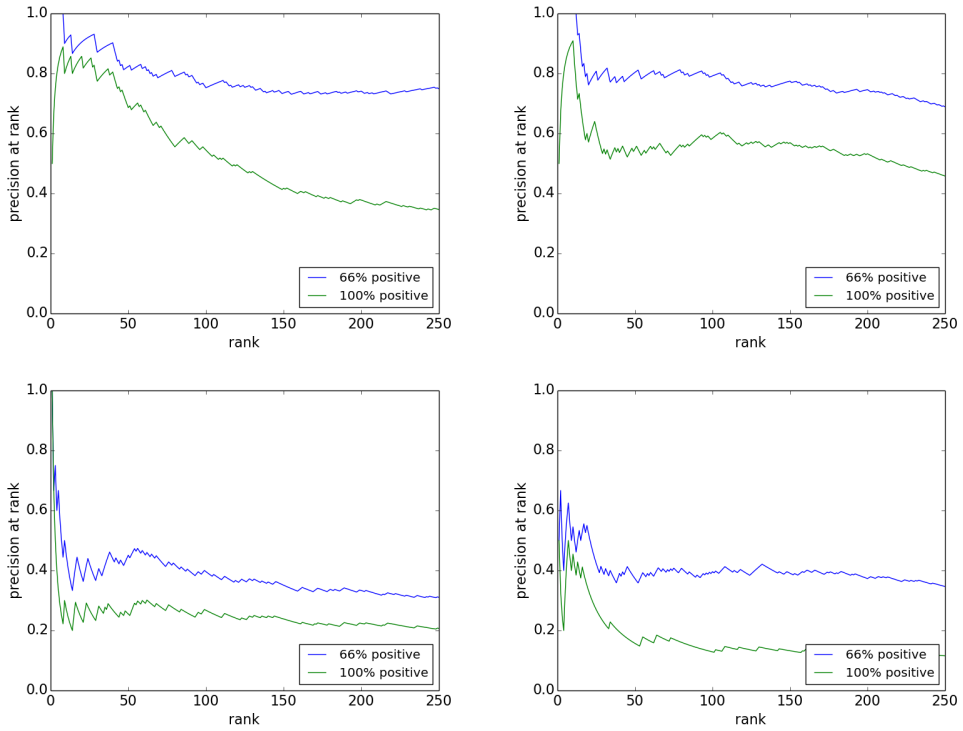


FIGURE 7.1: Precision at $\{1 \ldots 250\}$ on the classes #zinin (top left), #fml (top right), #geenzin (bottom left), and #omg (bottom right).

reasonable mutual F-scores between the annotators (67% for #zinin and 81% for #fml), although Cohen's Kappa is rather low in some cases. On the other hand, #geenzin lags behind with a majority precision of 31%. Also the top 250 for #omg does not often display any of the three most strongly linked emotions.

Plotting the annotations of the ranked tweets in precision-at curves, shown in Figure 7.1, provides further insight into the emotion detection quality in relation to the confidence ranks. Precisions at higher rank cut-offs tend to peak early (indicating that the first top-ranked tweets fit the hashtag best), and decrease slowly or reach a plateau.

The two-fold evaluation that was employed in this study underlines the difference between hashtag predictability and emotion detection. Regarding the three best performing hashtags in terms of predictability, only two, #zinin and #fml, provide utilisable data for emotion detection. Tweets retrieved based on #geenzin seem to have a less overt relation to the emotion of disinterest, although other cues (such as topical words indirectly related to the emotion) led to a fairly correct recovery of tweets that mention the hashtag. Comparing the two evaluations for #omg, scoring low on both, we may assume that hashtag predictability is a requirement for a proper emotion detection.

	% Adequate tweet-hashtag link (67% majority)	(100% majority)	Cohen's Kappa	Mutual F-score
#zinin	93%	80%	0.25	0.93
#fml	78%	54%	0.36	0.83
#geenzin	96%	88%	0.44	0.97
#omg	82%	50%	0.27	0.79

TABLE 7.3: Consistency of training data based on human annotation of a random sample of 250 tweets for each hashtag.

7.6 Analysis

7.6.1 Reliability of hashtag labels

Although W. Wang et al. (2012) state that leveraging labels given by the writers of tweets themselves is reliable, it is questionable whether Twitter users label their tweets accurately (González-Ibáñez et al., 2011). Purver and Battersby (2012) state that some emotions are used consistently (happiness, sadness, and anger), while others are not (fear, surprise, and disgust). They explain this finding by the ambiguity or vagueness of those emotions. To obtain further insights into the consistency of the hashtag-labeled training data, we extracted a sample of 250 hashtag-labelled training tweets for each of the four hashtags that were included both in the evaluation of hashtag predictability and emotion detection. For these tweets we annotated whether there was an adequate link between the hashtag and the rest of the message. The outcomes are given in Table 7.3. While the majority scores for #zinin and #geenzin (0.93 and 0.96) are in line with the respective AUC-scores, #omg has a majority score of 0.82, which is higher than the 0.78 of #fml. Presumably, #omg functions as a consistent intensifier of the exposed emotion in a tweet, but the range of suitable contexts is too big and non-exclusive for an automatic classifier to train a comprehensive model.

7.6.2 Feature categories

While classifier performance gives an indication of its ability to detect emotional tweets per hashtag, the strong indicators of those hashtags discovered by the classifiers may provide additional insight into the usage patterns of emotional hashtags by Twitter users. Having scored the emotion detection quality of four hashtags, we set out to analyse the predictive features of these hashtags. To this end we inspected the feature weights assigned by the Balanced Winnow classifier ranked by the strength of their connection to the emotion label, taking into

	Example	Percentage in top 150 features			
		#zinin	#fml	#geenzin	#omg
emotion hashtag	'#foreveralone'	6.7%	10.0%	2.7%	18.7%
emoticon	':S'	0.0%	4.7%	0.0%	6.7%
exclamation	'noooo'	0.0%	2.7%	0.0%	8.7%
state of being	'curious'	3.3%	7.3%	3.3%	0.7%
temporal reference	'moment'	26.0%	7.3%	10.0%	1.3%
topic	'dentist'	52.7%	48.7%	69.3%	25.3%
other	'ready_to'	11.3%	19.3%	14.7%	38.7%

TABLE 7.4: Shares (in percentages) of seven categories in the top-150 highest-weighted features for four hashtags.

account the 150 tokens and n-grams with the highest positive weight towards the hashtag.

Based on an analysis of the top 150 features for the four hashtags, we distinguished seven categories of features: other emotion-bearing hashtags, emoticons, exclamations, states of being, time expressions, topic reference, and remaining features. Example features for each category, as well as their share in the top 150 features for each hashtag, are presented in Table 7.4. The percentages give an impression of the most dominant types of features in the prediction of the hashtags.

A first observation is that the top features of the #geenzin classifier are predominantly topic related; the list hardly contains any feature that bears emotion. This is in line with the poor performance on the emotion detection evaluation, while the high AUC score can be explained by a relative consistency of the hashtag being used with topical words that have an indirect relation with the emotion, such as homework for school. The more accurate classifier for the opposite of #geenzin, #zinin, uses more temporal references pointing to the event that the person is looking forward to. Also, Dutch positive adjectives such as 'lekker' ('nice') and 'gezellig' (multiple translations[3]), which are strong predictors for #zinin, add to the accuracy of the classifier. There are no clear counterparts for the emotion linked to the opposing #geenzin.

The percentages for #omg display the largest shares of emotion hashtags, emoticons and exclamations, confirming our impression that #omg functions as an intensifying marker of different emotions; this is also reflected in the high percentage of features in the 'other' category.

[3]See `http://en.wikipedia.org/wiki/Gezelligheid`.

The most predictive features for the #fml classifier consist of quite some emoticons, emotional hashtags and exclamations. Furthermore, this classifier model contains most features in the 'state of being' category, mostly relating to the complex emotion of self pity.

7.7 Conclusion and Discussion

In our experiments we showed that machine learning classifiers can be relatively successful both in predicting the hashtag with tweets which were indeed tagged with them, and classifying tweets without the hashtag as exhibiting the emotion denoted by the hashtag, for two of the four fully analysed hashtags: #zinin and #fml. In contrast, the classifier of the hashtag #geenzin was only able to re-link tweets that are stripped from the target hashtag with this hashtag, but failed to capture the complex emotion behind the hashtag. The performance of the #omg classifier lags behind in both tasks.

These findings can be explained by the assumption we made that in order to be a proper emotion label, the context of the hashtag (the rest of the tweet) would need to convey the same emotion as the hashtag. This appears to be the case with #zinin and #fml. We may assume that the message in tweets with #zinin or #fml carries the emotion itself, which is intensified by the hashtag. The alternative relation between the hashtag and the text is that a hashtag adds emotion to an otherwise neutral message: a signalling function. It seems that most of the tweets tagged with #geenzin are examples of this second relation. The classifier performed well at the re-link task, which indicates that it was able to exploit the consistent use of predictive words and phrases, but less well as an emotion detector when we applied the classifier to unseen tweets that do not carry the hashtag. The topical words that the classifier used as predictive features appear to be used in several other settings in which no emotion is conveyed, or different emotions than the one expressed by #geenzin. The fourth hashtag that was fully analysed, #omg, turned out to be overall difficult for our classifier. We defined #omg as conveying an aroused level of indignation, fear or excitement. In comparison to the other three hashtags, this definition is less strictly linked to one emotion (Kim, Bak, & Oh, 2012). Rather, the hashtag is used in the context of three different emotions and is in itself not an emotion, but an emotion intensifier. Possibly, as a result thereof the tweets are more diverse and the hashtag #omg occurs more frequently with other linguistic elements to express emotion, such as emotional hashtags, emoticons and exclamations.

Although time restrictions prevented us from performing a similar analysis of more hashtags, we can conclude that hashtag predictability is fairly high for most of the 24 hashtags in our set. Interestingly, a considerable part of the synonymous and antonymous hashtags led to similar scores, indicating a relationship between the type of emotion conveyed by a hashtag and the degree of consistency by which the hashtag is employed by users.

The insights from this study, as well as the studies in Part One of this thesis, are brought together in the final study, in which we set out to score the degree of anticipointment conveyed in event tweets.

Part III

Expectations and retrospections on Twitter: Converging hashtags and time

I came in with high expectations. You are an anticipointment. I am anticipointed in you.
Armando Iannucci
(interviewing Stewart Lee as part of the 'Stewart Lee's comedy vehicle, season 1' extra content)

Chapter 8

Anticipointment detection in event tweets

Based on: Kunneman, F., van Mulken, M. & van den Bosch, A.. Anticipointment detection in event tweets. Submitted.

Based on the previous studies in this dissertation, in this chapter we research the detection of positive expectation, disappointment, and satisfaction in tweets that refer to events automatically discovered in the Twitter stream. The emotional content shared on Twitter when referring to public events can provide insights into the presumed and experienced quality of the event. We expected to find a connection between positive expectation and disappointment, a succession that is sometimes referred to as *anticipointment*. The application of computational approaches makes it possible to detect the presence and strength of this hypothetical relation for a large number of events. We extracted events from a longitudinal data set of Dutch Twitter posts, and modeled classifiers to recognise emotion in the tweets related to those events by means of hashtag-labeled training data. After classifying all tweets before and after the events in our data set, we summarised the collective emotions by calculating the percentage of tweets classified with an emotion as well as ranking tweets based on the classifier confidence score for an emotion and selecting the 90th percentile. Only a weak correlation of around 0.20 was found between positive expectation and disappointment, while a higher correlation of 0.60 was found between positive expectation and satisfaction. The most anticipointing events were events with a clear loss, such as a canceled event or when the favoured sports team had lost. We conclude that senders of Twitter posts might be more inclined to share satisfaction rather than disappointment after a much anticipated event.

8.1 Introduction

'What's happening?' is the question that the social media platform of Twitter asks its users in the text bar in which they can compose a new message. In line with this impetus, the content on Twitter reflects real-world events that are taking place and the feelings that people have with respect to these events. Part of the messages express feelings about events that are scheduled in the future or have already taken place. We research the extent to which positive expectations are followed by satisfaction or disappointment on Twitter, by means of automatic emotion detection on a large sample of Dutch event tweets.

The commonality between these three emotions, positive expectation, disappointment, and satisfaction, is their connection to the experience during an event; either the expected experience or the evaluation of the experience afterwards. In addition, it can be argued that dependencies between the expected and perceived experience are likely to occur. Disappointment may be more likely if expectations are high, while satisfaction is highest when expectations are exceeded (Miceli & Castelfranchi, 2014). Little is known, however, about the manifestation of these possibly interrelated emotions on Twitter. The increased usage of the word 'anticipointment', which is the sensation of a possibly long period of anticipation followed by a letdown,[1] [2] suggests a correlation between positive expectation and disappointment. Although the word was allegedly coined for the first time in the 1960s, its use has become increasingly relevant in recent years of hyped events and media releases. By targeting a large number of Twitter posts, we aim to quantify sequences of collective anticipointment, as well as its positive counterpart, positive expectation followed by satisfaction.

While computational approaches permit the detection of emotion in Twitter posts to some extent, the unstructured nature of Twitter makes it a challenging task. Based on the study described in Chapter 4, we extract a large number of events from the Dutch Twitter verse and classify the emotion in the event-referring tweets. By means of the observed emotion in tweets before and after events, we measure the extent to which positive expectation, disappointment, and satisfaction are correlated and analyse the events that are most exemplary of anticipointment, or of positive expectation followed by satisfaction. Although we focus on three emotions, the procedure may be applied on any emotion commonly expressed on Twitter.

[1] `http://nancyfriedman.typepad.com/away_with_words/2011/12/word-of-the-week-anticipointment.html`

[2] 'Anticipointment' might also be referred to as the 'anticipation of disappointment'; we do not adopt this definition here.

8.2 Related Work

8.2.1 Event-related emotions

In describing emotions related to anticipation, Miceli and Castelfranchi (2014) distinguish between *beliefs* and *goals*. They posit that so-called Cold Anticipatory Representations are only based on beliefs, while Interested Anticipatory Representations (IAR's) comprise of both beliefs and goals. The latter connect to emotion, as they entail a personal connection to the outcome. The emotion of positive expectation is categorized by Miceli and Castelfranchi as IAR, and defined as 'normative belief': the believed future state is prescribed to happen, and the positive expectation is associated with a subjective satisfaction of this expected outcome. Likewise, disappointment is defined by them as 'a negative emotional reaction to the invalidation of a positive IAR'. Hence, the goal that is connected to an event defines if someone might be disappointed. A person that has positive expectations of a music concert expects to be entertained during the concert. If the concert turns out to be boring, this negative outcome in relation to the goal to be excited leads to disappointment. The definition of anticipointment is similar to that of disappointment: both assume positive expectations prior to a disappointing outcome. Arguably, the difference between the two concepts is that anticipointment emphasizes a possibly prolonged period of intense positive expectations in anticipation of an event that fails to meet the expectations.

As disappointment after an outcome is affected by the expectations of the outcome beforehand, feelings of disappointment can be avoided by lowering expectations. van Dijk, Zeelenberg, and van der Pligt (2003) found that the personal importance of an outcome as well as the temporal proximity of its occurrence have an effect on the deployment of this strategy. They asked psychology students to speculate on their score for a recently finished exam, and observed significantly lower expectations from students who were told that the exam was important for their career and that the score would be revealed shortly. In contrast to important exams, expectations of events that are attended for relaxation or entertainment are often very positive. van Boven and Ashworth (2007) find that expectations of a future social event are more intense than the feelings about the event in retrospect. They argue this is likely due to the concreteness of an actual experience in comparison to an anticipated experience, when many aspects are open for imagination.

Following these studies it can be expected that the highest satisfaction will

be seen after events for which positive expectations are low, while disappointment is most likely after high expectations. It is not clear, however, whether the same processes will be observed in Twitter posts. The emotion that is conveyed through a tweet is not necessarily the emotion that is felt by the sender at that moment. Other processes might be at play, such as the social context and conversational goal of the sender (Thelwall & Kappas, 2014). In addition, in a social platform such as Twitter, emotion might operate at different levels, such as individual, group or cultural (van der Löwe & Parkinson, 2014). Although tweets might not convey a clear collective emotion before or after an event, some events might stir more emotional tweets than others. It is interesting to see the extent to which post-event emotions on Twitter can be explained by the emotion beforehand.

8.2.2 Emotion detection from tweets

The goal of emotion detection is to automatically determine the emotion in a message based on the words that are used. Many different approaches to emotion detection exist that vary in a number of ways, such as the emotions that are targeted, the way in which emotions are classified and the features that are used from a message. We will provide a brief discussion of these variations, and describe the approach that we will apply in this work.

Regarding the emotions that are targeted, a rough division of three approaches can be made. The first approach is to classify the sentiment of a message on a scale from negative to positive (Go et al., 2009; Pak & Paroubek, 2010; Davidov et al., 2010a; Montoyo et al., 2012). The second selects the most basic emotions as defined in psychological literature, such as the six basic emotions of Ekman (1971) (Purver & Battersby, 2012; Roberts et al., 2012; Balabantaray, Mohammad, & Sharma, 2012; Mohammad, 2012; Qadir & Riloff, 2013) or the wheel of eight primary bipolar emotions of Plutchik (1980) (Suttles & Ide, 2013). The third targets any emotion that might be present in the data that is studied (Mohammad & Kiritchenko, 2015; Liew Suet Yan, 2015). In an appeal for adopting this latter notion of emotion in the context of Twitter, Liew Suet Yan (2015) argues that basic emotions, such as Ekman's Six, do not necessarily reflect the range of emotions that are expressed in Twitter. As a result, a system that only focuses on the most basic emotions does not provide an accurate overview of public emotions on Twitter. Liew Suet Yan (2015) describes a bottom-up approach to emotion labeling, in which annotators are provided with a set of randomly selected tweets and individually decide which label best describes the

emotion of the sender. Subsequent discussion by the annotators results in a scheme of emotions that relates to the emotions that are conveyed in Twitter.

In comparison to the three approaches discussed above, we target emotions that occur within a specific selection of data (tweets that look forward and backward to an event) rather than primary emotions. Our focus on three emotions (positive expectation, disappointment, and satisfaction) is driven by theory rather than data. We do not aim to provide an exhaustive overview of the emotions that are expressed on Twitter before and after events. In this sense, our approach is most comparable to the second one.

When labeled data is used to train an emotion model, two dominant approaches exist to acquire the labelings. The first is to manually annotate instances (Roberts et al., 2012; Balabantaray et al., 2012). An advantage of manual annotations is that they generally result in reliable labelings. A disadvantage is that it requires substantial time and effort to generate a proper amount of labeled data. The second approach is to apply distant supervision (Snow et al., 2005) and deduce probable labelings from highly indicative features in the data. In tweets, hashtags are often used by the sender to stress the emotion connected to the message, and sometimes are good proxies for emotional labels (as we describe in Chapter 7). The lack of control over the process in which hashtags are added to tweets is arguably compensated by the large amount of data that can be acquired by means of hashtags. Hashtag-based emotion labels have shown to be useful in predicting the hashtag from the text in a tweet (Purver & Battersby, 2012; Suttles & Ide, 2013), collecting additional hashtags that convey the same emotion (Qadir & Riloff, 2013; Fraisse & Paroubek, 2014), and classifying emotion in tweets that do not contain a hashtag (Davidov et al., 2010a; Mohammad, 2012; Mohammad & Kiritchenko, 2015). In our work, we apply distant supervision to acquire a large number of labeled training tweets, and detect the related emotion in unseen training data.

Features assumed to hold cues for emotions vary from linguistically informed features to shallow surface features. Examples of the use of informed features for emotion detection are part-of-speech tags (Go et al., 2009), lexicons such as WordNet Affect (Mohammad & Kiritchenko, 2015) and General Inquirer (Roberts et al., 2012), and patterns of Highly Frequent Words and Content Words (Davidov et al., 2010a). The advantage is that such features highlight the parts of a text that are presumably most indicative of emotion, and might add useful information. Shallow surface features for emotion detection, typically word n-grams, are among others used by Purver and Battersby (2012), Suttles and Ide (2013) and Qadir and Riloff (2013). The advantage of such features is that

they do not impose unnecessary restrictions on the signals that might point to an emotion. A machine learning algorithm can figure out from the data which words are most indicative. We adopt this latter approach, and use word unigrams, bigrams and trigrams as features.

8.2.3 Emotion detection from real-world event reports on Twitter

Real-world events can have a strong influence on collective emotions on Twitter. Several works aim to measure such emotions automatically, for example to monitor public sentiment and search for correlations with popular events. Bollen et al. (2011) measure the mood state from the text in all public Twitter posts for five months in 2008, and find that fluctuations correlate to big economic, political and social events. Larsen et al. (2015) use a vocabulary of annotated emotion words to score a 10% stream of Twitter messages on the presence of 6 primary and 25 secondary emotion categories. They find correlations with two known events.

Other works focus on tweets that are exclusively related to selected events. Sintsova, Musat, and Pu Faltings (2013) collect tweets that refer to the 2012 Olympic games and use crowd sourcing to generate labels for complex emotions in these tweets. Torkildson, Starbird, and Aragon (2014) focus on the BP Gulf Oil Spill in 2010 and classify all tweets that mention #oilspill in this period on the presence of six emotions. Sykora, Jackson, O'Brien, Elayan, and Von Lunen (2014) apply a lexicon-based approach to identify eight different emotions in tweets reacting to 25 selected events. They analyse several events on the distribution of the detected emotions. Brooks, Robinson, Torkildson, and Aragon (2014) aim to facilitate collaborative visual analysis of event tweets and visualise the automatically labeled sentiment of tweets that refer to the Super Bowl.

Rather than querying tweets that refer to known events, Thelwall, Buckley, and Paltoglou (2011) and Chen, Argueta, and Chang (2015) analyse tweets that refer to automatically detected events. Thelwall et al. (2011) propose an approach to identify the top 30 words with the most 'bursty' time pattern as events from a month of English tweets. The sentiment strength of these tweets was classified and analysed for significant fluctuations in sentiment polarity. Chen et al. (2015) also apply burstiness-based event detection, and classify tweets that refer to the detected events into six emotion categories.

Like the works described above, we aim to automatically detect emotion in tweets referring to real-world events. The events will be collected automatically,

similar to Thelwall et al. (2011) and Chen et al. (2015). Unlike their burstiness-based approach, however, we leverage explicit references to the date of an event in tweets (i.e. not necessarily from a bursty set of tweets close in time, but potentially from a prolonged period of time), which results in the extraction of a diverse set of predominantly public social events. While most of the works presented in this section focus on the emotion in tweets during event time, we focus on tweets that were posted before and after event time. To the best of our knowledge, this work is the first to automatically analyse the relation between emotion in pre-event and post-event Twitter messages.

8.3 Data

We build on earlier work on open-domain event extraction to prepare a dataset of tweets in anticipation and hindsight of events. In this section, we will describe the output of this approach and the procedure to query tweets that refer to the extracted events.

8.3.1 Extracting open-domain events

We extracted events from all tweets available in the TwiNL database, following the same procedure as described in Chapter 5, Section 5.3.1. We applied this approach to TwiNL tweets between 01/01/11 and 31/10/15, resulting in a set of 97,885 events in total.

8.3.2 Harvesting additional event tweets

By definition, our approach to event extraction identifies tweets linked to an event posted before event time and containing a forward-pointing TIMEX. Consequently, the available tweets per event are often only a subset of all tweets that refer to the event: there are also those not containing a time reference and tweets posted after the event. As our aim is to detect emotion in all event tweets that we can identify, both before and after the event, we set out to collect additional tweets.

For each event we queried additional tweet IDs from the TwiNL database by means of the terms that describe the event. These terms have a strong link to the event, but are not all equally useful. (Becker et al., 2012) describe the task of querying additional social media content for known events as a precision and recall problem: some words that describe an event are too broad, such as the name of a city, whereas others might be too narrow, such as the full name of an

event in combination with its location and contents. They use several strategies to ensure both precision and recall, such as comparing queries with different combinations of event properties and ranking query terms by their specificity and time pattern. Many events in our set are characterised by only one event term, which excludes the strategy to query with different combinations. As an alternative, we will score the quality of individual event terms as event descriptors by inspecting their frequency in time.

For each event term, we first collect all tweets in which it is mentioned in a window of 30 days before and after the date of the event. To ensure original content, we stripped away all retweets from this set. From the resulting sequence of 61 days, we count the frequency of tweets per day and calculate the *burstiness* on the date of the event, by dividing the number of tweets on this date by the average number of tweets in the sequence. Research on automatic event detection (Weng & Lee, 2011; C. Li et al., 2012; Qin et al., 2013) shows that a bursty text phrase in Twitter is strongly related to the occurrence of an event. Based on this intuition, event terms that are not found to be bursty on the date of the event are likely not exclusively linked to this event and therefore not useful to query additional event tweets. To ensure a reliable burstiness calculation, we only selected event terms that were mentioned over 20 times on the date of the event. Event terms with a burstiness score of 10 or higher were selected as query term for the event. A burstiness of 10 is a fairly high threshold, by which we tried to ensure a high precision of event tweets. After applying this procedure on all events, 18,237 of them appeared to have useful event terms.

8.3.3 Selecting pre-event and post-event tweets

We separated the event tweets into pre-event and post-event tweets by comparing the post date of each tweet to the date of the event. As we did not know the specific hour and minute of the day at which the events started, we excluded tweets posted on the date of the event. This way we ensured that tweets posted during event time were not mixed with one of the two sets, albeit at the cost of tweets posted right before or after an event.

While a window of a month before and after the date of an event proved useful to calculate the burstiness of an event term, such a period may be too long for the analysis of emotion. Arguably, the felt emotion is stronger when the event is nearer in time. For this reason, we limited the data set to tweets that were posted within three days before and three days after the date of an event.

	#events	#tweets	maximum	median	mean	st. dev.
Pre-event	3,338	3,901,431	128,747	269	1,169	4,782
Post-event	3,338	2,925,069	63,669	216	876	3,080
Total	3,338	6,826,500	192,416	553	2,045	7,159

TABLE 8.1: General statistics of the collected tweets posted within three days before and three days after an event.

We chose to focus on events with a minimum of 50 pre-event tweets and 50 post-event tweets, which we deemed a reliable number to examine the predominant emotion connected to events. In addition, we removed events with over 10% overlapping tweets with other events, in order to avoid possible duplicate content. The resulting number of events and tweets are presented in Table 8.1. In total, the data set comprises of 3,338 events, with a median of 553 tweets per event. In general, more tweets are posted in anticipation of an event than tweets that look back to an event. Furthermore, the mean and median reveal that a small number of events are referred to in lots of tweets and that there is a long tail of events with minor popularity.

To assess the quality of the data set we extracted a random sample of 100 events and for each event randomly selected ten pre-event and ten post-event tweets. One of the authors assessed for all of these events if all pre-event and post-event tweets were linked to it. A third annotation category of partly related tweets was included for cases in which at least seven and at most nine of ten tweets were related to the event.

Of the 100 sets of ten pre-event tweets, 68 were completely related to the event, while thirteen were partly related. Nineteen of them were not related at all or only for a small part. Of the post-event tweets, 70 sets were completely related to the event, while 7 were partly related and 23 were poorly related or not related at all. Based on this evaluation, we can conclude that a substantial part of the event tweets in our data set have a proper connection to the event for which they were queried.

8.4 Emotion Classification

In this section we describe the procedure to train and test models of emotion, based on the approach studied in Part II of this thesis. Central to the approach is the use of hashtags as emotion label for any emotion of interest. In the following, we will motivate the selection of hashtags to model the emotions of interest, and provide a thorough evaluation and analysis of the quality of these models.

Although this procedure can be applied on many different emotions, we will focus here on the emotions of positive expectation, disappointment, and satisfaction. To save space, we will henceforth refer to these emotions as 'PE', 'D' and 'S'. We define them in the following way:

- PE - Anticipatory excitement for a future event. The sender is personally involved with the event: he will attend or follow it. He does not only announce the event, but also expresses in any way that he has positive expectations.
- D - The sender looks back to an event in a negative way. He indicates, implicitly or explicitly, that things have not turned out as hoped or expected.
- S - The sender looks back at an event, indicating that he or she enjoyed it. This is different from someone tweeting about an ongoing pleasant experience.

We explicitly defined S as positively looking back to an experience rather than reporting on an ongoing experience, in line with our aim to measure hindsight emotion. Although we do not include messages that were posted during an event, a sender might nonetheless describe something related to the past event that is currently happening, such as watching a rerun of a television broadcast.

8.4.1 Training models of emotion

In Section 8.2.2 we motivated our approach to train a machine learning classifier on n-grams in hashtag-labeled Twitter messages. An important requirement of this approach is the availability of a sufficient amount of tweets that contain the hashtag. In Chapter 6, we showed that multiple hashtags that convey the same meaning or emotion can be successfully combined to train a classifier to recognise this emotion. The advantage of combining multiple hashtags is that it results in a larger amount of tweets to train on. The hashtags might be manually selected based on their explicit reference to the concept, as we did in Chapter 6, or an initial selection can be expanded based on an empirical procedure (Qadir & Riloff, 2013). We follow the bootstrapping approach by (Qadir & Riloff, 2013) in order to identify hashtags with a good link to the target emotions.

For each of five basic emotions, (Qadir & Riloff, 2013) selected five seed hashtags that fit well to the emotion. They collected tweets that mention one of these hashtags as examples of the target emotion for a machine learning classifier.

Emotion	Seed hashtag	Gloss	# tweets	# random tweets
PE	#zinin	#excited	606,310	606,310
D	#teleurgesteld	#disappointed	11,138	11,138
S	#tevreden	#satisfied	17,459	17,459

TABLE 8.2: Overview of the hashtag and number of tweets on which the initial emotion models were trained.

After stripping away the hashtag itself from the feature space, (Qadir & Riloff, 2013) used these tweets to train a classifier and applied it to a pool of unseen tweets. They used the unseen tweets that were classified with the emotion to extract more emotion hashtags to train on. Each hashtag was ranked by the average classifier confidence that was assigned to the positively classified tweets in which it occurred. The intuition is that the values of these scores give an indication of the link between the hashtag and the target emotion: if the classifier is very certain that the emotion is expressed in tweets with the hashtag, the hashtag likely has a strong link to the emotion and can be used as an additional training label for the emotion. (Qadir & Riloff, 2013) selected the ten hashtags with the highest average score, and repeated the procedure for up to 100 times.

In contrast to (Qadir & Riloff, 2013) we started with only one seed hashtag, the best fitting one, for each of the three target emotions: '#zinin' (#excited) for PE, '#teleurgesteld' (#disappointed) for D, and '#tevreden' (#satisfied) for S. We collected all tweets that contained one of these hashtags from the TwiNL database, in the period from January 2011 until October 2015. We removed tweets in which the target hashtag was not placed at the end, as these are less reliable as emotion label (González-Ibáñez et al., 2011). In addition, we removed retweets to only include messages that were produced by the sender. The number of tweets after collection and filtering is presented in Table 8.2. We also collected a random sample of one million tweets from TwiNL to be used as negative training instances. We made sure that none of these tweets were retweets.

We applied Ucto[3] to tokenise the tweets. User names and URLs were stripped from the tweets, and all characters were lowercased. Punctuation was maintained, as these could be useful clues of emotion. We extracted word unigram, bigram, and trigram features from each tweet and weighted them as Boolean values. Any feature that included a target hashtag was removed from the feature space. Binary classifiers were trained on each of the three selected hashtags, by counterbalancing the hashtag-labeled tweets with an equal amount of random tweets. None of the random tweets contained the target hashtag.

[3] http://languagemachines.github.io/ucto/

Classification was performed by the Balanced Winnow algorithm (Littlestone, 1988). This algorithm is known to offer state-of-the-art results in text classification, and produces interpretable per-class weights that can be used to, for example, inspect the highest-ranking features for one class label. The α and β parameters were set to 1,05 and 0,95 respectively. The major threshold ($\theta+$) and the minor threshold ($\theta-$) were set to 2,5 and 0,5. The number of iterations was bounded to a maximum of six.

As a pool of tweets to extract additional emotion hashtags from, we randomly selected 20% of the events and their tweets from the event dataset described in Section 8.3. This resulted in 3,416,096 pre-event messages and 2,713,961 post-event messages. We applied the classifier trained on #zinin on the pre-event tweets and the classifiers trained on #teleurgesteld and #tevreden on the post-event tweets. We expected that these event-related messages might help to find hashtags that are related specifically to our target emotions.

After classification, the classifier confidence scores for the target emotion class were used to rank the hashtags by their average score. Where in (Qadir & Riloff, 2013) the ten highest ranked hashtags of this list were appended to the existing list of emotion hashtags, we found that a lot of the top-ranked hashtags actually referred to an event or topic. For this reason, we manually inspected the 50 top-ranking hashtags of each of the three classifiers, and selected hashtags that were strongly related to the target emotions manually. These hashtags, as well as the number of additional training tweets that could be collected by querying them from TwiNL, are shown in Table 8.3.

Only a few additional tweets were added to the already extensive training set for PE. On the other hand, the smaller training sets for D and S were expanded considerably, up to 283,399 and 301,420 tweets respectively. Due to the small number of useful hashtags in the ranked list, as well as the sufficient expansion of training tweets, we decided to stop the procedure after one run and train the final emotion classifiers on these collected tweets.

In contrast to the emotion of satisfaction, the other hashtags added as training labels for S seem to have a stronger link to happiness relating to an ongoing event. Nonetheless we expected that these hashtags are a valuable addition for emotion detection on tweets that were posted after an event.

8.4.2 Emotion model evaluation

We applied the same procedure as described in Section 8.4.1 for training the emotion classifiers on the hashtags listed in Table 8.3. We again balanced the

Emotion	Hashtag	Gloss	# tweets (filtered)	Total
PE	#zinin	#excited	606,310	610,576
	#klaarvoor	#readyforit	996	
	#heelveelzinin	#veryexcited	3,270	
D	#teleurgesteld	#disappointed	11,138	283,399
	#zonde	#pity	23,846	
	#balen	#bummer	96,651	
	#spijtig	#deplorable	4,630	
	#jammer	#shame	142,385	
	#teleurstelling	#disappointment	4,749	
S	#tevreden	#satisfied	17,459	301,420
	#blij	#happy	141,276	
	#dankbaar	#grateful	15,835	
	#genieten	#enjoy	126,850	

TABLE 8.3: Overview of the hashtags that were added as training label after classifying a pool of event tweets.

hashtag-labeled tweets with an equal amount of random tweets as training data. We applied the classifiers on a large sample of unlabelled tweets, in order to evaluate their quality in a real-world setting. We used the pool of tweets described in Section 8.4.1 as test set, with 3,416,096 pre-event messages and 2,713,961 post-event messages.

We follow the evaluation procedure applied in Part II of this thesis, assessing classifier performance by means of (a) a hashtag-based evaluation and (b) an evaluation of the top ranked tweets.

Hashtag-based evaluation

A clear indication of tweets that convey the target emotion in the test set are tweets that contain one of the hashtags on which the emotion was trained. During testing these hashtags are masked. Seeing how often the classifier is able to suggest a masked hashtag gives an impression of recall: the test tweets that contain a target hashtag can be seen as a subset of the test tweets that convey the emotion of interest and should be identified as such by the classifier.

In Table 8.4, we present the classifier performance on retrieving tweets that contain one of its target hashtags. For all three emotions, only a small part of the test tweets (about 0.1%) contain one of the target hashtags. In contrast, the classifiers predict the emotion in up to one-third of the test tweets. In line with these proportions, we evaluate performance by reporting the True Positive Rate (TPR, or Recall), False Positive Rate (FPR), and Area Under the Curve (AUC) (Fawcett,

Emotion	Test tweets	With target hashtag	Classified	Correct	TPR	FPR	AUC
PE	3,416,096	4,999	836,289	4,407	0.88	0.24	0.82
D	2,713,961	1,207	893,036	1,008	0.84	0.33	0.75
S	2,713,961	1,764	900,032	1,501	0.85	0.33	0.76

TABLE 8.4: Classifier performance on predicting whether a tweet contains one of a classifiers target hashtags (TPR = True Positive Rate, FPR = False Positive Rate, AUC = Area Under the Curve).

2004). The classifier that was trained to recognise PE manages to retrieve 88% of its target hashtags, while it labels about a quarter of the test tweets with the emotion. The result is an AUC score of 0.82, which is considerably above a score based on random decisions (0.50). The classifiers applied to the post-event test tweets also obtain decent recall scores of 0.84 and 0.85 respectively, but both label one-third of the test tweets with the emotion (i.e. they overpredict the emotion label), resulting in somewhat lower AUC scores of 0.75 and 0.76.

Evaluation of top-ranked tweets

An indication of classifier quality other than hashtag recall is the correctness of the classifier's confidence for an emotion. By ranking the tweets without a target hashtag by classifier confidence it is possible to manually assess whether the targeted emotion is present in the tweets. The outcome can be used to estimate the precision of the classifier at specific ranks. We extracted the top-ranked 250 tweets for each of the three classifiers. Two of the authors and a third annotator assessed for each tweet whether it conveyed the emotion of interest. The annotators had to make a binary decision, based on the definitions that we listed at the beginning of this section.

We evaluate the outcomes by calculating both the precision when two of three annotators labeled the presence of the emotion and when all three annotators did so. We report inter-annotator agreement by the average Cohen's Kappa (J. Cohen, 1960) for every annotator pair. In addition, we calculated the mutual F-score between the decisions of any two annotators, which is robust against class skew.

The precision-at-250 for the three classifiers based on human annotations is given in Table 8.5. For PE, the precision is 0.66 when two of three annotators positively rated a tweet, and 0.49 when all three annotators agreed. The scores for D are lower, with 51% of the top 250 tweets annotated with the emotion by two of the three annotators, while for 31% all three annotators agreed in annotating the emotion. The precision-at-250 for S is lowest, with 0.40 based on a

Emotion	Precision-at-250 66%	100%	Cohen's Kappa	Mutual F-score
PE	0.66	0.49	0.64	0.85
D	0.51	0.31	0.49	0.76
S	0.40	0.25	0.75	0.76

TABLE 8.5: Precision and inter-annotator agreement on the presence of the target emotion in the top ranked 250 tweets by classifier confidence.

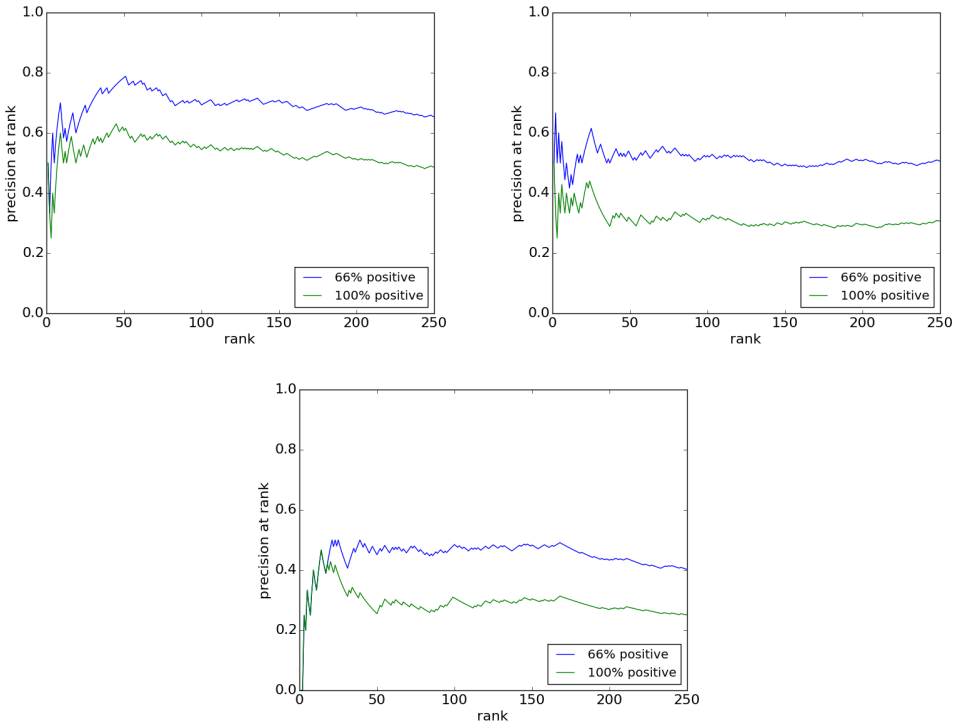


FIGURE 8.1: Precision at $\{1 \ldots 250\}$ on the classes 'Positive expectation' (top left), 'Disappointment' (top right) and 'Satisfaction' (bottom).

66% threshold and 0.25 with a 100% threshold. The Cohen's Kappa agreement of tweets classified as PE is substantial (Landis & Koch, 1977), at 0.64, the agreement on D is moderate, at 0.49 and the agreement on S is substantial, at 0.75. The mutual F-score is highest for PE (0.85), and slightly lower for D and S (both 0.76).

To obtain insights into the influence of tweet rank on precision we generated precision-at plots for the three emotions (see Figure 8.1). The quality of the classifications for PE indeed appears to be related to classifier confidence. The highest precision is reached around rank 50, after which the precision slowly decreases. This is different for D, where the precision is highest at rank 30, followed by a swift decrease. A plateau with minor fluctuations is reached after rank 50. The plot for S shows a decrease of precision after rank 170. The highest

precision is already reached at rank 30.

In conclusion, the evaluation of the emotion models shows that they all yield a decent recall while overshooting considerably in their classifications. The most confident classifications score a moderate precision on the presence of the emotion of interest. In the next section, we analyse the most indicative features of the models, to find out how their decisions come about.

Analysis of models

The Balanced Winnow algorithm returns per-class weights that we used to analyse the models of the classifiers. For the three models of emotion, we inspected the 200 features with the highest weight and divided them into eight categories. Any feature that did not fit into these categories was assigned to a ninth category 'Other'.

- Topic - The name of an event, entity, activity or object.
- Outcome - The outcome of an event or action. Might have an inherent positive ('graduated') or negative ('lost') connotation.
- Evaluation - Judgement of an outcome or topic.
- Exclamation - A word or phrase that expresses an emotion.
- Emo hashtag - A hashtag that explicitly refers to a certain emotion.
- Discrepancy - Word or phrase that expresses a contradiction between the expectation and outcome of an event.
- Conversational - A word or phrase that is explicitly directed to another person.
- Temp - A reference to a point in time.

In Table 8.6, we present the percentages of occurrences of the feature categories for the three models. The model for PE is characterised mostly by features that describe a topic and features that refer to a future point in time. Indeed, we find that PE in tweets is often expressed in relation to an announcement of a future event. As the event has not happened yet, the most common message is simply that the sender is looking forward to it. The features in D mostly comprise a (negative) outcome and an evaluation of the outcome. A small part of the features express a discrepancy with an expectation before the outcome. Features expressing an outcome are also prominent in S. In comparison to the other two

	P		D		S	
	Example	Percent	Example	Percent	Example	Percent
Topic	#ll12	39.5%	hamstring	12.0%	#elclasico	21.0%
Outcome	signed up	10.5%	erased	43.0%	hired	31.5%
Evaluation	#nice	7.5%	#lame	19.5%	#funny	16.0%
Exclamation	#goforit	2.5%	damn	6.0%	yesyesyes	13.5%
Emo hashtag	#curious	3.0%	#sad	3.5%	#relieved	12.0%
Discrepancy	-	0.0%	hoped	2.5%	-	0.0%
Affect	#seeyou	2.0%	#sorry	2.0%	-	0.0%
Temp	#tonight	28.5%	2 Feb	1.0%	#sunday	0.5%
Other	a bit	6.5%	there goes	10.5%	follows me	5.5%

TABLE 8.6: Overview of feature types in the 200 most indicative features per emotion, based on the model trained by Balanced Winnow. '#ll12' refers to Lowlands, a yearly music festival in The Netherlands. '#elclasico' refers to a football match in the Spanish Primera Division, between FC Barcelona and Real Madrid.

emotions, S is characterised mostly by features that express a personal feeling: an evaluation, exclamation or emotion hashtag.

The distribution of feature types in all three classifier models indicate that features describing the conditions for an emotion are as important as features that express the emotion itself. PE is typically related to an event at a future point in time, and this is reflected in the frequency of the 'Topic' and 'Temp' features in the model. A downside of such features is that they are not exclusively linked to PE: a future event might be announced without any emotion. The errors that were found during the annotation of the top 250 tweets were indeed of this type. Likewise, the tweets that were incorrectly classified as D were often characterised by a description of a negative outcome without an explicit mention of disappointment. In contrast, we found that the errors made by the classifier trained on S were often due to the absence of an outcome of a past event. While the aim is to detect tweets that look back to an event with satisfaction instead of tweets that convey satisfaction with the current situation, this difference is not clearly accounted for in the model.

The evaluation of the three classifiers shows that the classifiers recognise the trained emotion in many tweets, while these classifications are not highly accurate. This can be explained by the different types of n-grams that were found to be indicative of the emotion during training. Especially the prominence of n-grams referring to a topic or outcome seems to influence the high number of classifications. Nevertheless we think that the classifiers are sufficiently accurate to be applied to a large amount of event tweets and learn about PE, D and S in relation to social events on Twitter. Although they overshoot in their decision

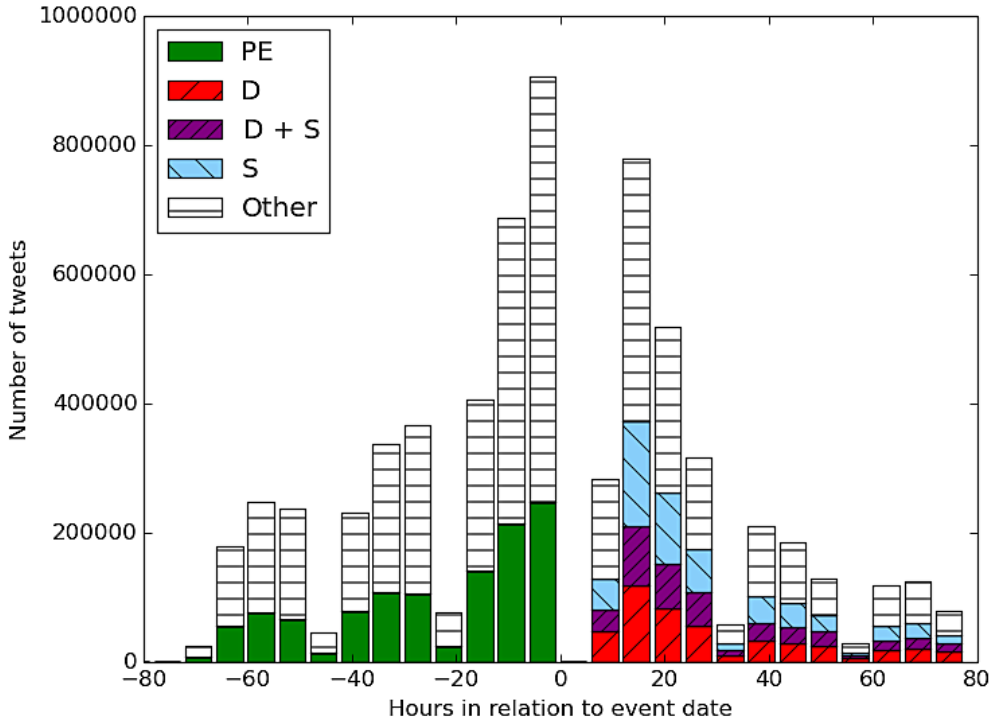


FIGURE 8.2: Overview of classifications of all tweets within three days before and after the date of the referred to event.

for the emotions, they will certainly be sensitive to marked differences in emotions voiced about different events in the set of tweets referring to these events, posted before and after.

8.5 Event Emotion

In this section we discuss the insights obtained after applying emotion classifiers on the set of event tweets described in Section 8.3. We start with an analysis of emotion before and after events in general, and then zoom in on events with distinctive patterns of emotion. Finally, we present a case study of the emotions linked to a sequence of related events: the matches played by the Dutch national football team during the 2014 World Championships.

8.5.1 General patterns

In Figure 8.2 we give an overview of the number of tweets and their classifications in the total set of 3,338 events, by hours before and after the date of the referred to event. Each bar represents a window of six hours and the presented numbers fall within a total window of 72 hours (three days) before and after the event date.

The number of tweets in time shows a clear day-night rhythm, where few tweets are posted during night time and increasingly more tweets are posted during the day. The highest number of tweets is posted one day before and one day after the event date. About one-third of all pre-event tweets are classified as PE. This proportion is seen in most time segments. Hence, positive expectations do not seem dominant at a particular point in time. The classifiers applied on post-event tweets show a similar pattern, labeling about one third of the tweets with the target emotion. Combined, the two emotions D and S make up half of the tweets. Part of the tweets is labeled both as D and S. Upon inspection, we found that the overlap is often due to incorrect decisions by one of the two classifiers. For example, the expression 'fijn' (nice) is sometimes used in a sarcastic way to express D, but could also be classified positively as S.

While Figure 8.2 shows the classifications of all tweets in time, we are most interested in the relation between emotions before and after the same event. To explore this relationship, we calculated the correlation between all three emotion pairs: PE and D, PE and S, and D and S. The latter pair is included for comparison.

The assumption is that all tweets that refer to the same event can be treated as a single unit of which the manifested emotion can be measured. The combination of scores for the three emotions that we target gives an indication of the emotion that an event triggers in the tweeting public, as well as the relation between emotions in general. Importantly, we can not say anything about the intensity of emotion. Rather, the classifications of a collective of tweets reveal the commonality of an emotion. In order to decide on this commonality for PE, D or S, we choose to use two types of information.

The first information type is the percentage of tweets that are classified with a certain label. We can assume that a set of event tweets of which 75% is classified as S represents a higher degree of public satisfaction with the event than a set of event tweets of which only 25% is classified as such. The second information type is the classifier confidence score for an emotion. As shown before, tweets that are more confidently classified with an emotion tend to be more strongly linked to the emotion. The tweets that refer to an event are likely classified with different confidence scores for an emotion label. Rather than calculating the mean or the median to summarise the combined confidence for an emotion, we choose to focus on the 90th percentile: the confidence score of the tweet that is higher than 90% of the confidence scores of the other tweets. This enables us to focus on the tweets in a set that are most indicative of an emotion. Arguably, an event of which 10% of the tweets are confidently classified as D indicates that

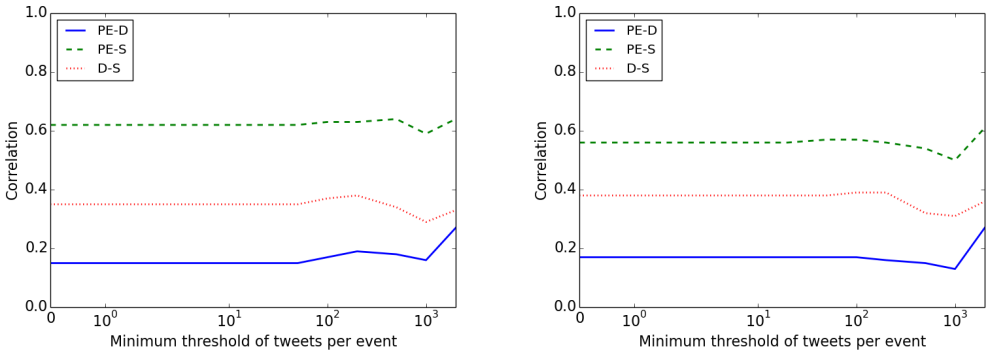


FIGURE 8.3: Correlation between the emotion scores of an event when scoring emotion by the percentage of classifications (left) and by the 90th percentile of classifier confidence for the emotion (right). All reported correlations are significant (< 0.05).

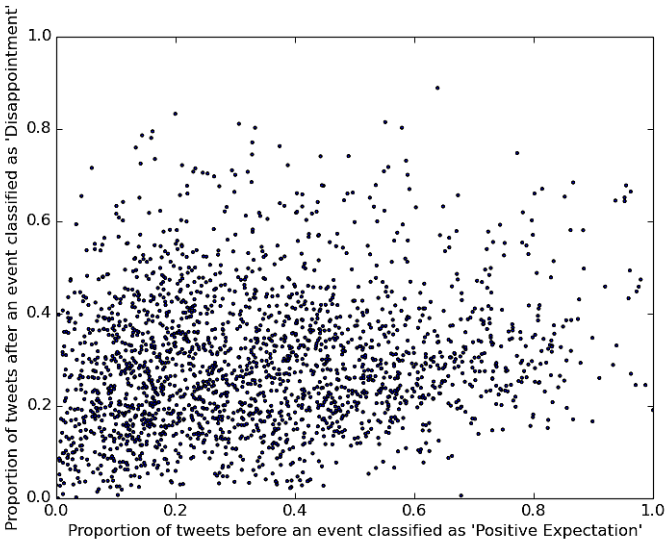


FIGURE 8.4: Scatterplot of degree of PE and D by proportion of classifications for all events with over 100 tweets posted before and after event time.

the event has triggered this emotion, albeit for part of the tweeting public. For a lower percentile of 50 (the median) this pattern would not have been observed. Likewise, the tweet with the highest confidence score might be an outlier.

For each event we calculated the percentage of tweets that were classified with any of the three emotions, as well as the 90th percentile of confidence scores for the three emotions. We calculated the Pearson correlation coefficient for both types of scores for all three emotion pairs. To examine the influence of event popularity on the correlation between emotions, we calculated the correlation for an increasing minimum threshold of tweets. For example, a threshold of 500 filters away any event with fewer than 500 pre-event tweets and/or post-event tweets.

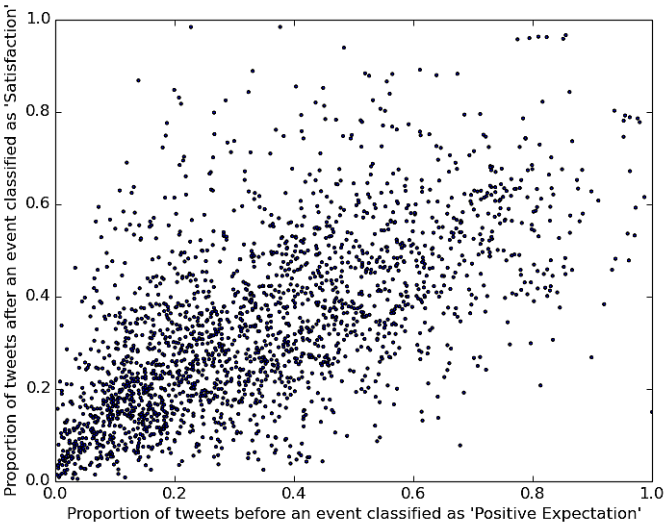


FIGURE 8.5: Scatterplot of degree of PE and S by proportion of classifications for all events with over 100 tweets posted before and after event time.

In Figure 8.3 we display the correlation scores between all emotion pairs, based on the proportion of classifications and the 90th percentile. All scores were found to be significant ($p < 0.05$). As can be seen from the graphs, the largest positive correlation, of around 0.60, exists between PE and S. In contrast, PE and D are very weakly correlated, with values around 0.20. A weak positive correlation also holds between D and S. Apparently, some events evoke strong emotions on both sides. The difference in correlation between the two types of emotion scores, the percentage of classifications and the 90th percentile, is negligible. The correlation between PE and D is somewhat higher when looking at the percentage of classifications. A small effect of the minimum threshold of tweets per event is seen for both types of emotion scores: the correlation is higher when only taking into account the more popular events with over 2000 tweets before and after event time.

In Figure 8.4, 8.5, and 8.6 we display scatterplots of events by their proportion of classifications for PE versus D, PE versus S, and D versus S, respectively. The first scatterplot shows a fuzzy dispersion of points within the values of 0.0 and 0.5 of both emotions. If a high percentage of tweets convey PE, this is as likely followed by a large as by a small number of tweets classified as D. Some events are characterised by a high degree of anticipointment, with a large proportion of both PE and DE. Conversely, some highly disappointing events did not follow high PE, while some events with little disappointment did.

The scatterplot of the percentage of PE and S, in Figure 8.5, shows a more

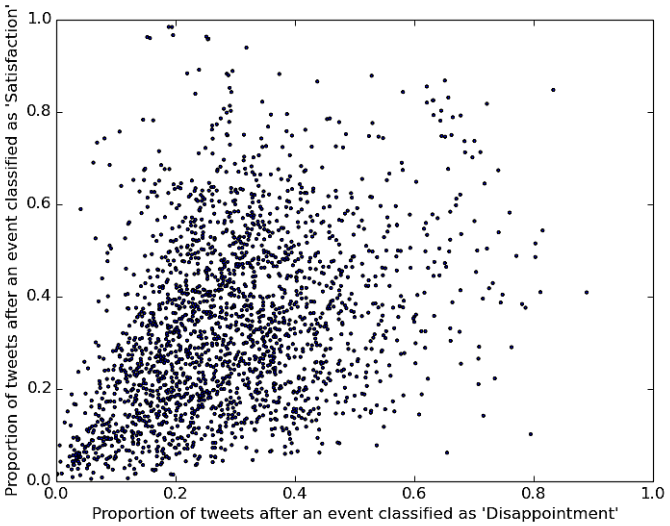


FIGURE 8.6: Scatterplot of degree of D and S by proportion of classifications for all events with over 100 tweets posted before and after event time.

coherent relation between PE and S values in the point cloud. To some extent, the percentage of S can be predicted from the percentage of PE. Finally, the scatterplot in Figure 8.6 shows a correlation that is somewhat clearer than the one in Figure 8.4 (PE and D), but that is not as obvious as in Figure 8.5 (PE and S). The plot is most dense between values of 0.2 and 0.4. Hence if a moderate percentage of tweets display disappointment, a moderate percentage of satisfaction could be seen as well.

In sum, we find that PE before an event does not make D after the event more likely. Higher positive expectations are more often followed by S. In the following, we zoom in on individual events that are prototypical of anticipointment and positive expectation followed by satisfaction.

8.5.2 Event profiles

The emotion scores that are assigned to the total of tweets before and after an event allow us to single out events that display a strong pattern of subsequent emotions. In addition, ranking the tweets most confidently classified with an emotion facilitate an inspection of the causes of PE, D, or S. In this section, we will highlight some of these patterns, and discuss the most exemplary events.

Anticipointing events

Based on the 90th percentile classifier confidence for each emotion, we can quantify the anticipointment of each event and rank them accordingly. We use the

Event	Date	PE	D	S	Anticipointment
#nedden	09/06/12	43.30	25.51	14.71	17.40
Concert at sea	17/06/11	35.83	36.05	20.99	14.95
#neddui	13/06/12	27.83	29.89	14.08	14.74
#ajatwe	24/09/11	29.20	38.81	19.03	14.30
#feytwe	27/01/13	25.74	27.50	13.28	13.31

TABLE 8.7: Top 5 anticipointing events.

following formula to calculate this score:

$$\text{Anticipointment} = \frac{2}{\frac{1}{PE} + \frac{1}{D}} - S \tag{8.1}$$

To avoid a strong influence of either the score for PE or D, we calculated the harmonic mean between them. The score for S is subtracted, in order to make sure that the highest scoring events are mostly disappointing.

The five most anticipointing events are shown in Table 8.7. Most of them are football matches, two of which are disappointing matches of the Dutch squad at the European Championships of 2012 (#nedden and #neddui). Before the first match between the Netherlands and Denmark, #nedden, expectations were very high with a score of 43.30. Expectations for the second match of the Netherlands, against Germany, #neddui, were more tempered, but again the outcome was disappointing. #ajatwe and #feytwe were matches where the team that is favourited by many on Twitter, Ajax and Feyenoord respectively, lost in the end. After the former match, Ajax fans mainly expressed their disappointment with a goal for Ajax that was declined. 'Concert at sea' is a multi-day festival in the Netherlands. The high anticipointment score is due to a canceled day because of stormy weather.

We conclude from these most anticipointing events that a clear loss is at the basis of strong anticipointment on Twitter. Positive expectations relate to a scenario in which the favourable team wins the match, or a fun day at a music festival is expected. Collective disappointment is caused by something preventing this scenario, such as a defeat of the favourite team or a canceled event.

Events with positive expectation followed by satisfaction

We calculated the degree of positive expectation followed by satisfaction in the same way as anticipointment:

$$\text{Anticipated satisfaction} = \frac{2}{\frac{1}{PE} + \frac{1}{S}} - D \tag{8.2}$$

Event	Date	PE	D	S	Anticipated satisfaction
Spring	20/03/12	32.25	11.83	45.55	25.93
#exactlive	02/10/13	33.63	6.30	30.23	25.54
#ad6	05/06/13	32.79	10.54	35.73	23.66
#ad6	07/06/12	32.74	14.95	42.15	21.90
#ad6	09/06/11	31.53	14.53	42.59	21.70

TABLE 8.8: Top 5 events by degree of anticipated satisfaction.

The harmonic mean between the 90th percentile classifier confidence score of PE and S is calculated, and subtracted by the score for D.

The top five of anticipated satisfaction is given in Table 8.8. In contrast to most of the anticipointing events, none of these events are characterised by a sports match. The highest score is obtained by the first day of spring in 2012, which appeared to have been a sunny day. Second is #exactlive, a career event with booths and presentations. The other events are three editions of the yearly charity event 'Alpe d'huzes', in which volunteers cycle up and down the Alpe d'Huez mountain, preferably six times, to collect money for the fight against cancer. The high degree of satisfaction is mostly related to the effort that was spent and the amount of money that was collected.

The highest ranking events by anticipated satisfaction seem to be predominantly characterised by a promotional element. Most tweets that expressed satisfaction after the Exact Live event were posted by organisations that had a booth during the event, and probably wanted to positively emphasise their presence. Likewise, participants of the charity event of Alpe d'huzes were motivated to share their satisfaction with the effort that they spent for a good cause. The high scores for the first day of spring represent an outcome of an uncontrolled condition, the weather, that is widely experienced as positive.

Events with overall high emotion

The weak correlation that was found between D and S (Figure 8.3) showed that the two emotions might be widely expressed in combination for some events. In order to find such occurrences, we rank events by the harmonic mean of all three emotions:

$$\text{Overall emotion} = \frac{3}{\frac{1}{PE} + \frac{1}{D} + \frac{1}{S}} \tag{8.3}$$

Event	Date	PE	D	S	Overall
Pinkpop line-up	19/03/11	51.22	25.64	42.26	36.50
Pinkpop	13/06/11	43.46	28.89	32.98	34.11
Pinkpop	11/06/11	47.59	26.44	34.01	34.06
Pinkpop	12/06/11	45.43	27.13	33.35	33.76
Pinkpop	14/06/13	40.40	26.53	31.21	31.40

TABLE 8.9: Top 5 events on scoring high on all three emotions.

The five events with the highest overall emotion are displayed in Table 8.9. Strikingly, all of the events are related to Pinkpop, a three-day music festival in the Netherlands. All three days of the 2011 edition are included in the top 5, as well as the day at which the line-up is announced. Due to the high status of the performers in the program, Pinkpop is a much anticipated event. The high degree of both disappointment and satisfaction after the event is due to the mixed reception of performances. As music festivals are a collection of many sub-events, the chance for both a high disappointment and satisfaction is substantial. Pinkpop is known for its line-up of predominantly well-known artists, which might be the reason that it scores high are on all three emotions.

8.5.3 Case study

While Sections 8.5.1 and 8.5.2 provide insights from the emotion scores obtained after classification, in this Section we start from a sequence of known events and study whether the classifiers return sensible outcomes. We selected the matches of the Dutch football team during the 2014 World Championships, which we expected to evoke mixed patterns of emotion. We judge the sensibility of the classifications by the known outcomes of these events, in the form of the final score, as well as the contents of the most confidently classified tweets and the values of the emotion scores for these matches in comparison.

The seven matches played by the Dutch squad, along with the 90th percentile scores for PE, D and S, are shown in Figure 8.7. The Netherlands played three first round matches, against Spain, Australia, and Chile, followed by an eighth final match against Mexico, a quarter-final match against Costa Rica, a semi-final match against Argentine and the match for the third place against Brazil.

The scores for the three emotions are generally sensible when considering the nature and outcomes of the events. The highest satisfaction score is seen in relation to #nedspa and #nedmex. The first ended in a surprising 5-1 victory against Spain, the then ruling world champions. The match against Mexico

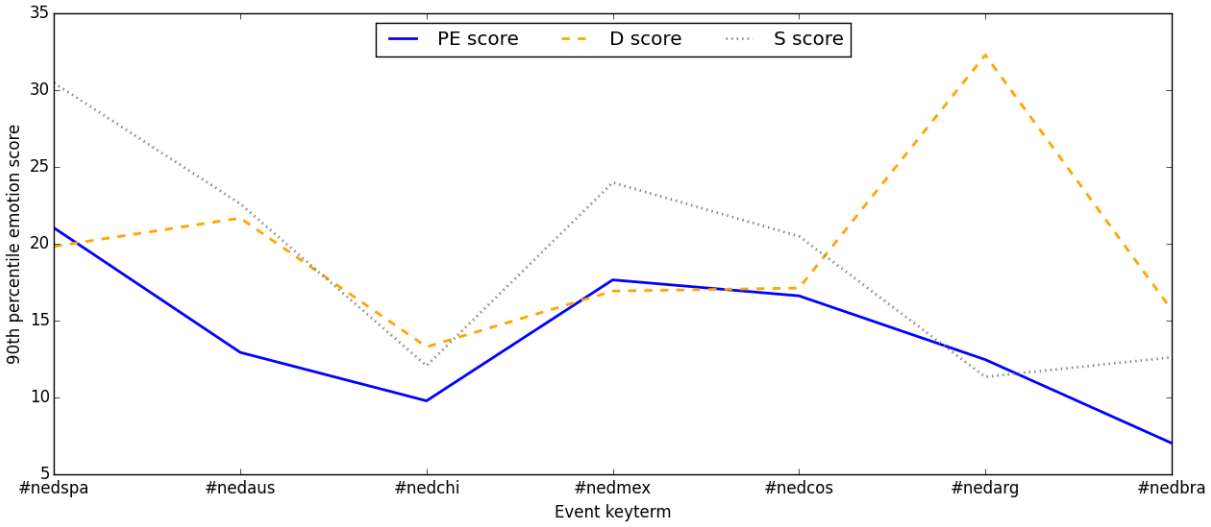


FIGURE 8.7: Overview of 90th percentile emotion scores for all matches of the Dutch football team during the 2014 World Cup, in the order in which they were played.

ended in 2-1 for the Netherlands, very late into the match. Hence, these high scores mark the surprise and relief after these matches. Expectantly, the lost match against Argentine stirred the lowest satisfaction and the highest disappointment. The lowest scores for all three emotions are connected to #nedchi, a less important match in the group stage at which point the Dutch squad was already through to the next round, and #nedbra, the last match with only the third place at stake. Finally, the highest positive expectation is scored before the first match, which relates to the excitement for the world cup campaign to start.

Other outcomes of the emotion classification might seem less intuitive. For example, the disappointment scores after #nedspa and #nedaus stand in contrast to their positive outcomes: a surprising victory and qualification for the next round, respectively. By ranking the tweets after these matches by the classifier confidence score for disappointment, we can inspect the correctness of these classifications and the nature of disappointment. It appears that most tweets indeed express disappointment, but this relates to issues other than the outcome. After the match between The Netherlands and Spain, tweets expressed disappointment in the service at the local pub, the reception of the television provider, the way in which the victory was celebrated by others, the commentary during the match and the analysis after the match. After the second match, disappointment mostly related to the quality of play of the Dutch squad and to this exact disappointment of others despite the qualification for the next round. Another puzzling outcome is the decreasing positive expectation as the Netherlands progress towards the semi-finals. Inspection of the tweets reveals that

an increasing number of tweets express tension about the outcome rather than excitement during anticipation of these matches.

In conclusion, the emotion classifications sketch a sensible story of the performance of the Dutch football team during the 2014 World Cup. This case study shows that our classifiers provide a useful handle to inspect the collective emotion that relates to an event, in the form of an emotion score for all tweets before or after an event, as well as rankings of tweets by the confidence score for an emotion.

8.6 Conclusion and Discussion

We applied classifiers trained to recognise positive expectation, disappointment and satisfaction on a data set that comprised more than three thousand events, automatically extracted from Twitter, with at least 50 forward referring and backward referring tweets. While we expected to find a correlation between positive expectation and disappointment, we observed the strongest correlation, of around 0.60, between positive expectation and satisfaction. The events that are most exemplary of anticipointment are events with a substantial risk of a negative outcome, such as football matches.

The expected correlation between disappointment and positive expectation followed from the insight that the felt disappointment is most likely when expectations are high (Miceli & Castelfranchi, 2014). A possible explanation for the absence of this outcome in our study is that cognitive dissonance might be at play when a much anticipated event appears to be disappointing (Festinger, 1962). After a disappointing experience, the discrepancy with the preliminary high expectations might be made consistent by underlining the good parts or trivialising the positive expectations. It seems easier to tweet about felt satisfaction after issuing positive expectations beforehand, than tweeting about felt disappointment in such a case. Indeed, in our data set the strongest anticipointment can be observed after a clear negative outcome, such as a defeat of the favourable side in a competition or a complete cancellation of an event. A moderate performance of an artist or sports team might stir some disappointment on Twitter, but this is not widespread. The correlation between positive expectation and satisfaction can also be explained by the Pollyanna hypothesis (Boucher & Osgood, 1969), which states that there is a universal human tendency to use positive evaluative words more frequently and diversely than negative evaluative words in communicating. Following this hypothesis, a succession of positive expectation by satisfaction is by default more likely.

It should also be noted that the emotion scores for an event in our study are a reflection of the relative frequency of tweets in which the emotion is detected, as well as the classifier confidence that the emotion is conveyed in these tweets. Based on this information, it is not directly possible to say anything about the *intensity* of the emotions in order to study, for example, whether disappointment is more likely preceded by high expectations than by low expectations.

Although the classifiers that we trained have proven to be useful in comparing events on their pre-event and post-event emotions, evaluation has shown that they yield a sub-optimal performance on detecting the emotion in individual tweets. We find that the classifiers tend to be overly sensitive to the context of an emotion, such as a topic, outcome or an announcement of an event. A possible reason for this is that we contrasted hashtag-labeled emotion tweets with random tweets during training. As a result, a lot of positive classifications were made when we applied the classifiers on tweets that refer to an event, a context that relates to the trained emotion. A higher precision on such data might be obtained by contrasting emotion tweets with random tweets that specifically refer backward or forward to an event.

Another limitation is that the component to automatically extract events from Twitter only retrieves the date of an event. This prevented us from detecting emotion from tweets right before or after event time on the day of the event. Furthermore, our approach to harvest additional tweets for an event was effective for only about three percent of the available events, with a threshold aimed at high precision. The recall may be improved by training a classifier to distinguish event tweets from non-event tweets.

Chapter 9

Conclusions

The past chapters have reported studies on processing the language in a stream of Twitter messages, ultimately leading to the assembly of a system that detects popular events ahead of time and analyses some of their characteristics. In this final section, I will discuss the lessons that can be learned from these studies as well as their contributions, and sketch the avenues of future work that are still open.

9.1 Answers to Research Questions

The first two research questions concern the problem of event detection from tweets. RQ 1 is related to the identification of an event start date when processing tweets that refer to a single event:

> **RQ 1:** Given a stream of tweets that refer to the same event, how accurately and early can we infer the number of days until the start of this event?

In order to infer an event date from tweets, we tested several combinations of settings, features and approaches on 60 football events and subsequently on five events of other types. We list four findings from this study. First, the best performance, close to zero days off, is yielded when employing informed time-based features in a majority-voting set-up. Second, the start time of some events is hardly mentioned in tweets, in which case machine learning based on word n-gram features is a useful alternative. A third finding is that performance improves as an event draws nearer in time. Fourth, it is most beneficial to base a decision on as many tweets as possible, using a sliding window with many tweets and including past estimates. It is important to note, however, that the

inclusion of many tweets in a window comes at the cost of an early estimate. A window of fifty tweets might span all tweets that are posted in anticipation, resulting in a very late or even obsolete estimation of the event date.

Based on these findings, we can conclude that the number of days until the start of an event can be estimated rather precisely from a given stream of event tweets. Overall, the most effective approach is to make use of common-sense rules to infer the time-to-event from temporal markers, and select the most frequently inferred time-to-event. A trade-off exists between accuracy and speed, where it takes a certain amount of tweets, and thus time, to make a reliable time-to-event estimate.

Aiming to detect events from tweets, we subsequently studied the effectiveness of two approaches: burstiness-based event detection and time reference-based event detection.

> **RQ 2a:** To what extent can events be detected from Twitter by means of burstiness-based event detection?

Building on the work by C. Li et al. (2012) and Qin et al. (2013), we found that many bursty terms can be detected in two months of tweets in TwiNL, which we used as our testing grounds. However, after clustering them by similarity, only part of these clusters represented a significant event. The challenge of this approach is to highlight these significant events. We applied a machine learning procedure to make this distinction, which required the effort to manually annotate for a selection of bursty clusters whether they represent a significant event. Manual annotations of a sample of the 33K bursty clusters uncovered about 35% significant events, while the bursty clusters most confidently classified as significant event showed a precision-at-1000 up to 0.80. We found that the number of tweets and the degree of term burstiness are the most useful features to recognise a significant event by. With respect to the output, this approach seemed biased towards news reports that are forwarded on Twitter.

Thus, we found that many bursty terms can be detected from a stream of tweets and clustered by similarity, but that it takes a considerable effort to distill significant events from these. Another point of notice is that the outcome of burstiness-based event detection is effected by several parameter settings, such as the likeliness of a term to be qualified as 'bursty', the time window size at which bursty terms are compared and the conditions for clustering. While we

based these settings on previous literature, a more elaborate answer to this research question would require a comparison of a grid of parameter settings.

> **RQ 2b:** To what extent can significant events be detected from Twitter by means of time reference-based event detection?

To answer this question, we reproduced the work by Ritter et al. (2012) and identified events as entities that link well to a date, as evidenced by tweets that predominantly refer to one specific date while also mentioning an entity. Crucially, this procedure serves to value significant events over mundane events, based on the idea that many entities that represent a mundane event, such as a visit to the dentist, are typically mentioned with a diversity of dates. In contrast, public events are typically referred to in the context of one specific date. We adapted this procedure to Dutch, by manually formulating rules to identify time expressions, and identifying entities as hashtags and concepts on Wikipedia that are commonly linked to on this platform. In line with our research on timely identification of event start dates, we focused on the detection of future events.

Applying the approach to a month of tweets, 87% was assessed by at least two of four human annotators as representing a significant event and 80% by at least three annotators. An additional recall evaluation, where we compared the output to several curated event calendars as gold standard, returned a score of 0.40 on events that were mentioned over five times in the data set. As we focused on future referring time expressions, the detected output predominantly comprised social events such as sports matches, music festivals and holiday celebrations. An error analysis revealed that a common pitfall for the system was to aggregate tweets that refer to different events but share a common entity, such as a city where these events take place in parallel. In addition, similar mundane events obtained a high rank if a lot of them took place on the same date, e.g. carpooling.

Based on these outcomes, we conclude that significant events can be detected sufficiently accurately based on explicit time references. However, in order to provide end users with a sensible and intuitive overview of events, two additional challenges needed to be coped with. The first is the reduction of duplicate events that reside in the output due to the different words or phrases that an event might be referenced with. We managed to resolve part of these duplicates by a clustering procedure. The second challenge is the presentation of events. Mostly, the output did not comply with an ideal picture of event terms

that provide a relevant and adequate event description and tweets that tell a versatile story and exclusively refer to the event. We added several components to improve upon this. Most importantly, we extended the terms that describe an event by selecting additional descriptors from the event tweets. A human evaluation revealed that part of the descriptions were improved by this procedure, but an even higher amount led to redundant information. Apart from this component, we added components to reduce overlapping event terms, present event terms in an intuitive order and rank event tweets by their informativeness.

By experimenting with two essentially different approaches to event detection, one based on burstiness and one based on time references, we can make a comparison between them. Looking at the characteristics of the approaches, time reference-based event detection seems the more elegant of the two, with a minimum of parameter settings. In addition, this approach enables the tracking of events before they take place, while a bursty signal only serves to detect events as they occur. With respect to output, burstiness-based event detection seems biased towards events reported by the news media, while time reference-based event detection mostly results in the detection of social events. The two approaches are complementary: while the output of time reference-based event detection is restricted to events that might be planned, burstiness-based event detection enables the detection of sudden, unplanned events.

In sum, we find that the use of time references in tweets leads to the most accurate and diverse overview of events that will take place in the future, and thereby provides a robust basis to, for example, identify periodic events and detect emotion before and after events.

> **RQ 3:** Given a set of detected events from Twitter over an extended period of time, to what extent can periodic events be identified?

By applying time reference-based event detection to the long span of tweets in TwiNL, we obtained an extensive overview of events from 2011 until 2015. The detection of periodic events from Twitter being an unpaved path, we tested a time interval-based and a calendar-based approach to identify periodically recurring event terms. An assessment of the top-ranked 500 periodic events from the two approaches yielded respective precision scores of 0.63 and 0.76. Wrongly identified periodicities were due to recurring event terms that described a property of an event rather than the event itself. For example, the term 'firecrackers',

recurring for several years on the 31st of December, does not sufficiently describe New Year's Eve. Also, some similarly named events that happened to recur periodically, such as the change in different types of taxes as per July 1 of each year, were mistaken for a periodic event. Arguably, they do represent the recurring governmental change in taxes and rates set on July 1.

The two approaches have complementing strengths, evidenced by the minor overlap of their output. The interval-based approach is lenient to sequences that do not show perfect periodicity. Interestingly, this quality did help the detection of events with a seemingly irregular periodicity, such as Easter. On the other hand, the calendar-based approach is robust against missing or redundant dates in a sequence. The latter approach seems the most favourable of the two. In addition to a better performance, this approach led to the identification of a fairly high number of periodic patterns; a total of 7,018, against 5,301 by the approach based on time intervals. Furthermore, it returns the exact calendar pattern of periodicity, which has proven effective in identifying the future dates of a periodic event.

Based on these findings, we conclude that many periodic patterns can accurately be identified after applying event detection over an extended period of time. The best performance is obtained by searching for recurring calendar features from a sequence of dates.

> **RQ 4:** To what extent can figurative speech or emotion in tweets be detected based on hashtag-annotations?

Hashtag annotations can provide a useful quantity of training data to recognise the textual context of the hashtag, and potentially enable the detection of the figurative speech or emotion that this hashtag refers to. We put this to the test for the tasks of sarcasm detection, by training on a combination of four sarcasm denoting hashtags, and emotion detection, by training on 24 separate hashtags. For each of these 25 targets we trained a binary classifier by contrasting the hashtag-labeled training tweets with an equal amount of random tweets. We employed word n-grams as features and stripped the training hashtag(s) from the feature space as a label.

In a first evaluation, we tested how well each classifier could predict their own training hashtag(s) from a full day of unseen tweets. The classifier that was trained on sarcasm-denoting hashtags yielded an Area under the ROC-curve (AUC) score of 0.85, by correctly identifying 307 of the 353 tweets that carried one

of these hashtags and classifying about 375,000 tweets (17%) in total as sarcasm. The 24 emotion classifiers revealed a mixed hashtag predictability. About half of them retrieved over 75% the tweets that carried their training hashtag, scoring an AUC of over 0.80. We could not observe a clear a priori predictor of hashtag predictability, such as the amount of training data or the emotion that a hashtag links to.

For a fair assessment of the tasks of sarcasm detection and emotion detection, the classification of the tweets that did not contain a training hashtag were of higher interest. As a second evaluation, therefore, we ranked these tweets by classifier confidence and annotated for the top 250 whether they conveyed the figurative speech or emotion that was trained. Applying this more labour-intensive second evaluation to the sarcasm hashtags and a selection of emotion hashtags, we found mixed results. About one-third of the top 250 tweets classified as sarcastic indeed fitted this label. In the remainder of the tweets, the use of highly positive words were confused for sarcasm. A decent performance was yielded for two of four evaluated emotions, anticipatory excitement and self-pity. The emotion of disinterest, for which the training hashtag was well predictable during the first evaluation, was hardly reflected in the most confidently classified 250 tweets. Apparently, Twitter users included this hashtag predominantly to add this emotion, rather than strengthen the emotion reflected in the remainder of words.

In conclusion, we found that hashtags can be exploited to train a classifier for emotion detection, provided that they are more or less consistently deployed in tweets to reflect this emotion as expressed in the words. We could assess this quality based on the performance of hashtag-trained classifiers, but could not find inherent properties of a hashtag that give an a priori estimate of its potential as label for emotion detection.

> **RQ 5:** What is the strength of the correlation between the collective expression of positive expectation before events and disappointment and satisfaction after events on Twitter?

The aforementioned insights on event detection and hashtag-based emotion detection enabled us to research the emotion expressed in tweets before and after a large number of events. We collected tweets that refer to over 3,000 automatically detected events and applied hashtag-based classifiers to detect the tweets that express positive expectation before these events and disappointment and

satisfaction after them. In order to inquire the correlation between these emotions, we scored the classifications of the collective of tweets before and after event time based on the percentage of positive classifications for an emotion. As can be inferred from the title that we chose for the chapter that relates to this research question, we expected to find a correlation between positive expectation and disappointment. However, the outcomes themselves turned out to be an anticipointment, as positive expectations showed a stronger correlation, of around 0.60, with satisfaction than with disappointment. Positive expectation and disappointment were weakly correlated at 0.17. These outcomes might be explained by the possible ease of the Twitter user to share an experience of positive emotion, while there is a weaker inclination to share negative emotion, especially after expressing positive expectations.

Apart from the limited manifestation of anticipointment on Twitter, this study gave insight into the effectivity of our combined approaches to event detection and emotion detection. We found that their scalability made it possible to disclose interesting patterns from the Twitter stream. Both components did leave room for improvement in the context of this application. For instance, a valuable improvement of event detection for this task would be to identify the precise start and end times of events, in order to better distinguish tweets that are posted before, during or after them. Emotion classification might be further improved by more carefully selecting contrasting training instances and including more elaborate features, possibly tuned towards the specific emotion.

9.2 Answer to Problem Statement

> **PS:** How can we discover time-anchored events and their characteristics from a stream of Twitter messages?

Central to our approaches to discover time-anchored events and their characteristics, e.g. periodicity and emotion, is that we make use of contextual anchors. As a distributed information source, characterised by a high number of participants that independently share information, feelings and thoughts, Twitter offers a vast diversity of information. By filtering the platform for information that shares a context of time or a common hashtag, we aimed at a *broad* range of events and words that express figurative speech and certain emotions. These approaches are essentially data-driven. Only the contextual anchors are defined, while the knowledge emerges from the tweets.

Based on our studies, we can conclude that these contextual anchors are indeed feasible to yield the desired diverse outcomes. Sufficiently many tweets explicitly share the time at which an event takes place. This information assisted the timely inference of the date of a given event in Chapter 2, and likewise the detection of events based on consistent mentions of a phrase in combination with the same future day in Chapter 4. In addition, based on the widespread usage of specific hashtags, some emotions could accurately be modelled by using hashtag-labeled tweets as training data.

The high volume of posts on Twitter is hence turned to account by these data-driven approaches, by retrieving a diversity of events and modelling emotions with a diversity of markers. However, this diversity comes at the cost of accuracy, where the contextual anchors lead to a mix of intended and unintended output. In this light we found that the best implementation of the contextual anchors, keeping the amount of unintended output at a low, includes knowledge-driven components. Such components could be tuned more towards the intended output, and led to a higher accuracy. This is most clear for the task of event detection, where we implemented the contextual anchor of time in two ways, as words with a similar time stamp in Chapter 3 and words that are referenced with a similar point in time in Chapter 4. Better results were yielded by the latter variant, which was implemented by a set of manually composed rules to identify time expressions. Explicit time references are more exclusively linked to events than bursty behaviour of words. Likewise, periodic events were best detected by incorporating knowledge of the Gregorian calendar scheme. Knowing that many periodic social events are anchored in this scheme helped to obtain a precision considerably higher than the competing approach based on the consistency of time intervals.

Analysis of the output of our studies suggests that further improvement through knowledge-driven components is possible. For example, a systematic error of time reference-based event detection was to consider the name of a city as an event when it hosted several events on the same date. Such a pattern can be excluded from the output in a post-processing procedure, by recognising names of locations. The same is true for our hashtag-based detection of figurative speech and emotions, which does not incorporate a knowledge-driven component other than the initial selection of hashtags. Specifically, we found that many incorrect classifications were made when the topical context of an emotion was present in a tweet that did not express the emotion itself. Such errors could be corrected by a more selective collection of negative examples to

train the classifier, such as tweets that mention the same topics as the tweets containing the emotion hashtag, but do not convey the emotion itself. In addition, it might be beneficial to scan tweets for the presence of n-grams that explicitly express the emotion, which can be driven by analysis of an initial hashtag-based classifier model.

In conclusion, we found that contextual anchors provide a sufficient data-driven basis to enable a high recall of time-anchored events and their characteristics, while the addition of knowledge-driven components is requisite for going the last mile and further boost the precision of output.

9.3 Thesis Contributions

This thesis offers the following contributions:

1. We replicated work on event detection, sarcasm detection and emotion detection from Twitter to a context of Dutch tweets. Complementary to existing work on burstiness-based event detection, we showed that a Hidden Markov Model could be used to identify bursty terms to be combined into bursty clusters, and gained new insights into the most indicative features for recognising significant events. We added to work on time reference-based event detection by performing a thorough error analysis and adding components to reduce duplicate events in the output and improve on the presentation of events. In comparison to existing work on sarcasm detection and emotion detection from tweets, we took a novel focus by analysing the potential of any hashtag as training label for classification. We also gained insight into the expression of sarcasm and several emotions on Twitter, by inspecting the resulting classification models.

2. We present the first research on identifying the number of days until an event starts from a stream of event tweets, leading to new insight into the temporal information that resides in Twitter messages.

3. We present the detection of periodic events from Twitter as a new task, proposing two approaches and providing insight into their strengths.

4. We conduct the first research on comparing the emotion before and after events as expressed on Twitter. Previous studies on detecting event-related emotion focus on tweets posted during event time, where the targeted events are often manually selected. In contrast, we perform emotion

detection on tweets posted before and after automatically detected events and make a comparison between the patterns in these two sets.

5. We evaluate most of these tasks on all tweets that could be collected in TwiNL within specific but often broad time frames, obtaining a realistic impression of performance when the approaches would be applied to streaming Twitter messages.

6. Based on our studies, we assembled a novel system that processes streaming Dutch tweets to detect events and describe the emotion and periodicity of these events.

9.4 Future work

Although our approaches assemble well into a system of practical use, the studies leave enough avenues for future work and additional functionality. We list the extensions that we see as the most promising.

The event detection component is currently focused on date-anchored events, but we found several multi-day events in the output. For example, all three days in a music festival might be detected as a separate event. It would be valuable to recognise such longer events and link the daily fragments together. In addition, the recall of events can be improved by extending the set of time expressions that are extracted, as event detection is only based on tweets that are found to have a time expression. Finally, event clustering can be improved to further reduce the number of duplicate events. The output of calendar-based periodicity detection can be enhanced by including several types of calendars, such as the lunar and lunisolar calendar in addition to the gregorian calendar, so as to enable the detection of periodic patterns that relate to Easter, the Ramadan, and Hindu festivals, for example. Also, the identification of weekly and monthly periodicity can be increased by chunking sequences of event terms and applying periodicity detection on the chunks. The three classifiers that we applied to label the emotion in pre-event and post-event tweets can be extended with additional emotions, helped by the flexibility of hashtag-annotated emotion detection. In particular, emotions that apply to tweets posted during event time would be a valuable addition.

Two obvious functionalities to enrich the characterisation of events, in addition to the detection of event periodicity and emotion, are to distinguish the location and type of events. Such components would make it possible to narrow the overview of detected events down to events with a specified time frame,

location and type, and cast a warning upon detecting an event with a type of particular interest, such as a social action. Apart from such characterisations, the event description can be contextualised by linking events to related news messages and Wikipedia pages. Another extension worth exploring is to apply burstiness-based event detection to tweets related to separate events in order to detect sub-events before, during and after any main event. Such sub-events can assist the description of separate events, but can also be used to cluster events by typical sub-events and help to understand the dynamics of event types.

The system could be put to use to assist specified user groups. Security services can be helped by receiving an overview of social actions and monitoring whether emotions of aggression are increasing for any event. Journalists might consult the future overview of events to select potential news items based on events that display extreme patterns on Twitter, such as a strong contrast between the emotion before and after event time. Also, rankings such as 'the ten most anticipointing events of the year' might be of interest to them. Tourists can be appealed by providing an overview of events by date, location and type, and listing the events that cause most preliminary excitement. Businesses and event organisers are likely interested in the expectation and reception of the product or event of their competitors and themselves on Twitter.

A final aspect to expand is the data that the system is build on. The TwiNL dataset has provided us with a realistic sample of tweets, but the full Twitter stream would enable a more complete overview of events and event tweets. In addition, the data might be expanded to tweets that are written in different languages than Dutch. The approaches that we apply are largely language-independent, which means that it would take little effort to apply them to other languages as well. For example, event detection would only require the deployment of a proper tagger of time expressions in a language or an effort to manually formulate rules. The Wikipedia version of the given language could be used to extract entities from tweets. Hashtag-based emotion detection might also apply to many languages. Encouragingly, in Chapter 6 we showed that the hashtag '#sarcasme' is deployed in French and Dutch tweets in a similar manner.

Although we are positive about the opportunities for future work, the future of the communication platform on which it is build is uncertain. The worldwide usage of Twitter currently seems stable at 320 Million active users per month,[1]

[1] http://www.adweek.com/socialtimes/heres-how-many-people-are-on-facebook-instagram-twitter-other-big-social-networks/637205

but this number can easily change for the worse, as was demonstrated by a drastic decrease since a peak of over 500 Million users in the summer of 2014.[2] It is therefore sensible to examine the extent to which our studies apply to other platforms, in preemption of a possible shut-down of Twitter. The most prominent microblogging platform apart from Twitter is Sina Weibo[3], although it is mostly connected to a Chinese user base. Social networking platforms like Facebook[4], widely used forum platforms such as Reddit[5] or a diverse blogging platform like Tumblr[6] can also provide rich input to extract information about the world. Our approaches are worth studying on these platforms, where platform-specific features such as emotion stickers in Facebook and subreddits in Reddit show potential as contextual anchors.

To obtain the same kind of output as our current studies, however, there will likely be two difficulties. First, the structure and policy of these platforms prevent access to a more or less open stream of messages. Posts would have to be collected from the networks of specified users or by manually selecting a set of topics, which likely poses a considerable bias on the collected messages. Second, the messages are not bounded by a strict number of characters, like the 140 characters for Twitter. Linking anchors to complete messages might be less successful if messages consist of several sentences. For example, an emotion sticker in Facebook might only apply to the last sentence of a message. Hence, while an extension of our studies to such other types of social media is a logical avenue for future work, we also hope to remain leveraging tweets to extract a continuous snapshot of the world. The access to an open stream as well as the 140 character limit make that Twitter is one of the most suitable platforms for such a pursuit. For this reason, we hope that a considerable amount of people will continue to share their event anticipointment in microblogging format.

[2] http://uk.businessinsider.com/twitter-users-may-be-in-decline-2015-4
[3] http://www.weibo.com/
[4] https://www.facebook.com/
[5] https://www.reddit.com/
[6] https://www.tumblr.com/

Appendices

Appendix A

Instruction letter for the evaluation of burstiness-based event detection (translated from Dutch)

Explanation of annotation task

The units of annotation comprise the output of a system that detects 'bursty' terms in tweets, clusters them together and links them to tweets. Your annotations serve to evaluate this output. For each output, you will answer the following questions:

1. Does the output represent an event?
2. Does the output represent a social event?

Definitions

We define an event as: 'something significant that happens at a specific time and place'. Something is significant if it interests a large group of people, or in other words, if the news media are likely to pay attention to it. Examples of such events are sports matches and transfers of players between sports teams, announcements or actions of a famous person, governmental policies, concerts or performances, television broadcasts, releases of music albums or video games, a natural or humanitarian disaster, etc. Typical output that does not fall under the given definition of an event is a discussion about the hairstyle of Christiano Ronaldo, a conversation between a select company of Twitter users, tweets about playing a video game, tweets that refer to a domestic party, food advises and commercials.

Question number 2 follows our specific interest in social events. A social event can be defined as: 'a scheduled event that is attended by several persons at a specific time and place'. Examples are demonstrations, movie premieres, concerts, football matches and big parties. An event like the Eurovision Songfestival is predominantly followed on television, but this event is held at one specific

time and place and therefore falls within the definition of a social event. Television series are not seen as social event.

Units of annotation and task description

Each row in your annotation file represents an output unit. Column 5 contains one or more terms that describe the output unit. Column 6 contains ten tweets that refer to the output unit. It is advised to broaden both these columns. The tweets in column six are separated by a line break. They can be exposed by double-clicking the cell.

To decide whether the terms and tweets refer to an event, the first question to answer is if the terms and tweets fit together. Terms might describe several events, in which case the output unit does not represent a coherent event. The tweets are included as context to interpret the terms. If the tweets describe different events, or do not describe a clear event, the output unit does not represent an event. Be aware that your task is not to decide whether the tweets and terms match well together, or to assess the extent to which the tweets provide a diverse picture of the event. If many of the tweets are similar in content but do describe one clear event, this is desirable output.

Given an output unit, you are first requested to indicate whether it refers to an event. This can be done in the column titled 'event?', by filling in a '0' (no event), '1' (event) or '2' (doubtful). If you filled in a '1' or '2' in this cell, you are also asked to indicate whether the output is a social event, by filling in a '1' in the column titled 'social?'. If the event is not of a social nature, nothing has to be filled in.

In case of doubt it might be beneficial to consult hyperlinks in tweets, if given.

Examples

North Kenia, grenade attack, refugee camp

331 At least 10 dead persons after grenade attack North Kenia As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 At least 10 dead persons after grenade attack North Kenia CHOROKO As a result of a grenade attack at a refugee camp in Nor URL
331 Massacre in Kenia As a result of a grenade attack at a refugee camp in North-East Kenia on Sunday URL NL_Nieuws_NL

331 RTL Refugee Camp Kenia attacked ten dead As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 At least 10 dead persons after grenade attack North Kenia As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 At least 10 dead persons after grenade attack North Keniaa As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 #nieuws At least 10 dead persons after grenade attack North Kenia As a result of a grenade attack at a refugee camp URL #trouw
331 Refugee Camp Kenia attacked ten dead As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 At least 10 dead persons after grenade attack North Kenia As a result of a grenade attack at a refugee camp in North-East Kenia URL
331 At least 10 dead persons after grenade attack North Kenia As a result of a grenade attack at a refugee camp in North-East Kenia URL

event? 1
social?

ice hockey players, blackhawks, stanley

131 #news Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanle URL #Netherlands
131 #news Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup URL #AD
131 Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup or t. URL
131 Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup URL NL_Nieuws_NL
131 Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup or t. URL
131 Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup or t. URL
131 Stanley Cup Blackhawks need only one more win The ice hockey players of the Chicago Blackhawks are removed one victory URL
131 Stanley Cup in reach for Blackhawks The ice hockey players of the Chicago Blackhawks are close to reaching the Stanley Cup or t. URL

event? 1
social? 1

ehuno, watersports practitioners, kitesurfers, inexperienced, USER, #tatalines

279 RT USER #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL
279 RT USER #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL?
279 #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL #europa
279 #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL #europa
279 KNRM wind too strong for inexperienced watersports practitioners URL
279 RT USER #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL
279 RT USER #Entertainment KNRM wind too strong for inexperienced watersports practitioners Volkskrant VolkskrantKNRM wind too URL?
279 RT USER Is one not allowed a sip of water during ramadan #tatalines
279 RT USER Is one not allowed a sip of water during ramadan #tatalines
279 RT USER Is one not allowed a sip of water during ramadan #tatalines

event? 2 (User en #tatalines do not seem to link to kite surfers)
social? 1

#remainingsportsnews, #darts, police aid, cheaptickets, billion dollar acquisition, samenspender, night bus, hobbyus

129 England Open 2013 Geert De Vos plays semi finals tomorrow URL #darts #remainingsportsnews #sport
129 Peter Wright wins fifth edition of PDC Players Championship URL #darts #remainingsportsnews #sport
129 Sunday Darts Championship 4 located at Ammerzoden URL #darts #remainingsportsnews #sport
129 BDO player Phil Nixon struck by liver, lung and stomach cancer URL #darts #remainingsportsnews #sport
129 Aileen in the finals today at the England Open 2013 URL #darten #remainingsportsnews #sport

129 Bye Adam CS walking to the night bus baibai
129 In the train home again Does anyone know by heart the departure times of the night bus between 2 and 3
129 USER Waiting for the night bus in a breezy bus station Still not rolling Did see a great concert by The Boss
129 USER A night bus departs from Schiphol
129 Achilles C 1 defeats Oranje Wit C 1 during great final at NKD http://t.co/5fvmW35r99 #korfbal #remainingsportsnews #sport

event? 0 (No coherent event seems to underly the terms and tweets)

Appendix B

Rules for the extraction of time expressions

day value	hyphen (optional)	month value	hyphen (optional)	year value (optional)
[1-31]	-	[1-12]	-	20[14-99]
een		januari		
twee		februari		
drie		maart		
vier		april		
vijf		mei		
...		juni		
zevenentwintig		juli		
achtentwintig		augustus		
negenentwintig		september		
dertig		oktober		
eenendertig		november		
		december		

Table A1: Date-related rules for the extraction of time expressions. Values in the columns can be combined sequentially.

indication of future moment	optional part	number of days	time unit	optional part	optional part
over	minimaal	[1-365]	dag(je)	(nog)te	tot
(met)nog	maximaal		dagen	-gaan	
	tenminste		daagjes	slapen	
	bijna		nacht(je)		
	ongeveer		nachtjes		
	maar		nachten		
	slechts		weken		
	pakweg		week(je)		
	ruim		weekjes		
	krap		maand(je)		
	(maar)een		maandjes		
	-kleine		maanden		
	(maar)iets				
	-(meer/minder)				
	-dan				

Table A2: Exact rules for the extraction of time expressions. Values in the columns can be combined sequentially.

time indication (optional)	weekday	part of day (optional)
volgende week	maandag	ochtend
	dinsdag	middag
	woensdag	avond
	donderdag	nacht
	vrijdag	
	zaterdag	
	zondag	

Table A3: Rules for the extraction of time expressions that contain a weekday. Values in the columns can be combined sequentially.

Appendix C

Instruction letter for the evaluation of time reference-based event detection (translated from Dutch)

We have developed a system that fully automatically detects events from the big stream of Dutch tweets. You will test the output of this system. You will judge 50 events in total. This will take about 20 minutes. You can close this survey at any moment and at a later time click the link to repeat the survey. As a start, read the instructions below thoroughly.

You will get to see 5 tweets each time. We ask you to indicate whether they all refer to the same event. To identify an event. you should make use of the following definition:

An event is something that happens at a specific time and is important to a larger group of people.

Sports matches and law amendments qualify as event in this definition, while a holiday to Turkey is too personal to qualify as event.

Warning: sometimes several events are described in a tweet, such as an initiative by the supporters of a football club during a match. If all five tweets indirectly refer to the same football match in this way, they do refer to the same overarching event. However, if five tweets describe different events in the city of Amsterdam, this does not qualify as the same event. These different events are not linked by a common event.

In case of a positive answer, a second question will appear. You will get to see one or more terms that describe the event, and are asked if these terms are a good, moderate or bad representation of the event.

Good luck!

References

Aggarwal, C. C., & Subbian, K. (2012). Event detection in social streams. In J. Ghosh, L. Huan, I. Davidson, C. Domeniconi, & C. Kamath (Eds.), *Proceedings of the 2012 SIAM International Conference on Data Mining* (pp. 624–635).

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37–45). New York, NY, USA: ACM.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, (pp. 579–586). Stroudsburg, PA, USA: ACL.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek & P. Mautner (Eds.), *Proceedings of the 10th Iinternational Conference Text, Speech and Dialogue* (pp. 196–205). Springer.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32*(6), 793–826.

Attardo, S. (2007). Irony as relevant inappropriateness. In R. W. Gibbs, R. W. G. Jr., & H. Colston (Eds.), *Irony in Language and Thought: A Cognitive Science Reader* (pp. 135–170). New York, NY, USA: Lawrence Erlbaum.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Visual markers of irony and sarcasm. *Humor*, *16*(2), 243–260.

Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, *4*(1), 48–53.

Becker, H., Iter, D., Naaman, M., & Gravano, L. (2012). Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 533–542). New York, NY, USA: ACM. doi: 10.1145/2124295.2124360

Bennett-Kastor, T. (1992). Relevance relations in discourse: A study with special reference to Sissala. *Journal of Linguistic Anthropology*, *2*(2), 240–242.

Benson, E., Haghighi, A., & Barzilay, R. (2011). Event discovery in social media feeds.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 389–398). Stroudsburg, PA, USA: Association for Computational Linguistics.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *ICML '06 Proceedings of the 23rd international conference on Machine learning* (pp. 113–120). New York, NY, USA: ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 450–453). Menlo Park, CA, USA: The AAAI Press.

Boucher, J., & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*(1), 1–8.

Bowers, J. W. (1964). Some correlates of language intensity. *Quarterly Journal of Speech*, *50*(4), 415–420.

Brooks, M., Robinson, J. J., Torkildson, M. K., & Aragon, C. R. (2014). Collaborative visual analysis of sentiment in twitter events. In Y. Luo (Ed.), *Proceedings of the 11th International Conference on Cooperative Design, Visualization, and Engineering* (pp. 1–8). Berlin, Germany: Springer-Verlag.

Brown, R. L. (1980). The pragmatics of verbal irony. In R. W. Shuy & A. Shnukal (Eds.), *Language use and the uses of language* (pp. 111–127). Washington, DC, USA: Georgetown University Press.

Bryant, G. A., & Tree, J. E. F. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277.

Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 161–164). Stroudsburg, PA, USA: ACL.

Burgers, C., van Mulken, M., & Schellens, P. J. (2011). Finding irony: an introduction of the verbal irony procedure (vip). *Metaphor and Symbol*, *26*(3), 186–205.

Burgers, C., van Mulken, M., & Schellens, P. J. (2012a). Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics*, *44*(3), 231–242.

Burgers, C., van Mulken, M., & Schellens, P. J. (2012b). Verbal irony differences in usage across written genres. *Journal of Language and Social Psychology*, *31*(3), 290–310.

Carlson, A., Cumby, C., Rosen, J., & Roth, D. (1999). *The SNoW learning architecture* (Tech. Rep. No. UIUCDCS-R-99-2101). Urbana, IL, USA: University of Illinois.

Chakrabarti, D., & Punera, K. (2011). Event summarization using tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 66–73). Menlo

Park, CA, USA: AAAI Press.

Chang, H. (2010). A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, *47*, 1–4.

Chatfield, C. (2013). *The analysis of time series: an introduction*. Boca Raton, FL, USA: CRC press.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1–6.

Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, *126*(3), 1394–1405.

Chen, Y.-S., Argueta, C., & Chang, C.-H. (2015). Emotrend: Emotion trends for events. In M. Renz, C. Shahabi, X. Zhou, & M. A. Cheema (Eds.), *Database systems for advanced applications* (pp. 522–525). Heidelberg, Germany: Springer.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on twitter. In F. Bao, P. Samarati, & J. Zhou (Eds.), *Applied cryptography and network security* (pp. 455–472). Berlin, Germany: Springer-Verlag.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, M. J., van den Brink, G. J. M., Adang, O. M. J., van Dijk, J. A. G. M., & Boeschoten, T. (2013). *Twee werelden, you only live once: Hoofdrapport commissie 'project x' haren.*

Colston, H. L. (2007). What figurative language development reveals about the mind. *Mental States: Volume 2: Language and cognitive structure*, *93*, 191–212.

Cordeiro, M. (2012). Twitter event detection: Combining wavelet analysis and topic inference summarization. In E. Oliveira, G. David, & A. A. Sousa (Eds.), *Doctoral Symposium on Informatics Engineering, DSIE.* Porto, Portugal: Faculdade de Engenharia da Universidade do Porto.

Dann, S. (2010). Twitter content classification. *First Monday*, *15*(12), 1-13.

Davidov, D., Tsur, O., & Rappoport, A. (2010a). Enhanced sentiment learning using twitter hashtags and smileys. In C.-R. Huang (Ed.), *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241–249). Beijing, China: Tsinghua University Press.

Davidov, D., Tsur, O., & Rappoport, A. (2010b). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on*

Computational Natural Language Learning (p. 107). Stroudsburg, PA, USA: ACL.

Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, *1*(1), 7–24.

Dedaić, M. N. (2005). Ironic denial: Tobože in Croatian political discourse. *Journal of pragmatics*, *37*(5), 667–683.

Diao, Q., Jiang, J., Zhu, F., & Lim, E.-P. (2012). Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1* (pp. 536–544). Stroudsburg, PA, USA: ACL.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, *8*(11), e79449.

Eisinger, R. M. (2000). Questioning cynicism. *Society*, *37*(5), 55–60.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. In J. K. Cole (Ed.), *Nebraska symposium on motivation.* Omaha, NE, USA: University of Nebraska Press.

Elfeky, M. G., Aref, W. G., & Elmagarmid, A. K. (2005). Periodicity detection in time series databases. *IEEE Transactions on Knowledge and Data Engineering*, *17*(7), 875–887.

Fan, R., Zhao, J., Feng, X., & Xu, K. (2014). Topic dynamics in weibo: Happy entertainment dominates but angry finance is more periodic. In X. Wu, M. Ester, & G. Xu (Eds.), *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 230–233). Piscataway, NJ, USA: IEEE.

Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for researchers* (Tech. Rep. No. HPL-2003-4). Palo Alto, CA, USA: HP Laboratories.

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Redwood City, CA, USA: Stanford university press.

Forney, G., & David, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Fraisse, A., & Paroubek, P. (2014). Twitter as a comparable corpus to build multilingual affective lexicons. In P. Zweigenbaum, S. Sharoff, R. Rapp, A. Aker, & S. Vogel (Eds.), *The 7th Workshop on Building and Using Comparable Corpora* (pp. 26–31).

Fung, G. P. C., Yu, J. X., Yu, P. S., & Lu, H. (2005). Parameter free bursty events detection in text streams. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-Å. Larson, & B. C. Ooi (Eds.), *Proceedings of the 31st International Conference on Very Large Data Bases* (pp. 181–192). New York, NY, USA: ACM.

Gayo-Avello, D. (2012). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" – A Balanced Survey on Election Prediction using Twitter Data. *CoRR*, *abs/1204.6441*.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, *115*(1), 3–15.

Gibbs, R. W. (2007). On the psycholinguistics of sarcasm. In R. W. Gibbs, R. W. G. Jr., & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 173–200). New York, NY, USA: Lawrence Erlbaum.

Gibbs, R. W., & Colston, H. (2007). Irony as persuasive communication. In R. W. Gibbs, R. W. G. Jr., & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 581–595). New York, NY, USA: Lawrence Erlbaum.

Gibbs, R. W., & O'Brien, J. (1991). Psychological aspects of irony understanding. *Journal of pragmatics*, *16*(6), 523–530.

Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford, UK: Oxford University Press.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1–12.

Goddard, C. (2006). "Lift your game Martina!": Deadpan jocular irony and the ethnopragmatics of Australian English. *Applications of Cognitive Linguistics*, *3*, 65–97.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 581–586). Stroudsburg, PA, USA: Association for Computational Linguistics.

Grice, H. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Pragmatics: syntax and semantics* (pp. 113–127). New York, NY, USA: Academic Press.

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD, USA: StatAxis Publishing Company.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2), 107–145.

Holtgraves, T. (2005). Social psychology, cognitive psychology, and linguistic politeness. *Journal of Politeness Research. Language, Behaviour, Culture*, *1*(1), 73–93.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A practical guide to support vector classification* (Tech. Rep.). Taipei, Taiwan: Department of Computer Science, National Taiwan University.

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (pp. 173–178). New York, NY, USA: ACM.

Hürriyetoğlu, A., Kunneman, F., & van den Bosch, A. (2013). Estimating the time between twitter messages and future events. In C. Eickhoff & A. P. de Vries (Eds.), *Proceedings of the 13th Dutch-Belgian Workshop on Information Retrieval* (pp. 20–23). Delft, The Netherlands: Delft University of Technology.

Hürriyetoğlu, A., Oostdijk, N., & van den Bosch, A. (2014). Estimating time to event from tweets using temporal expressions. In *Proceedings of the 5th Workshop on Language*

Analysis for Social Media (LASM) (pp. 8–16). Stroudsburg, PA, USA: ACL.

Jackoway, A., Samet, H., & Sankaranarayanan, J. (2011). Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 25–32). New York, NY, USA: ACM.

Jahandarie, K. (1999). *Spoken and written discourse: A multi-disciplinary perspective*. Santa Barbara, CA, USA: Greenwood Publishing Group.

James, G., & Hastie, T. (1998). The Error Coding Method and PICTs. *Journal of Computational and Graphical Statistics*, *7*(3), 377–387.

Jarvis, R. A., & Patrick, E. A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, *100*(11), 1025–1034.

Kamvar, S. D., & Harris, J. (2011). We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 117–126). New York, NY, USA: ACM.

Kim, S., Bak, J., & Oh, A. H. (2012). Do you feel what I feel? social aspects of emotions in twitter conversations. In *Proceedings of the Sixth International Conference on Weblogs and Social Media.* Palo Alto, CA, USA: The AAAI Press.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*(4), 373–397.

Kreuz, R., Roberts, R., Johnson, B., & Bertus, E. (1996). Figurative language occurrence and co-occurrence in contemporary literature. In R. Kreuz & M. MacNealy (Eds.), *Empirical approaches to literature and aesthetics* (pp. 83–97). Norwood, NJ, USA: Ablex.

Kreuz, R. J., & Roberts, R. M. (1993). The empirical study of figurative language in literature. *Poetics*, *22*(1), 151–169.

Kreuz, R. J., & Roberts, R. M. (1995). Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and symbol*, *10*(1), 21–31.

Kumar, S., Liu, H., Mehta, S., & Venkata Subramaniam, L. (2014). From tweets to events: Exploring a scalable solution for twitter streams. *CoRR*, *abs/1405.1392*.

Kunneman, F., & van den Bosch, A. (2012). Leveraging unscheduled event prediction through mining scheduled event tweets. In J. W. Uiterwijk, N. Roos, & M. H. Winands (Eds.), *Proceedings of the 24th Benelux Conference on Artificial Intelligence* (pp. 147–154). Maastricht, The Netherlands.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Larsen, M., Boonstra, T., Batterham, P., O'Dea, B., Paris, C., & Christensen, H. (2015). We feel: Mapping emotion on twitter. *IEEE journal of biomedical and health informatics*, *19*(4), 1246–1252.

Leigh, J. H. (1994). The use of figures of speech in print ad headlines. *Journal of Advertising*, 17–33.

Li, C., Sun, A., & Datta, A. (2012). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 155–164). New York, NY, USA: ACM.

Li, Y., Wang, X. S., & Jajodia, S. (2001). Discovering temporal patterns in multiple granularities. In *Proceedings of the First International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers* (pp. 5–19). London, UK: Springer-Verlag.

Liebrecht, C. (2015). *Intens krachtig. stilistische intensiveerders in evaluatieve teksten* (Doctoral dissertation, Radboud University Nijmegen). Retrieved from `http://hdl.handle.net/2066/141116`

Liebrecht, C., Kunneman, F., & van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 29–37). Stroudsburg, PA, USA: ACL.

Liew Suet Yan, J. (2015). Discovering emotions in the wild: An inductive method to identify fine-grained emotion categories in tweets. In I. Russell & W. Eberle (Eds.), *Proceedings of the twenty-eighth international florida artificial intelligence research society conference* (pp. 317–323). Menlo Park, CA, USA: The AAAI Press.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, *2*, 285–318.

Livnat, Z. (2004). On verbal irony, meta-linguistic knowledge and echoic interpretation. *Pragmatics & Cognition*, *12*(1), 57–70.

Mahanta, A. K., Mazarbhuiya, F. A., & Baruah, H. K. (2008). Finding calendar-based periodic patterns. *Pattern Recognition Letters*, *29*(9), 1274–1284.

McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd acm international conference on conference on information & knowledge management* (pp. 409–418). New York, NY, USA: ACM.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889–892). New York, NY, USA: ACM.

Meij, E., Weerkamp, W., & de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 563–572). New York, NY, USA: ACM.

Miceli, M., & Castelfranchi, C. (2014). *Expectancy and emotion*. Oxford, UK: Oxford University Press.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Proceedings of the Fifth International*

Conference on Weblogs and Social Media (pp. 554–557). Menlo Park, CA, USA: The AAAI Press.

Mizzau, M. (1984). *L'ironia: la contraddizione consentita*. Milan, Italy: Feltrinelli.

Mohammad, S. M. (2012). #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 246–255). Stroudsburg, PA, USA: ACL.

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301–326.

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, *53*(4), 675–679.

Muecke, D. C. (1969). *The compass of irony*. Oxford, UK: Oxford University Press.

Muecke, D. C. (1978). Irony markers. *Poetics*, *7*(4), 363–375.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses: An introduction*. Hoboken, NJ, USA: Wiley-Interscience.

Ou, G., Chen, W., Wang, T., Wei, Z., Li, B., Yang, D., & Wong, K.-F. (2014). Exploiting community emotion for microblog event detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1159–1168). Stroudsburg, PA, USA: ACL.

Ozdikis, O., Senkul, P., & Oguztuzun, H. (2012). Semantic expansion of hashtags for enhanced event detection in twitter. In *The First International Workshop on Online Social Systems (WOSS 2012).*

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics (naacl hlt 2010)* (pp. 181–189). Stroudsburg, PA, USA: ACL.

Petrovic, S., Osborne, M., & Lavrenko, V. (2013). I wish I didn't say that! analyzing and predicting deleted messages in twitter. *CoRR*, *abs/1305.3107*.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York, NY, USA: Harper & Row.

Preoţiuc-Pietro, D., & Cohn, T. (2013). A temporal model of text periodicities using gaussian processes. In *Proceedings of the 2013 Conference on Empirical Methods in*

Natural Language Processing (pp. 977–988). Stroudsburg, PA, USA: ACL.

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482–491). Stroudsburg, PA, USA: ACL.

Qadir, A., & Riloff, E. (2013). Bootstrapped learning of emotion hashtags #hashtags4you. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 2–11). Stroudsburg, PA, USA: ACL.

Qin, Y., Zhang, Y., Zhang, M., & Zheng, D. (2013). Feature-rich segment-based news event detection on twitter. In *Proceedings of the sixth international joint conference on natural language processing* (pp. 302–310). AFNLP.

Quezada, M., & Poblete, B. (2013). *Understanding Real-World Events via Multimedia Summaries Based on Social Indicators*. Berlin, Germany: Springer-Verlag.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846–850.

Reuter, T., & Cimiano, P. (2012). Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (p. 22). New York, NY, USA: ACM.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, *47*(1), 239–268.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 172–180). Stroudsburg, PA, USA: ACL.

Ritter, A., Clark, S., & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524–1534). Stroudsburg, PA, USA: ACL.

Ritter, A., Mausam, Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104–1112). New York, NY, USA: ACM.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). Empatweet: Annotating and detecting emotions on twitter. In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3806–3813). Istanbul, Turkey: European Language Resources Association (ELRA).

Rockwell, P. (2003). Empathy and the expression and recognition of sarcasm by close relations or strangers. *Perceptual and motor skills*, *97*(1), 251–256.

Rockwell, P. (2007). Vocal features of conversational sarcasm: A comparison of methods. *Journal of psycholinguistic research*, *36*(5), 361–369.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ, USA: Prentice-Hall.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860). New York, NY, USA: ACM.

Sanders, E., & van den Bosch, A. (2013). Relating political party mentions on twitter with polls and election results. In K. Eickhoff & A. de Vries (Eds.), *Proceedings of the 13th dutch-belgian workshop on information retrieval* (pp. 68–71). Delft, The Netherlands: Delft University of Technology.

Sappelli, M., Verberne, S., & Kraaij, W. (2013). Combining textual and non-textual features for e-mail importance estimation. In K. Hindriks, M. de Weerdt, B. van Riemsdijk, & M. Warnier (Eds.), *Proceedings of the 25th benelux conference on artificial intelligence* (pp. 147–154). Delft, The Netherlands: Delft University of Technology.

Sethares, W. A., & Staley, T. W. (1999). Periodicity transforms. *IEEE Transactions on Signal Processing*, *47*(11), 2953–2964.

Siegel, S., & Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. New York, NY, USA: McGraw Hill.

Sintsova, V., Musat, C.-C., & Pu Faltings, P. (2013). Fine-grained emotion recognition in olympic tweets based on human computation. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 12–20). Stroudsburg, PA, USA: ACL.

Snow, R., Jurafsky, D., & Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17 (NIPS 2004)* (pp. 1297–1304). Cambridge, MA, USA: MIT Press.

Srinarawat, D. (2005). Indirectness as a politeness strategy of Thai speakers. In R. Lakoff & S. Ide (Eds.), *Broadening the horizon of linguistic politeness* (pp. 175–193). Amsterdam, The Netherlands: John Benjamins.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 525–526). New York, NY, USA: ACM.

Strötgen, J., & Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 321–324). Stroudsburg, PA, USA: ACL.

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY, USA: Anchor Books.

Suttles, J., & Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 121–136). Berlin, Germany: Springer-Verlag.

Sykora, M. D., Jackson, T., O'Brien, A., Elayan, S., & Von Lunen, A. (2014). Twitter based analysis of public, fine-grained emotional reactions to significant events. In *The Proceedings of the European Conference on Social Media* (pp. 540–548). Brighton, UK: University of Brighton.

Tao, K., Abel, F., Hauff, C., Houben, G.-J., & Gadiraju, U. (2013). Groundhog day: near-duplicate detection on twitter. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1273–1284). New York, NY, USA: ACM.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Thelwall, M., & Kappas, A. (2014). The role of sentiment in the social web. In C. Von Scheve & M. Salmela (Eds.), *Collective emotions: Perspectives from psychology, philosophy, and sociology* (pp. 375–388). Oxford, UK: OUP Oxford.

Tjong Kim Sang, E. (2011). Het gebruik van twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap*, *39*(1/2), 62–72.

Tjong Kim Sang, E., & van den Bosch, A. (2013). Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal*, *3*, 121-134.

Tops, H., van den Bosch, A., & Kunneman, F. (2013). Predicting time-to-event from twitter messages. In K. Hindriks, M. de Weerdt, B. van Riemsdijk, & M. Warnier (Eds.), *Proceedings of the 25th benelux conference on artificial intelligence* (pp. 207–2014). Delft, The Netherlands: Delft University of Technology.

Torkildson, M. K., Starbird, K., & Aragon, C. (2014). Analysis and visualization of sentiment and emotion on crisis tweets. In *International conference on cooperative design, visualization and engineering* (pp. 64–67). Springer International Publishing.

Tsur, O., Davidov, D., & Rappoport, A. (2010). A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 162–169). Menlo Park, CA, USA: The AAAI Press.

Valkanas, G., & Gunopulos, D. (2013). How the live web feels about events. In *Proceedings of the 22nd ACM international conference on Conference on Information & Knowledge Management* (pp. 639–648). New York, NY, USA: ACM.

van Boven, L., & Ashworth, L. (2007). Looking forward, looking back: anticipation is more evocative than retrospection. *Journal of Experimental Psychology: General*, *136*(2), 289–300.

van den Bosch, A. (2004). Wrapped progressive sampling search for optimizing learning algorithm parameters. In R. Verbrugge, N. Taatgen, & L. Schomaker (Eds.), *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence* (pp. 219–226). Groningen, The Netherlands.

van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for dutch. In *Computational*

linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting (pp. 99–114).

van der Löwe, I., & Parkinson, B. (2014). Relational emotions and social networks. In C. Von Scheve & M. Salmela (Eds.), *Collective emotions: Perspectives from psychology, philosophy, and sociology* (pp. 125–140). Oxford, UK: OUP Oxford.

van Dijk, W. W., Zeelenberg, M., & van der Pligt, J. (2003). Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment. *Journal of Economic Psychology*, *24*(4), 505–516.

van Mulken, M., & Schellens, P. J. (2012). Over loodzware bassen en wapperende broekspijpen. gebruik en perceptie van taalintensiverende stijlmiddelen. *Tijdschrift voor taalbeheersing*, *34*(1), 26–53.

Wang, D., Abdelzaher, T., & Kaplan, L. (2015). *Social sensing: building reliable systems on unreliable data*. Burlington, MA, USA: Morgan Kaufmann.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)* (pp. 587–592). Los Alamitos, CA, USA: IEEE CPS.

Wang, X., & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 424–433). New York, NY, USA: ACM.

Wasserman, A. I. (1980). Information system design methodology. *Journal of the American Society for Information Science*, *31*(1), 1–24.

Weerkamp, W., & de Rijke, M. (2012). Activity prediction: A twitter-based exploration. In *SIGIR 2012 Workshop on Time-aware Information Access: #TAIA2012. Accepted papers.* Microsoft Research.

Weiler, A., Scholl, M. H., Wanner, F., & Rohrdantz, C. (2013). Event identification for local areas using social media streaming data. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks* (pp. 1–6). New York, NY, USA: ACM.

Weng, J., & Lee, B.-S. (2011). Event detection in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 401–408). Menlo Park, CA, USA: The AAAI Press.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79.

Yang, T., Lee, D., & Yan, S. (2013). Steeler nation, 12th man, and boo birds: classifying Twitter user interests using time series. In T. Özyer, P. Carrington, & E.-P. LIM (Eds.), *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 684–691). New York, NY, USA: ACM.

Yoos, G. E. (1985). The rhetoric of cynicism. *Rhetoric Review*, *4*(1), 54–62.

Zhang, M., Kao, B., Cheung, D. W., & Yip, K. Y. (2007). Mining periodic patterns with gap requirement from sequences. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(2), 7.

Zhao, S., Zhong, L., Wickramasuriya, J., & Vasudevan, V. (2011). *Human as Real-Time sensors of social and physical events: A case study of Twitter and sports games* (Tech. Rep. No. TR0620-2011). Houston, TX, USA: Rice University and Motorola Labs.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough et al. (Eds.), *Advances in Information Retrieval* (pp. 338–349). Berlin, Germany: Springer.

Zhou, X., & Chen, L. (2013). Event detection over Twitter social media streams. *The VLDB Journal*, *23*(3), 381-400.

Samenvatting

Wanneer mensen samenkomen om te vieren, demonstreren of vermaakt te worden laten ze vaak indrukken over deze gebeurtenissen achter door erover te communiceren op sociale media. Deze dissertatie bundelt een aantal studies die als doel hebben om dit soort indrukken zoals ze op Twitter worden achtergelaten automatisch te identificeren en te interpreteren. De uitkomsten van deze studies zijn geïntegreerd in een informatiesysteem dat inzichten biedt in gebeurtenissen vanuit het gezichtspunt van het twitterende publiek. Een deel van de studies sluit aan op ook elders onderzochte taken, zoals het detecteren van gebeurtenissen, stijlfiguren en emoties in Twitterberichten. Daarnaast verkent deze dissertatie nieuwe terreinen van onderzoek, door gedetecteerde gebeurtenissen te analyseren op de aanwezigheid van periodiciteit en de relatie tussen emoties ervoor en erna. De studies kunnen onderverdeeld worden in de eenheid van informatie waar ze op gebaseerd zijn: tijd en hashtags.

Veel van de woorden in tweets zeggen iets over tijd, hetzij impliciet of expliciet. Hoofdstukken 2, 3 en 4 beschrijven studies naar de waarde van zulke informatie voor het detecteren van gebeurtenissen en het tijdstip waarop ze plaatsvinden. Hoofdstuk 2 beschrijft een studie naar de taal in Twitterberichten die vooruitverwijzen naar een wedstrijd in de Eredivisie, om de relatie te onderzoeken tussen de woorden in een bericht en het aantal dagen tot een wedstrijd plaatsvindt. Expliciete verwijzingen naar de datum of het aantal dagen tot de wedstrijd blijken het meest waardevol te zijn om automatisch de datum in te schatten, maar wanneer zulke informatie ontbreekt blijken ook inhoudswoorden van waarde. Als men tweet over het kopen van een kaartje is de wedstrijd bijvoorbeeld gemiddeld nog vier dagen verwijderd. De studies in hoofdstuk 3 en 4 hebben beiden tot doel om automatisch gebeurtenissen van velerlei typen te detecteren uit Twitter, en doen dit vanuit een verschillend uitgangspunt. De eerste studie scant de open stroom van tweets voor woorden die plotseling sterk in aantal stijgen, wat mogelijk verwijst naar een gebeurtenis die gerelateerd is aan dat woord. De tweede speurt Twitterberichten af voor de aanwezigheid van tijdsexpressies, en beschouwt het vaak voorkomen van entiteiten

(zoals eigennamen) met één specifieke toekomstige dag in tweets als een sterke aanwijzing voor een gebeurtenis. Deze laatste methode bleek het meest effectief voor het detecteren van sociale events, waaronder muziekfestivals, de uitgave van nieuwe producten en nationale feestdagen. De studie in hoofdstuk 5 maakt gebruik van deze methode als tussenstap om automatisch periodiek voorkomende gebeurtenissen te identificeren, daarin geholpen door een archivering van Twitterberichten die meer dan vier jaar teruggaat. Het speuren naar repetitieve kalenderpatronen in terugkerende gebeurtenissen, zoals de tuinvogeltelling op elke derde zondag van januari, is daarbij het meest effectief gebleken.

Het tweede deel van de dissertatie rapporteert over het gebruik van hashtags om de aanwezigheid van stijlfiguren en emoties te voorspellen in tweets. Hashtags zoals '#sarcasme' en '#leuk' worden door Twittergebruikers ingezet als ruimtebesparende manier om het sentiment van hun bericht te verduidelijken. Zelfs wanneer deze context al begrepen kan worden uit de woorden die in het bericht staan, wordt een dergelijke hashtag soms toegevoegd om miscommunicatie uit te sluiten of de al aanwezige emotie te versterken. Als deze laatste toepassing vaak voorkomt kunnen hashtags van waarde zijn als anker voor het verzamelen van tweets die de taal van een stijlfiguur of emotie bevatten. Door een *machine learning classifier* te trainen met zulke voorbeelden, vaak tienduizenden of zelfs honderdduizenden, kan deze de stijlfiguur of emotie gaan detecteren in tweets zonder de hashtag. In hoofdstuk 6 is de effectiviteit van deze methode getoetst voor de taak van sarcasmedetectie. Het trainen van een model op basis van sarcasme-gerelateerde hashtags toonde aan dat met name positieve markeerders, zoals 'geweldig' en 'charmant', een sterke indicator zijn van sarcasme. Dit model bleek slechts ten dele effectief in het detecteren van sarcastische tweets, doordat het zeer positieve berichten vaak aanzag voor sarcastisch. De procedure als toegepast in hoofdstuk 6 is in hoofdstuk 7 uitgebreid naar een veelvoud aan hashtags die een emotie beschrijven. Deze studie toonde een wisselend succes, afhankelijk van hashtag. Sommige hashtags, zoals '#omg' (afkorting voor 'oh my god'), werden met verschillende communicatieve doelen ingezet door de Twittergebruiker, waardoor hier geen eenduidig model op getraind kon worden. Andere hashtags toonden een consistenter gebruik, maar dit bleek voornamelijk het toevoegen van de emotie en niet het versterken van de reeds aanwezige emotie. Een voorbeeld van een hashtag die wel vruchtbaar bleek was '#zinin', waarmee de emotie van positief vooruitkijken succesvol gemodelleerd kon worden.

De twee richtingen van onderzoek, gerelateerd aan gebeurtenissen en emotie, komen samen in de laatste studie die beschreven wordt in hoofdstuk 8: het

detecteren van uitingen van *anticipointment* (of *anticipeleurstelling* in het Nederlands). Dit concept benadrukt het contrast tussen positieve verwachting vooraf en teleurstelling achteraf. De mate van *anticipointment* werd in kaart gebracht door eerst automatisch sociale gebeurtenissen te detecteren, om vervolgens met behulp van hashtags te detecteren welke tweets voor en na deze gebeurtenissen een uiting zijn van positieve verwachting, tevredenheid of teleurstelling. De studie toonde aan dat de Twittergebruiker in het algemeen de neiging heeft om positief te tweeten over gebeurtenissen, zowel in de aanloop ernaartoe als na afloop. De zeldzame gevallen waarin anticipointment wel voorkwam werden gekenmerkt door een sterk gevoel van onrecht, zoals bij een oneerlijk ervaren verlies van het favoriete sportteam of bij een annulering van een optreden. Gecombineerd geven deze studies inzicht in het Twitter landschap en de verscheidenheid aan toepassingen voor het detecteren van gebeurtenissen en emoties in tweets.

Summary

As people come together to celebrate, demonstrate or be entertained, they often leave traces about these events through communicating via online social media. The studies reported in this thesis are aimed at the automatic identification and interpretation of such traces in the Dutch Twitter sphere. The outcomes are integrated into an information system that provides insight into real-world events and into the contemplations of their crowds. The studies connect to existing work on detecting events, figurative speech or emotion in Twitter messages. The thesis covers new ground in analysing detected events for patterns of periodicity and for patterns of prior and subsequent emotion. Two prominent units of information in tweets are at the basis of all studies: time and hashtags.

Many words in Twitter messages reveal temporal information, either implicitly or explicitly. Chapter 2, 3 and 4 study the value of such information for detecting events and the time at which they will take place in the future. In Chapter 2, Twitter messages that refer to a future football match are isolated in order to analyse the relation between the words that are used and the number of days until a match takes place. Explicit references to the start time are most valuable for automatically estimating the time-to-event, but topical mentions appeared indicative as well. For example, tweets tend to report on purchasing tickets when the match is about four days in the future. The studies described in Chapter 3 and 4 work with different signals to detect open-domain events. The first scans an open stream of Twitter messages for words that show a sudden increase in frequency, which is likely to reflect that something related to the word is happening. In Chapter 4 a strategy is explored in which tweets are analysed for references to future dates, and for frequent co-occurrences of events and future dates. This latter study was found to be most effective for detecting social events such as music festivals, product releases and national celebrations. The study in Chapter 5, drawing on a record of Twitter messages going back over four years in the past, takes the output of this approach as an intermediate step to identify periodically recurring events. Scanning events for repetitive calendar patterns was found to be most effective to this end.

The second part of the Thesis reports on the use of hashtags for predicting the presence of figurative speech or emotion in tweets. Hashtags like '#sarcasm' and '#loveit' tend to be employed by Twitter users as a space-efficient means to clarify, respectively, the valence and emotion of their message. Even when this context could be understood from the contents of the message itself, the user might include a hashtag to prevent miscommunication or strengthen the emotion. If this latter use occurs sufficiently often, hashtags could serve as valuable aggregators of tweets that exemplify the language of the figurative speech or emotion. By training a machine learning classifier on such examples, typically tens or even hundreds of thousands, the figurative speech or emotion can be detected in tweets that do not include such a hashtag. First, the effectiveness of this approach for the detection of sarcasm is studied in Chapter 6. Training a model based on sarcasm-related hashtags revealed a group of predominantly positive markers, such as 'geweldig' ('great') and 'charmant' ('charming'), as strong indicators of sarcasm. This model was only partly effective in detecting sarcastic tweets, often confusing intense positivity for sarcasm. Extending the procedure of Chapter 6 to various emotion hashtags in Chapter 7 revealed a variable potential for emotion detection. Some hashtags could not be predicted well from their context and others were mostly included in tweets to add the emotion rather than to strengthen it. The hashtag '#zinin' ('#lookingforwardtoit') was an example of a useful hashtag for modelling its emotion of positive anticipation.

The two strands of research, related to events and emotion, come together in the final study reported in Chapter 8: searching for manifestations of public *anticipointment*. This concept stresses the contrast between positive expectations and disappointment, which was analysed by detecting events and using hashtags to detect positive expectation, disappointment and satisfaction in event-referring tweets. The study showed that the average Twitter user is mainly inclined to tweet positively about events, both in anticipation and hindsight. The rare cases in which anticipointment was detected were characterised by a perceived injustice, such as the unfair loss of the favoured sports team or the cancellation of a music concert. Combined, these studies provide a clear overview of the Twitter landscape and the myriad possibilities for detecting events and emotions in tweets.

Curriculum Vitae

Florian Kunneman (Monnickendam, 7 July 1987) obtained his BA in Language and Culture Studies at Utrecht University in 2008. He proceeded to obtain an MA in Communication and Information Sciences at the University of Groningen, graduating in 2010 with a thesis on the future of the Semantic Web. He then moved to Nijmegen, obtaining a MA degree in Language and Speech Technology in 2011 at Radboud University, with a thesis on topic segmentation and clustering of television broadcasts. In the fall of 2011 he started PhD project at Radboud University as part of the national COMMIT/ research program. During this project he was visiting researcher at the Computational Linguistics and Psycholinguistics Research Group in Antwerp in 2015. He currently works as a postdoctoral researcher at the Centre for Language Studies at Radboud University, without anticipointment.

SIKS Dissertation Series

1998

1. Johan van den Akker (CWI) *DEGAS - An Active, Temporal Database of Autonomous Objects*
2. Floris Wiesman (UM) *Information Retrieval by Graphically Browsing Meta-Information*
3. Ans Steuten (TUD) *A Contribution to the Linguistic Analysis of Busine within the Language/Action Perspective*
4. Dennis Breuker (UM) *Memory versus Search in Games*
5. E.W. Oskamp (RUL) *Computerondersteuning bij Straftoemeting*

1999

1. Mark Sloof (VU) *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*
2. Rob Potharst (EUR) *Classification using decision trees and neural nets*
3. Don Beal (UM) *The Nature of Minimax Search*
4. Jacques Penders (UM) *The practical Art of Moving Physical Objects*
5. Aldo de Moor (KUB) *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*
6. Niek J.E. Wijngaards (VU) *Re-design of compositional systems*
7. David Spelt (UT) *Verification support for object database design*
8. Jacques H.J. Lenting (UM) *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation*

2000

1. Frank Niessink (VU) *Perspectives on Improving Software Maintenance*
2. Koen Holtman (TUE) *Prototyping of CMS Storage Management*
3. Carolien M.T. Metselaar (UVA) *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief*
4. Geert de Haan (VU) *ETAG, A Formal Model of Competence Knowledge for User Interface Design*
5. Ruud van der Pol (UM) *Knowledge-based Query Formulation in Information Retrieval*
6. Rogier van Eijk (UU) *Programming Languages for Agent Communication*
7. Niels Peek (UU) *Decision-theoretic Planning of Clinical Patient Management*
8. Veerle Coup (EUR) *Sensitivity Analyis of Decision-Theoretic Networks*
9. Florian Waas (CWI) *Principles of Probabilistic Query Optimization*
10. Niels Nes (CWI) *Image Database Management System Design Considerations, Algorithms and Architecture*
11. Jonas Karlsson (CWI) *Scalable Distributed Data Structures for Database Management*

2001

1. Silja Renooij (UU) *Qualitative Approaches to Quantifying Probabilistic Networks*
2. Koen Hindriks (UU) *Agent Programming Languages: Programming with Mental Models*
3. Maarten van Someren (UvA) *Learning as problem solving*
4. Evgueni Smirnov (UM) *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*
5. Jacco van Ossenbruggen (VU) *Processing Structured Hypermedia: A Matter of Style*
6. Martijn van Welie (VU) *Task-based User Interface Design*
7. Bastiaan Schonhage (VU) *Diva: Architectural Perspectives on Information Visualization*
8. Pascal van Eck (VU) *A Compositional Semantic Structure for Multi-Agent Systems Dynamics*
9. Pieter Jan 't Hoen (RUL) *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*
10. Maarten Sierhuis (UvA) *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and*

simulation language for work practice analysis and design

11. Tom M. van Engers (VUA) *Knowledge Management: The Role of Mental Models in Business Systems Design*

2002

1. Nico Lassing (VU) *Architecture-Level Modifiability Analysis*
2. Roelof van Zwol (UT) *Modelling and searching web-based document collections*
3. Henk Ernst Blok (UT) *Database Optimization Aspects for Information Retrieval*
4. Juan Roberto Castelo Valdueza (UU) *The Discrete Acyclic Digraph Markov Model in Data Mining*
5. Radu Serban (VU) *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*
6. Laurens Mommers (UL) *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*
7. Peter Boncz (CWI) *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*
8. Jaap Gordijn (VU) *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*
9. Willem-Jan van den Heuvel(KUB) *Integrating Modern Business Applications with Objectified Legacy Systems*
10. Brian Sheppard (UM) *Towards Perfect Play of Scrabble*
11. Wouter C.A. Wijngaards (VU) *Agent Based Modelling of Dynamics: Biological and Organisational Applications*
12. Albrecht Schmidt (Uva) *Processing XML in Database Systems*
13. Hongjing Wu (TUE) *A Reference Architecture for Adaptive Hypermedia Applications*
14. Wieke de Vries (UU) *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*
15. Rik Eshuis (UT) *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*
16. Pieter van Langen (VU) *The Anatomy of Design: Foundations, Models and Applications*
17. Stefan Manegold (UVA) *Understanding, Modeling, and Improving Main-Memory Database Performance*

2003

1. Heiner Stuckenschmidt (VU) *Ontology-Based Information Sharing in Weakly Structured Environments*
2. Jan Broersen (VU) *Modal Action Logics for Reasoning About Reactive Systems*
3. Martijn Schuemie (TUD) *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*
4. Milan Petkovic (UT) *Content-Based Video Retrieval Supported by Database Technology*
5. Jos Lehmann (UVA) *Causation in Artificial Intelligence and Law - A modelling approach*
6. Boris van Schooten (UT) *Development and specification of virtual environments*
7. Machiel Jansen (UvA) *Formal Explorations of Knowledge Intensive Tasks*
8. Yongping Ran (UM) *Repair Based Scheduling*
9. Rens Kortmann (UM) *The resolution of visually guided behaviour*
10. Andreas Lincke (UvT) *Electronic Business Negotiation: Some experimental studies on t between medium, innovation context and culture*
11. Simon Keizer (UT) *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*
12. Roeland Ordelman (UT) *Dutch speech recognition in multimedia information retrieval*
13. Jeroen Donkers (UM) *Nosce Hostem - Searching with Opponent Models*
14. Stijn Hoppenbrouwers (KUN) *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*
15. Mathijs de Weerdt (TUD) *Plan Merging in Multi-Agent Systems*
16. Menzo Windhouwer (CWI) *Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses*
17. David Jansen (UT) *Extensions of Statecharts with Probability, Time, and Stochastic Timing*
18. Levente Kocsis (UM) *Learning Search Decisions*

2004

1. Virginia Dignum (UU) *A Model for Organizational Interaction: Based on Agents, Founded in Logic*
2. Lai Xu (UvT) *Monitoring Multi-party Contracts for E-business*
3. Perry Groot (VU) *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*
4. Chris van Aart (UVA) *Organizational Principles for Multi-Agent Architectures*
5. Viara Popova (EUR) *Knowledge discovery and monotonicity*
6. Bart-Jan Hommes (TUD) *The Evaluation of Business Process Modeling Techniques*
7. Elise Boltjes (UM) *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*
8. Joop Verbeek(UM) *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise*
9. Martin Caminada (VU) *For the Sake of the Argument; explorations into argument-based reasoning*

10. Suzanne Kabel (UVA) *Knowledge-rich indexing of learning-objects*
11. Michel Klein (VU) *Change Management for Distributed Ontologies*
12. The Duy Bui (UT) *Creating emotions and facial expressions for embodied agents*
13. Wojciech Jamroga (UT) *Using Multiple Models of Reality: On Agents who Know how to Play*
14. Paul Harrenstein (UU) *Logic in Conflict. Logical Explorations in Strategic Equilibrium*
15. Arno Knobbe (UU) *Multi-Relational Data Mining*
16. Federico Divina (VU) *Hybrid Genetic Relational Search for Inductive Learning*
17. Mark Winands (UM) *Informed Search in Complex Games*
18. Vania Bessa Machado (UvA) *Supporting the Construction of Qualitative Knowledge Models*
19. Thijs Westerveld (UT) *Using generative probabilistic models for multimedia retrieval*
20. Madelon Evers (Nyenrode) *Learning from Design: facilitating multidisciplinary design teams*

2005

1. Floor Verdenius (UVA) *Methodological Aspects of Designing Induction-Based Applications*
2. Erik van der Werf (UM)) *AI techniques for the game of Go*
3. Franc Grootjen (RUN) *A Pragmatic Approach to the Conceptualisation of Language*
4. Nirvana Meratnia (UT) *Towards Database Support for Moving Object data*
5. Gabriel Infante-Lopez (UVA) *Two-Level Probabilistic Grammars for Natural Language Parsing*
6. Pieter Spronck (UM) *Adaptive Game AI*
7. Flavius Frasincar (TUE) *Hypermedia Presentation Generation for Semantic Web Information Systems*
8. Richard Vdovjak (TUE) *A Model-driven Approach for Building Distributed Ontology-based Web Applications*
9. Jeen Broekstra (VU) *Storage, Querying and Inferencing for Semantic Web Languages*
10. Anders Bouwer (UVA) *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*
11. Elth Ogston (VU) *Agent Based Matchmaking and Clustering - A Decentralized Approach to Search*
12. Csaba Boer (EUR) *Distributed Simulation in Industry*
13. Fred Hamburg (UL) *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*
14. Borys Omelayenko (VU) *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*
15. Tibor Bosse (VU) *Analysis of the Dynamics of Cognitive Processes*
16. Joris Graaumans (UU) *Usability of XML Query Languages*
17. Boris Shishkov (TUD) *Software Specification Based on Re-usable Business Components*
18. Danielle Sent (UU) *Test-selection strategies for probabilistic networks*
19. Michel van Dartel (UM) *Situated Representation*
20. Cristina Coteanu (UL) *Cyber Consumer Law, State of the Art and Perspectives*
21. Wijnand Derks (UT) *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*

2006

1. Samuil Angelov (TUE) *Foundations of B2B Electronic Contracting*
2. Cristina Chisalita (VU) *Contextual issues in the design and use of information technology in organizations*
3. Noor Christoph (UVA) *The role of metacognitive skills in learning to solve problems*
4. Marta Sabou (VU) *Building Web Service Ontologies*
5. Cees Pierik (UU) *Validation Techniques for Object-Oriented Proof Outlines*
6. Ziv Baida (VU) *Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling*
7. Marko Smiljanic (UT) *XML schema matching – balancing efficiency and effectiveness by means of clustering*
8. Eelco Herder (UT) *Forward, Back and Home Again - Analyzing User Behavior on the Web*
9. Mohamed Wahdan (UM) *Automatic Formulation of the Auditor's Opinion*
10. Ronny Siebes (VU) *Semantic Routing in Peer-to-Peer Systems*
11. Joeri van Ruth (UT) *Flattening Queries over Nested Data Types*
12. Bert Bongers (VU) *Interactivation - Towards an ecology of people, our technological environment, and the arts*
13. Henk-Jan Lebbink (UU) *Dialogue and Decision Games for Information Exchanging Agents*
14. Johan Hoorn (VU) *Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change*
15. Rainer Malik (UU) *CONAN: Text Mining in the Biomedical Domain*
16. Carsten Riggelsen (UU) *Approximation Methods for Efficient Learning of Bayesian Networks*
17. Stacey Nagata (UU) *User Assistance for Multitasking with Interruptions on a Mobile Device*

18. Valentin Zhizhkun (UVA) *Graph transformation for Natural Language Processing*
19. Birna van Riemsdijk (UU) *Cognitive Agent Programming: A Semantic Approach*
20. Marina Velikova (UvT) *Monotone models for prediction in data mining*
21. Bas van Gils (RUN) *Aptness on the Web*
22. Paul de Vrieze (RUN) *Fundaments of Adaptive Personalisation*
23. Ion Juvina (UU) *Development of Cognitive Model for Navigating on the Web*
24. Laura Hollink (VU) *Semantic Annotation for Retrieval of Visual Resources*
25. Madalina Drugan (UU) *Conditional log-likelihood MDL and Evolutionary MCMC*
26. Vojkan Mihajlović (UT) *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*
27. Stefano Bocconi (CWI) *Vox Populi: generating video documentaries from semantically annotated media repositories*
28. Borkur Sigurbjornsson (UVA) *Focused Information Access using XML Element Retrieval*

2007

1. Kees Leune (UvT) *Access Control and Service-Oriented Architectures*
2. Wouter Teepe (RUG) *Reconciling Information Exchange and Confidentiality: A Formal Approach*
3. Peter Mika (VU) *Social Networks and the Semantic Web*
4. Jurriaan van Diggelen (UU) *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*
5. Bart Schermer (UL) *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*
6. Gilad Mishne (UVA) *Applied Text Analytics for Blogs*
7. Natasa Jovanovic' (UT) *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*
8. Mark Hoogendoorn (VU) *Modeling of Change in Multi-Agent Organizations*
9. David Mobach (VU) *Agent-Based Mediated Service Negotiation*
10. Huib Aldewereld (UU) *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*
11. Natalia Stash (TUE) *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*
12. Marcel van Gerven (RUN) *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*
13. Rutger Rienks (UT) *Meetings in Smart Environments; Implications of Progressing Technology*
14. Niek Bergboer (UM) *Context-Based Image Analysis*
15. Joyca Lacroix (UM) *NIM: a Situated Computational Memory Model*
16. Davide Grossi (UU) *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*
17. Theodore Charitos (UU) *Reasoning with Dynamic Networks in Practice*
18. Bart Orriens (UvT) *On the development an management of adaptive business collaborations*
19. David Levy (UM) *Intimate relationships with artificial partners*
20. Slinger Jansen (UU) *Customer Configuration Updating in a Software Supply Network*
21. Karianne Vermaas (UU) *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*
22. Zlatko Zlatev (UT) *Goal-oriented design of value and process models from patterns*
23. Peter Barna (TUE) *Specification of Application Logic in Web Information Systems*
24. Georgina Ramírez Camps (CWI) *Structural Features in XML Retrieval*
25. Joost Schalken (VU) *Empirical Investigations in Software Process Improvement*

2008

1. Katalin Boer-Sorbán (EUR) *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*
2. Alexei Sharpanskykh (VU) *On Computer-Aided Methods for Modeling and Analysis of Organizations*
3. Vera Hollink (UVA) *Optimizing hierarchical menus: a usage-based approach*
4. Ander de Keijzer (UT) *Management of Uncertain Data - towards unattended integration*
5. Bela Mutschler (UT) *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*
6. Arjen Hommersom (RUN) *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*
7. Peter van Rosmalen (OU) *Supporting the tutor in the design and support of adaptive e-learning*
8. Janneke Bolt (UU) *Bayesian Networks: Aspects of Approximate Inference*
9. Christof van Nimwegen (UU) *The paradox of the guided user: assistance can be counter-effective*
10. Wauter Bosma (UT) *Discourse oriented summarization*
11. Vera Kartseva (VU) *Designing Controls for Network Organizations: A Value-Based Approach*

12. Jozsef Farkas (RUN) *A Semiotically Oriented Cognitive Model of Knowledge Representation*
13. Caterina Carraciolo (UVA) *Topic Driven Access to Scientific Handbooks*
14. Arthur van Bunningen (UT) *Context-Aware Querying; Better Answers with Less Effort*
15. Martijn van Otterlo (UT) *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains*
16. Henriette van Vugt (VU) *Embodied agents from a user's perspective*
17. Martin Op 't Land (TUD) *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*
18. Guido de Croon (UM) *Adaptive Active Vision*
19. Henning Rode (UT) *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*
20. Rex Arendsen (UVA) *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven*
21. Krisztian Balog (UVA) *People Search in the Enterprise*
22. Henk Koning (UU) *Communication of IT-Architecture*
23. Stefan Visscher (UU) *Bayesian network models for the management of ventilator-associated pneumonia*
24. Zharko Aleksovski (VU) *Using background knowledge in ontology matching*
25. Geert Jonker (UU) *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*
26. Marijn Huijbregts (UT) *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*
27. Hubert Vogten (OU) *Design and Implementation Strategies for IMS Learning Design*
28. Ildiko Flesch (RUN) *On the Use of Independence Relations in Bayesian Networks*
29. Dennis Reidsma (UT) *Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans*
30. Wouter van Atteveldt (VU) *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*
31. Loes Braun (UM) *Pro-Active Medical Information Retrieval*
32. Trung H. Bui (UT) *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes*
33. Frank Terpstra (UVA) *Scientific Workflow Design; theoretical and practical issues*
34. Jeroen de Knijf (UU) *Studies in Frequent Tree Mining*
35. Ben Torben Nielsen (UvT) *Dendritic morphologies: function shapes structure*

2009

1. Rasa Jurgelenaite (RUN) *Symmetric Causal Independence Models*
2. Willem Robert van Hage (VU) *Evaluating Ontology-Alignment Techniques*
3. Hans Stol (UvT) *A Framework for Evidence-based Policy Making Using IT*
4. Josephine Nabukenya (RUN) *Improving the Quality of Organisational Policy Making using Collaboration Engineering*
5. Sietse Overbeek (RUN) *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*
6. Muhammad Subianto (UU) *Understanding Classification*
7. Ronald Poppe (UT) *Discriminative Vision-Based Recovery and Recognition of Human Motion*
8. Volker Nannen (VU) *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*
9. Benjamin Kanagwa (RUN) *Design, Discovery and Construction of Service-oriented Systems*
10. Jan Wielemaker (UVA) *Logic programming for knowledge-intensive interactive applications*
11. Alexander Boer (UVA) *Legal Theory, Sources of Law & the Semantic Web*
12. Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin) *perating Guidelines for Services*
13. Steven de Jong (UM) *Fairness in Multi-Agent Systems*
14. Maksym Korotkiy (VU) *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*
15. Rinke Hoekstra (UVA) *Ontology Representation - Design Patterns and Ontologies that Make Sense*
16. Fritz Reul (UvT) *New Architectures in Computer Chess*
17. Laurens van der Maaten (UvT) *Feature Extraction from Visual Data*
18. Fabian Groffen (CWI) *Armada, An Evolving Database System*
19. Valentin Robu (CWI) *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*
20. Bob van der Vecht (UU) *Adjustable Autonomy: Controling Influences on Decision Making*
21. Stijn Vanderlooy (UM) *Ranking and Reliable Classification*
22. Pavel Serdyukov (UT) *Search For Expertise: Going beyond direct evidence*
23. Peter Hofgesang (VU) *Modelling Web Usage in a Changing Environment*
24. Annerieke Heuvelink (VUA) *Cognitive Models for Training Simulations*

25. Alex van Ballegooij (CWI) *"RAM: Array Database Management through Relational Mapping"*
26. Fernando Koch (UU) *An Agent-Based Model for the Development of Intelligent Mobile Services*
27. Christian Glahn (OU) *Contextual Support of social Engagement and Reflection on the Web*
28. Sander Evers (UT) *Sensor Data Management with Probabilistic Models*
29. Stanislav Pokraev (UT) *Model-Driven Semantic Integration of Service-Oriented Applications*
30. Marcin Zukowski (CWI) *Balancing vectorized query execution with bandwidth-optimized storage*
31. Sofiya Katrenko (UVA) *A Closer Look at Learning Relations from Text*
32. Rik Farenhorst (VU) and Remco de Boer (VU) *Architectural Knowledge Management: Supporting Architects and Auditors*
33. Khiet Truong (UT) *How Does Real Affect Affect Affect Recognition In Speech?*
34. Inge van de Weerd (UU) *Advancing in Software Product Management: An Incremental Method Engineering Approach*
35. Wouter Koelewijn (UL) *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*
36. Marco Kalz (OUN) *Placement Support for Learners in Learning Networks*
37. Hendrik Drachsler (OUN) *Navigation Support for Learners in Informal Learning Networks*
38. Riina Vuorikari (OU) *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*
39. Christian Stahl (TUE, Humboldt-Universitaet zu Berlin) *Service Substitution – A Behavioral Approach Based on Petri Nets*
40. Stephan Raaijmakers (UvT) *Multinomial Language Learning: Investigations into the Geometry of Language*
41. Igor Berezhnyy (UvT) *Digital Analysis of Paintings*
42. Toine Bogers *Recommender Systems for Social Bookmarking*
43. Virginia Nunes Leal Franqueira (UT) *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*
44. Roberto Santana Tapia (UT) *Assessing Business-IT Alignment in Networked Organizations*
45. Jilles Vreeken (UU) *Making Pattern Mining Useful*
46. Loredana Afanasiev (UvA) *Querying XML: Benchmarks and Recursion*

2010

1. Matthijs van Leeuwen (UU) *Patterns that Matter*
2. Ingo Wassink (UT) *Work flows in Life Science*
3. Joost Geurts (CWI) *A Document Engineering Model and Processing Framework for Multimedia documents*
4. Olga Kulyk (UT) *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*
5. Claudia Hauff (UT) *Predicting the Effectiveness of Queries and Retrieval Systems*
6. Sander Bakkes (UvT) *Rapid Adaptation of Video Game AI*
7. Wim Fikkert (UT) *Gesture interaction at a Distance*
8. Krzysztof Siewicz (UL) *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*
9. Hugo Kielman (UL) *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*
10. Rebecca Ong (UL) *Mobile Communication and Protection of Childr*
11. Adriaan Ter Mors (TUD) *The world according to MARP: Multi-Agent Route Planning*
12. Susan van den Braak (UU) *Sensemaking software for crime analysis*
13. Gianluigi Folino (RUN) *High Performance Data Mining using Bio-inspired techniques*
14. Sander van Splunter (VU) *Automated Web Service Reconfiguration*
15. Lianne Bodenstaff (UT) *Managing Dependency Relations in Inter-Organizational Models*
16. Sicco Verwer (TUD) *Efficient Identification of Timed Automata, theory and practice*
17. Spyros Kotoulas (VU) *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*
18. Charlotte Gerritsen (VU) *Caught in the Act: Investigating Crime by Agent-Based Simulation*
19. Henriette Cramer (UvA) *People's Responses to Autonomous and Adaptive Systems*
20. Ivo Swartjes (UT) *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*
21. Harold van Heerde (UT) *Privacy-aware data management by means of data degradation*
22. Michiel Hildebrand (CWI) *End-user Support for Access to Heterogeneous Linked Data*
23. Bas Steunebrink (UU) *The Logical Structure of Emotions*
24. *Designing Generic and Efficient Negotiation Strategies*
25. Zulfiqar Ali Memon (VU) *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*
26. Ying Zhang (CWI) *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*
27. Marten Voulon (UL) *Automatisch contracteren*
28. Arne Koopman (UU) *Characteristic Relational Patterns*

29. Stratos Idreos(CWI) *Database Cracking: Towards Auto-tuning Database Kernels*
30. Marieke van Erp (UvT) *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval*
31. Victor de Boer (UVA) *Ontology Enrichment from Heterogeneous Sources on the Web*
32. Marcel Hiel (UvT) *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*
33. Robin Aly (UT) *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*
34. Teduh Dirgahayu (UT) *Interaction Design in Service Compositions*
35. Dolf Trieschnigg (UT) *Proof of Concept: Concept-based Biomedical Information Retrieval*
36. Jose Janssen (OU) *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification*
37. Niels Lohmann (TUE) *Correctness of services and their composition*
38. Dirk Fahland (TUE) *From Scenarios to components*
39. Ghazanfar Farooq Siddiqui (VU) *Integrative modeling of emotions in virtual agents*
40. Mark van Assem (VU) *Converting and Integrating Vocabularies for the Semantic Web*
41. Guillaume Chaslot (UM) *Monte-Carlo Tree Search*
42. Sybren de Kinderen (VU) *Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach*
43. Peter van Kranenburg (UU) *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*
44. Pieter Bellekens (TUE) *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*
45. Vasilios Andrikopoulos (UvT) *A theory and model for the evolution of software services*
46. Vincent Pijpers (VU) *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*
47. Chen Li (UT) *Mining Process Model Variants: Challenges, Techniques, Examples*
48. Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*
49. Jahn-Takeshi Saito (UM) *Solving difficult game positions*
50. Bouke Huurnink (UVA) *Search in Audiovisual Broadcast Archives*
51. Alia Khairia Amin (CWI) *Understanding and supporting information seeking tasks in multiple sources*
52. Peter-Paul van Maanen (VU) *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*
53. Edgar Meij (UVA) *Combining Concepts and Language Models for Information Access*

2011

1. Botond Cseke (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*
2. Nick Tinnemeier(UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*
3. Jan Martijn van der Werf (TUE) *Compositional Design and Verification of Component-Based Information Systems*
4. Hado van Hasselt (UU) *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms*
5. Base van der Raadt (VU) *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline*
6. Yiwen Wang (TUE) *Semantically-Enhanced Recommendations in Cultural Heritage*
7. Yujia Cao (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*
8. Nieske Vergunst (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*
9. Tim de Jong (OU) *Contextualised Mobile Media for Learning*
10. Bart Bogaert (UvT) *Cloud Content Contention*
11. Dhaval Vyas (UT) *Designing for Awareness: An Experience-focused HCI Perspective*
12. Carmen Bratosin (TUE) *Grid Architecture for Distributed Process Mining*
13. Xiaoyu Mao (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*
14. Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*
15. Marijn Koolen (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*
16. Maarten Schadd (UM) *Selective Search in Games of Different Complexity*
17. Jiyin He (UVA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*
18. Mark Ponsen (UM) *Strategic Decision-Making in complex games*
19. Ellen Rusman (OU) *The Mind ' s Eye on Personal Profiles*
20. Qing Gu (VU) *Guiding service-oriented software engineering - A view-based approach*
21. Linda Terlouw (TUD) *Modularization and Specification of Service-Oriented Systems*
22. Junte Zhang (UVA) *System Evaluation of Archival Description and Access*
23. Wouter Weerkamp (UVA) *Finding People and their Utterances in Social Media*
24. Herwin van Welbergen (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

25. Syed Waqar ul Qounain Jaffry (VU) *Analysis and Validation of Models for Trust Dynamics*
26. Matthijs Aart Pontier (VU) *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*
27. Aniel Bhulai (VU) *Dynamic website optimization through autonomous management of design patterns*
28. Rianne Kaptein(UVA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*
29. Faisal Kamiran (TUE) *Discrimination-aware Classification*
30. Egon van den Broek (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*
31. Ludo Waltman (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*
32. Nees-Jan van Eck (EUR) *Methodological Advances in Bibliometric Mapping of Science*
33. Tom van der Weide (UU) *Arguing to Motivate Decisions*
34. Paolo Turrini (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*
35. Maaike Harbers (UU) *Explaining Agent Behavior in Virtual Training*
36. Erik van der Spek (UU) *Experiments in serious game design: a cognitive approach*
37. Adriana Burlutiu (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*
38. Nyree Lemmens (UM) *Bee-inspired Distributed Optimization*
39. Joost Westra (UU) *Organizing Adaptation using Agents in Serious Games*
40. Viktor Clerc (VU) *Architectural Knowledge Management in Global Software Development*
41. Luan Ibraimi (UT) *Cryptographically Enforced Distributed Data Access Control*
42. Michal Sindlar (UU) *Explaining Behavior through Mental State Attribution*
43. Henk van der Schuur (UU) *Process Improvement through Software Operation Knowledge*
44. Boris Reuderink (UT) *Robust Brain-Computer Interfaces*
45. Herman Stehouwer (UvT) *Statistical Language Models for Alternative Sequence Selection*
46. Beibei Hu (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*
47. Azizi Bin Ab Aziz(VU) *Exploring Computational Models for Intelligent Support of Persons with Depression*
48. Mark Ter Maat (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*
49. Andreea Niculescu (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

1. Terry Kakeeto (UvT) *Relationship Marketing for SMEs in Uganda*
2. Muhammad Umair(VU) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*
3. Adam Vanya (VU) *Supporting Architecture Evolution by Mining Software Repositories*
4. Jurriaan Souer (UU) *Development of Content Management System-based Web Applications*
5. Marijn Plomp (UU) *Maturing Interorganisational Information Systems*
6. Wolfgang Reinhardt (OU) *Awareness Support for Knowledge Workers in Research Networks*
7. Rianne van Lambalgen (VU) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*
8. Gerben de Vries (UVA) *Kernel Methods for Vessel Trajectories*
9. Ricardo Neisse (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*
10. David Smits (TUE) *Towards a Generic Distributed Adaptive Hypermedia Environment*
11. J.C.B. Rantham Prabhakara (TUE) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*
12. Kees van der Sluijs (TUE) *Model Driven Design and Data Integration in Semantic Web Information Systems*
13. Suleman Shahid (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*
14. Evgeny Knutov(TUE) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*
15. Natalie van der Wal (VU) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*
16. Fiemke Both (VU) *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment*
17. Amal Elgammal (UvT) *Towards a Comprehensive Framework for Business Process Compliance*
18. Eltjo Poort (VU) *Improving Solution Architecting Practices*
19. Helen Schonenberg (TUE) *What's Next? Operational Support for Business Process Execution*
20. Ali Bahramisharif (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*
21. Roberto Cornacchia (TUD) *Querying Sparse Matrices for Information Retrieval*

22. Thijs Vis (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*
23. Christian Muehl (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*
24. Laurens van der Werff (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*
25. Silja Eckartz (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*
26. Emile de Maat (UVA) *Making Sense of Legal Text*
27. Hayrettin Gurkok (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*
28. Nancy Pascall (UvT) *Engendering Technology Empowering Women*
29. Almer Tigelaar (UT) *Peer-to-Peer Information Retrieval*
30. Alina Pommeranz (TUD) *Designing Human-Centered Systems for Reflective Decision Making*
31. Emily Bagarukayo (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*
32. Wietske Visser (TUD) *Qualitative multi-criteria preference representation and reasoning*
33. Rory Sie (OUN) *Coalitions in Cooperation Networks (COCOON)*
34. Pavol Jancura (RUN) *Evolutionary analysis in PPI networks and applications*
35. Evert Haasdijk (VU) *Never Too Old To Learn – Online Evolution of Controllers in Swarm- and Modular Robotics*
36. Denis Ssebugwawo (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*
37. Agnes Nakakawa (RUN) *A Collaboration Process for Enterprise Architecture Creation*
38. Selmar Smit (VU) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*
39. Hassan Fatemi (UT) *Risk-aware design of value and coordination networks*
40. Agus Gunawan (UvT) *Information Access for SMEs in Indonesia*
41. Sebastian Kelle (OU) *Game Design Patterns for Learning*
42. Dominique Verpoorten (OU) *Reflection Amplifiers in self-regulated Learning*
43. Anna Tordai (VU) *On Combining Alignment Techniques*
44. Benedikt Kratz (UvT) *A Model and Language for Business-aware Transactions*
45. Simon Carter (UVA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*
46. Manos Tsagkias (UVA) *Mining Social Media: Tracking Content and Predicting Behavior*
47. Jorn Bakker (TUE) *Handling Abrupt Changes in Evolving Time-series Data*
48. Michael Kaisers (UM) *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*
49. Steven van Kervel (TUD) *Ontologogy driven Enterprise Information Systems Engineering*
50. Jeroen de Jong (TUD) *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*

2013

1. Viorel Milea (EUR) *News Analytics for Financial Decision Support*
2. Erietta Liarou (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*
3. Szymon Klarman (VU) *Reasoning with Contexts in Description Logics*
4. Chetan Yadati(TUD) *Coordinating autonomous planning and scheduling*
5. Dulce Pumareja (UT) *Groupware Requirements Evolutions Patterns*
6. Romulo Goncalves(CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*
7. Giel van Lankveld (UvT) *Quantifying Individual Player Differences*
8. Robbert-Jan Merk(VU) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*
9. Fabio Gori (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*
10. Jeewanie Jayasinghe Arachchige(UvT) *A Unified Modeling Framework for Service Design*
11. Evangelos Pournaras(TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*
12. Marian Razavian(VU) *Knowledge-driven Migration to Services*
13. Mohammad Safiri(UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*
14. Jafar Tanha (UVA) *Ensemble Approaches to Semi-Supervised Learning Learning*
15. Daniel Hennes (UM) *Multiagent Learning - Dynamic Games and Applications*
16. Eric Kok (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*
17. Koen Kok (VU) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*
18. Jeroen Janssens (UvT) *Outlier Selection and One-Class Classification*
19. Renze Steenhuizen (TUD) *Coordinated Multi-Agent Planning and Scheduling*
20. Katja Hofmann (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*
21. Sander Wubben (UvT) *Text-to-text generation by monolingual machine translation*

22. Tom Claassen (RUN) *Causal Discovery and Logic*
23. Patricio de Alencar Silva(UvT) *Value Activity Monitoring*
24. Haitham Bou Ammar (UM) *Automated Transfer in Reinforcement Learning*
25. Agnieszka Anna Latoszek-Berendsen (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*
26. Alireza Zarghami (UT) *Architectural Support for Dynamic Homecare Service Provisioning*
27. Mohammad Huq (UT) *Inference-based Framework Managing Data Provenance*
28. Frans van der Sluis (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*
29. Iwan de Kok (UT) *Listening Heads*
30. Joyce Nakatumba (TUE) *Resource-Aware Business Process Management: Analysis and Support*
31. Dinh Khoa Nguyen (UvT) *Blueprint Model and Language for Engineering Cloud Applications*
32. Kamakshi Rajagopal (OUN) *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development*
33. Qi Gao (TUD) *User Modeling and Personalization in the Microblogging Sphere*
34. Kien Tjin-Kam-Jet (UT) *Distributed Deep Web Search*
35. Abdallah El Ali (UvA) *Minimal Mobile Human Computer Interaction*
36. Than Lam Hoang (TUe) *Pattern Mining in Data Streams*
37. Dirk Bürner (OUN) *Ambient Learning Displays*
38. Eelco den Heijer (VU) *Autonomous Evolutionary Art*
39. Joop de Jong (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*
40. Pim Nijssen (UM) *Monte-Carlo Tree Search for Multi-Player Games*
41. Jochem Liem (UVA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*
42. Léon Planken (TUD) *Algorithms for Simple Temporal Reasoning*
43. Marc Bron (UVA) *Exploration and Contextualization through Interaction and Concepts*

2014

1. Nicola Barile (UU) *Studies in Learning Monotone Models from Data*
2. Fiona Tuliyano (RUN) *Combining System Dynamics with a Domain Modeling Method*
3. Sergio Raul Duarte Torres (UT) *Information Retrieval for Children: Search Behavior and Solutions*
4. Hanna Jochmann-Mannak (UT) *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*
5. Jurriaan van Reijsen (UU) *Knowledge Perspectives on Advancing Dynamic Capability*
6. Damian Tamburri (VU) *Supporting Networked Software Development*
7. Arya Adriansyah (TUE) *Aligning Observed and Modeled Behavior*
8. Samur Araujo (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*
9. Philip Jackson (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*
10. Ivan Salvador Razo Zapata (VU) *Service Value Networks*
11. Janneke van der Zwaan (TUD) *An Empathic Virtual Buddy for Social Support*
12. Willem van Willigen (VU) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*
13. Arlette van Wissen (VU) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*
14. Yangyang Shi (TUD) *Language Models With Meta-information*
15. Natalya Mogles (VU) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*
16. Krystyna Milian (VU) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*
17. Kathrin Dentler (VU) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*
18. Mattijs Ghijsen (UVA) *Methods and Models for the Design and Study of Dynamic Agent Organizations*
19. Vinicius Ramos (TUE) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*
20. Mena Habib (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*
21. Kassidy Clark (TUD) *Negotiation and Monitoring in Open Environments*
22. Marieke Peeters (UU) *Personalized Educational Games - Developing agent-supported scenario-based training*
23. Eleftherios Sidirourgos (UvA/CWI) *Space Efficient Indexes for the Big Data Era*
24. Davide Ceolin (VU) *Trusting Semi-structured Web Data*
25. Martijn Lappenschaar (RUN) *New network models for the analysis of disease interaction*
26. Tim Baarslag (TUD) *What to Bid and When to Stop*

27. Rui Jorge Almeida (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*
28. Anna Chmielowiec (VU) *Decentralized k-Clique Matching*
29. Jaap Kabbedijk (UU) *Variability in Multi-Tenant Enterprise Software*
30. Peter de Cock (UvT) *Anticipating Criminal Behaviour*
31. Leo van Moergestel (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*
32. Naser Ayat (UvA) *On Entity Resolution in Probabilistic Data*
33. Tesfa Tegegne (RUN) *Service Discovery in eHealth*
34. Christina Manteli(VU) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*
35. Joost van Ooijen (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*
36. Joos Buijs (TUE) *Flexible Evolutionary Algorithms for Mining Structured Process Models*
37. Maral Dadvar (UT) *Experts and Machines United Against Cyberbullying*
38. Danny Plass-Oude Bos (UT) *Making brain-computer interfaces better: improving usability through post-processing.*
39. Jasmina Maric (UvT) *Web Communities, Immigration, and Social Capital*
40. Walter Omona (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*
41. Frederic Hogenboom (EUR) *Automated Detection of Financial Events in News Text*
42. Carsten Eijckhof (CWI/TUD) *Contextual Multidimensional Relevance Models*
43. Kevin Vlaanderen (UU) *Supporting Process Improvement using Method Increments*
44. Paulien Meesters (UvT) *Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*
45. Birgit Schmitz (OUN) *Mobile Games for Learning: A Pattern-Based Approach*
46. Ke Tao (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*
47. Shangsong Liang (UVA) *Fusion and Diversification in Information Retrieval*

2015

1. Niels Netten (UvA) *Machine Learning for Relevance of Information in Crisis Response*
2. Faiza Bukhsh (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*
3. Twan van Laarhoven (RUN) *Machine learning for network data*
4. Howard Spoelstra (OUN) *Collaborations in Open Learning Environments*
5. Christoph B sch(UT) *Cryptographically Enforced Search Pattern Hiding*
6. Farideh Heidari (TUD) *Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes*
7. Maria-Hendrike Peetz(UvA) *Time-Aware Online Reputation Analysis*
8. Jie Jiang (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*
9. Randy Klaassen(UT) *HCI Perspectives on Behavior Change Support Systems*
10. Henry Hermans (OUN) *OpenU: design of an integrated system to support lifelong learning*
11. Yongming Luo(TUE) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*
12. Julie M. Birkholz (VU) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*
13. Giuseppe Procaccianti(VU) *Energy-Efficient Software*
14. Bart van Straalen (UT) *A cognitive approach to modeling bad news conversations*
15. Klaas Andries de Graaf (VU) *Ontology-based Software Architecture Documentation*
16. Changyun Wei (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*
17. André' van Cleeff (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*
18. Holger Pirk (CWI) *Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories*
19. Bernardo Tabuenca (OUN) *Ubiquitous Technology for Lifelong Learners*
20. Loïs Vanhée(UU) *Using Culture and Values to Support Flexible Coordination*
21. Sibren Fetter (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*
22. Zhemin Zhu(UT) *Co-occurrence Rate Networks*
23. Luit Gazendam (VU) *Cataloguer Support in Cultural Heritage*
24. Richard Berendsen (UVA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*
25. Steven Woudenberg (UU) *Bayesian Tools for Early Disease Detection*
26. Alexander Hogenboom (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*
27. Sándor Héman (CWI) *Updating compressed colomn stores*
28. Janet Bagorogoza(TiU) *KNOWLEDGE MANAGEMENT AND HIGH PERFORMANCE; The Uganda Financial Institutions Model for HPO*

29. Hendrik Baier (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*
30. Kiavash Bahreini(OU) *Real-time Multimodal Emotion Recognition in E-Learning*
31. Yakup Koç (TUD) *On the robustness of Power Grids*
32. Jerome Gard(UL) *Corporate Venture Management in SMEs*
33. Frederik Schadd (TUD) *Ontology Mapping with Auxiliary Resources*
34. Victor de Graaf(UT) *Gesocial Recommender Systems*
35. Jungxao Xu (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

1. Syed Saiden Abbas (RUN) *Recognition of Shapes by Humans and Machines*
2. Michiel Christiaan Meulendijk (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*
3. Maya Sappelli (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*
4. Laurens Rietveld (VU) *Publishing and Consuming Linked Data*
5. Evgeny Sherkhonov (UVA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*
6. Michel Wilson (TUD) *Robust scheduling in an uncertain environment*
7. Jeroen de Man (VU) *Measuring and modeling negative emotions for virtual training*
8. Matje van de Camp (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*
9. Archana Nottamkandath (VU) *Trusting Crowdsourced Information on Cultural Artefacts*
10. George Karafotias (VUA) *Parameter Control for Evolutionary Algorithms*
11. Anne Schuth (UVA) *Search Engines that Learn from Their Users*
12. Max Knobbout (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*
13. Nana Baah Gyan (VU) *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach*
14. Ravi Khadka (UU) *Revisiting Legacy Software System Modernization*
15. Steffen Michels (RUN) *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments*
16. Guangliang Li (UVA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*
17. Berend Weel (VU) *Towards Embodied Evolution of Robot Organisms*
18. Albert Mero o Pe uela (VU) *Refining Statistical Data on the Web*
19. Julia Efremova (Tu/e) *Mining Social Structures from Genealogical Data*
20. Daan Odijk (UVA) *Context & Semantics in News & Web Search*
21. Alejandro Moreno C lleri (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*
22. Grace Lewis (VU) *Software Architecture Strategies for Cyber-Foraging Systems*
23. Fei Cai (UVA) *Query Auto Completion in Information Retrieval*
24. Brend Wanders (UT) *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*
25. Julia Kiseleva (TU/e) *Using Contextual Information to Understand Searching and Browsing Behavior*
26. Dilhan Thilakarathne (VU) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*
27. Wen Li (TUD) *Understanding Geo-spatial Information on Social Media*
28. Mingxin Zhang (TUD) *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*
29. Nicolas Höning (TUD) *Peak reduction in decentralised electricity systems -Markets and prices for flexible planning*
30. Ruud Mattheij (UvT) *The Eyes Have It*
31. Mohammad Khelghati (UT) *Deep web content monitoring*
32. Eelco Vriezekolk (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*
33. Peter Bloem (UVA) *Single Sample Statistics, exercises in learning from just one example*
34. Dennis Schunselaar (TUE) *Configurable Process Trees: Elicitation, Analysis, and Enactment*
35. Zhaochun Ren (UVA) *Monitoring Social Media: Summarization, Classification and Recommendation*
36. Daphne Karreman (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*
37. Giovanni Sileno (UvA) *Aligning Law and Action - a conceptual and computational inquiry*
38. Andrea Minuto (UT) *MATERIALS THAT MATTER - Smart Materials meet Art & Interaction Design*
39. Merijn Bruijnes (UT) *Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect*
40. Christian Detweiler (TUD) *Accounting for Values in Design*

41. Thomas King (TUD) *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*
42. Spyros Martzoukos (UVA) *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*
43. Saskia Koldijk (RUN) *Context-Aware Support for Stress Self-Management: From Theory to Practice*
44. Thibault Sellam (UVA) *Automatic Assistants for Database Exploration*
45. Bram van de Laar (UT) *Experiencing Brain-Computer Interface Control*
46. Jorge Gallego Perez (UT) *Robots to Make you Happy*
47. Christina Weber (UL) *Real-time foresight - Preparedness for dynamic innovation networks*
48. Tanja Buttler (TUD) *Collecting Lessons Learned*
49. Gleb Polevoy (TUD) *Participation and Interaction in Projects. A Game-Theoretic Analysis*
50. Yan Wang (UVT) *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*

2017

1. Jan-Jaap Oerlemans (UL) *Investigating Cybercrime*
2. Sjoerd Timmer (UU) *Designing and Understanding Forensic Bayesian Networks using Argumentation*
3. Daniël Harold Telgen (UU) *Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*
4. Mrunal Gawade (CWI) *MULTI-CORE PARALLELISM IN A COLUMN-STORE*
5. Mahdieh Shadi (UVA) *Collaboration Behavior*
6. Damir Vandic (EUR) *Intelligent Information Systems for Web Product Search*
7. Roel Bertens (UU) *Insight in Information: from Abstract to Anomaly*
8. Rob Konijn (VU) *Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*
9. Dong Nguyen (UT) *Text as Social and Cultural Data: A Computational Perspective on Variation in Text*
10. Robby van Delden (UT) *(Steering) Interactive Play Behavior*
11. Florian Kunneman (RUN) *Modelling patterns of time and emotion in Twitter #anticipointment*