Estimating Time to Event from Tweets Using Temporal Expressions

Ali Hürriyetoğlu, Nelleke Oostdijk, and Antal van den Bosch
Centre for Language Studies
Radboud University Nijmegen
P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands
{a.hurriyetoglu,n.oostdijk,a.vandenbosch}@let.ru.nl

Abstract

Given a stream of Twitter messages about an event, we investigate the predictive power of temporal expressions in the messages to estimate the time to event (TTE). From labeled training data we learn average TTE estimates of temporal expressions and combinations thereof, and define basic rules to compute the time to event from temporal expressions, so that when they occur in a tweet that mentions an event we can generate a prediction. We show in a case study on soccer matches that our estimations are off by about eight hours on average in terms of mean absolute error.

1 Introduction

Textual information streams such as those produced by news media and by social media reflect what is happening in the real world. These streams often contain explicit pointers to future events that may interest or concern a potentially large amount of people. Besides media-specific markers such as event-specific hashtags in messages on Twitter[1], these messages may contain explicit markers of place and time that help the receivers of the message disambiguate and pinpoint the event on the map and calendar.

The automated analysis of streaming text messages can play a role in catching these important events. Part of this analysis may be the identification of the future start time of the event, so that the event can be placed on the calendar and appropriate action may be taken by the receiver of the message, such as ordering tickets, planning a security operation, or starting a journalistic investigation. The automated identification of the time to event (TTE) should be as accurate and come early as possible. In this paper we explore a hybrid rule-based and data-driven method that exploits the explicit mentioning of temporal expressions to arrive at accurate and early TTE estimations.

The idea of publishing future calendars with potentially interesting events gathered (semi-) automatically for subscribers, possibly with personalization features and the option to harvest both social media and the general news, has been implemented already and is available through services such as Zapaday[2], Daybees[3], and Songkick[4]. 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.

In this study we do not use a rule-based temporal tagger such as the HeidelTime tagger (Strötgen and Gertz, 2013), which searches for only a limited set of temporal expressions. Instead, we propose an approach that uses a large set of temporal expressions, created by using seed terms and generative rules, and a training method that automatically determines the TTE estimate to be associated with each temporal expression sequence in a data-driven way. Typically, rule-based systems do not use the implicit information provided by adverbs ('more' in 'three more days') and relations between non-subsequent elements, while machine-learning-based systems do not make use of the temporal logic inherent to temporal expressions; they may identify 'three more days' as a temporal expression but they lack the logical apparatus to compute that this implies a TTE of about 3×24 hours. To make use of the best of both worlds we propose a hybrid system which uses information about the distribution of temporal expressions

[1] http://twitter.com

[2] http://www.zapaday.com

[3] http://daybees.com/

[4] https://www.songkick.com/

Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) @ EACL 2014, pages 8–16,
Gothenburg, Sweden, April 26-30 2014.

as they are used in forward-looking social media messages in a training set of known events, and combines this estimation method with an extensive set of regular expressions that capture a large space of possible Dutch temporal expressions.

Thus, our proposed system analyzes social media text to find information about future events, and estimates how long it will take before the event takes place. The service offered by this system will be useful only if it generates accurate estimations of the time to event. Preferably, these accurate predictions should come as early as possible. Moreover, the system should be able, in the long run, to freely detect relevant future events that are not yet on any schedule we know in any language represented on social media. For now, in this paper we focus on estimating the starting time of scheduled events, and use past and known events for a controlled experiment involving Dutch twitter messages.

For our experiment we collected tweets referring to scheduled Dutch premier league soccer matches. This type of event generally triggers many anticipatory discussions on social media containing many temporal expressions. Given a held-out soccer match not used during training, our system predicts the time to the event based on individual tweets captured in a range from eight days before the event to the event time itself. Each estimation is based on the temporal expressions which occur in a particular twitter message. The mean absolute error of the predictions for each of the 60 soccer matches in our data set is off by about eight hours. The results are generated in a leave-one-out cross-validation setup[5].

This paper starts with describing the relation of our work to earlier research in Section 2. Section 3 describes the overall experimental setup, including a description of the data, the temporal expressions that were used, our two baselines, and the evaluation method used. Next, in Section 4 the results are presented. The results are analyzed and discussed in Section 5. We conclude with a summary of our main findings and make suggestions for the direction future research may take (Section 6).

[5]Tweet ID's, per tweet estimations, occurred time expressions and rules can be found at `http://www.ru.nl/lst/resources/`

2 Related Work

Future-reference analysis in textual data has been studied from different angles. In the realm of information retrieval the task is more commonly defined as seeking future temporal references in large document collections such as the Web by means of time queries (Baeza Yates, 2005). Various studies have used temporal expression elements as features in an automatic setting to improve the relevance estimation of a web document (Dias et al., 2011; Jatowt and Au Yeung, 2011). Information relevant to event times has been the focus of studies such as those by Becker *et al*. (2012) and Kawai et al. (2010).

Our research is aimed at estimating the time to event of an upcoming event as precisely as possible. Radinsky et al. (2012) approach this problem by learning from causality pairs in texts from long-ranging news articles. Noro et al. (2006) describe a machine-learning-based system for the identification of the time period in which an event will happen, such as in the morning or at night.

Some case studies are focused on detecting events as early as possible as their unfolding is fast. The study by Sakaki et al. (2010) describes a system which analyzes the flow of tweets in time and place mentioning an earthquake, to predict the unfolding quake pattern which may in turn provide just-in-time alerts to people residing in the locations that are likely to be struck shortly. Zielinski et al. (2012) developed an early warning system to detect natural disasters in a multilingual fashion and thereby support crisis management. The quick throughput of news in the Twitter network is the catalyst in these studies focusing on natural disasters. In our study, we rather rely on the slower build-up of clues in messages in days before an event, at a granularity level of hours.

Ritter et al. (2012) aim to create a calendar of events based on explicit date mentions and words typical of the event. They train on annotated open domain event mentions and use a rule-based temporal tagger. We aim to offer a more generic solution that makes use of a wider range of temporal expressions, including indirect and implicit expressions.

Weerkamp and De Rijke (2012) study this type of more generic patterns of anticipation in tweets, but focus on personal future activities, while we aim to predict as early as possible the time to event of events that affect and interest many users.

Our estimations do not target time periods such as mornings or evenings but on the number of hours remaining to the event.

TTE estimation of soccer 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 soccer match. Estimating the time to event of future matches from tweet streams has been studied by Hürriyetoglu et al. (2013), using local regression over word time series. In a related study, Tops et al. (2013) use support vector machines to classify the time to event in automatically discretized categories. At best these studies are about a day off in their predictions. Both studies investigate the use of temporal expressions, but fail to leverage the utility of this information source, most likely because they use limited sets of less than 20 regular expressions. In this study we scale up the number of temporal expressions.

3 Experimental Set-Up

We carried out a controlled case study in which we focused on Dutch premier league soccer matches as a type of scheduled event. These types of games have the advantage that they occur frequently, have a distinctive hashtag by convention, and often generate thousands to several tens of thousands of tweets per match.

Below we first describe the collection and composition of our data sets (Subsection 3.1) and the temporal expressions which were used to base our predictions upon (Subsection 3.2). Then, in Subsection 3.3, we describe our baselines and evaluation method.

3.1 Data Sets

We harvested tweets from twiqs.nl[6], a database of Dutch tweets collected from December 2010 onwards. We selected the six best performing teams of the Dutch premier league in 2011 and 2012[7], and queried all matches in which these teams played against each other in the calendar years 2011 and 2012. The collection procedure resulted in 269,999 tweets referring to 60 individual matches. The number of tweets per event ranges from 321 to 35,464, with a median of 2,723 tweets. Afterwards, we restricted the data to tweets sent within eight days before the match[8] and eliminated all retweets. This reduced the number of tweets in our final data set to 138,141 tweets.

In this experiment we are working on the assumption that the presence of a hashtag can be used as proxy for the topic addressed in a tweet. Inspecting a sample of tweets referring to recent soccer games not part of our data set, we developed the hypothesis that the position of the hashtag may have an effect as regards the topicality of the tweet. Hashtags that occur in final position (i.e. they are tweet-final or are only followed by one or more other hashtags) are typically metatags and therefore possibly more reliable as topic identifiers than tweet non-final hashtags which behave more like common content words in context. In order to be able to investigate the possible effect that the position of the hashtag might have, we split our data in the following two subsets:

FIN – comprising tweets in which the hashtag occurs in final position (as defined above); 84,533 tweets.

NFI – comprising tweets in which the hashtag occurs in non-final position; 53,608 tweets.

Each tweet in our data set has a time stamp of the moment (in seconds) it was posted. Moreover, for each soccer match we know exactly when it took place. This information is used to calculate for each tweet the actual time that remains to the start of the event and the absolute error in estimating the time to event.

3.2 Temporal Expressions

In the context of this paper temporal expressions are considered to be words or phrases which point to the point in time, the duration, or the frequency of an event. These may be exact, approximate, or even right out vague. Although in our current experiment we restrict ourselves to an eight-day period prior to an event, we chose to create a gross list of all possible temporal expressions we could think of, so that we would not run the risk of overlooking any items and the list can be used on future occasions even when the experimental setting is different. Thus the list also includes temporal expressions that refer to points in time outside the time span under investigation here, such

[6] http://twiqs.nl

[7] Ajax, Feyenoord, PSV, FC Twente, AZ Alkmaar, and FC Utrecht.

[8] An analysis of the tweet distribution shows that the eight-day window captures about 98% of the tweets in the larger data set from which it was derived.

as *gisteren* 'yesterday' or *over een maand* 'in a month from now', and items indicating duration or frequency such as *steeds* 'continuously'/'time and again'. No attempt has been made to distinguish between items as regards time reference (future time, past time) as many items can be used in both fashions (compare for example *vanmiddag* in *vanmiddag ga ik naar de wedstrijd* 'this afternoon I'm going to the match' vs *ik ben vanmiddag naar de wedstrijd geweest* 'I went to the match this afternoon'.

The list is quite comprehensive. Among the items included are single words, e.g. adverbs such as *nu* 'now', *zometeen* 'immediately', *straks* 'later on', *vanavond* 'this evening', nouns such as *zondagmiddag* 'Sunday afternoon', and conjunctions such as *voordat* 'before'), but also word combinations and phrases such as *komende woensdag* 'next Wednesday. Temporal expressions of the latter type were obtained by means of a set of 615 seed terms and 70 rules, which generated a total of around 53,000 temporal expressions. In addition, there are a couple of hundred thousand temporal expressions relating the number of minutes, hours, days, or time of day;[9] they include items containing up to 9 words in a single temporal expression. Notwithstanding the impressive number of items included, the list is bound to be incomplete.

We included prepositional phrases rather than single prepositions so as to avoid generating too much noise. Many prepositions have several uses: they can be used to express time, but also for example location. Compare *voor* in *voor drie uur* 'before three o'clock' and *voor het stadion* 'in front of the stadium'. Moreover, prepositions are easily confused with parts of separable verbs which in Dutch are abundant.

Various items on the list are inherently ambiguous and only in one of their senses can be considered temporal expressions. Examples are *week* 'week' but also 'weak' and *dag* 'day' but also 'goodbye'. For items like these, we found that the different senses could fairly easily be distinguished whenever the item was immediately preceded by an adjective such as *komende* and *volgende* (both meaning 'next'). For a few highly frequent items this proved impossible. These are words like *zo* which can be either a temporal adverb ('in a minute'; cf. *zometeen*) or an intensifying adverb ('so'), *dan* 'then' or 'than', and *nog* 'yet' or 'another'. As we have presently no way of distinguishing between the different senses and these items have at best an extremely vague temporal sense so that they cannot be expected to contribute to estimating the time to event, we deciced to discard these.[10]

In order to capture event targeted expressions, we treated domain terms such as *wedstrijd* 'soccer match' as parts of temporal expressions in case they co-occur with a temporal expression.

For the items on the list no provisions were made for handling any kind of spelling variation, with the single exception of a small group of words (including *'s morgens* 'in the morning', *'s middags* 'in the afternoon' and *'s avonds* 'in the evening') which use in their standard spelling the archaic *'s* and abbreviations. As many authors of tweets tend to spell these words as *smorgens*, *smiddags* and *savonds* we decided to include these forms as well.

The items on the list that were obtained through generation include temporal expressions such as *over 3 dagen* 'in 2 days', *nog 5 minuten* 'another 5 minutes', but also fixed temporal expressions such as clock times.[11] The rules handle frequently observed variations in their notation, for example *drie uur* 'three o'clock' may be written in full or as *3:00*, *3:00 uur*, *3 u*, *15.00*, etc.

Table 1 shows example temporal expression estimates and applicable rules. The median estimations are mostly lower than the mean estimations. The distribution of the time to event (TTE) for a single temporal expression often appears to be skewed towards lower values. The final column of the table displays the applicable rules. The first six rules subtract the time the tweet was posted (TT) from an average marker point, heuristically determined, such as 'today 20.00' (i.e. 8 pm) for *vanavond* 'tonight'. The second and third rules from below state a TTE directly, again heuristically set – *over 2 uur* 'in 2 hours' is directly translated to a TTE of 2.

3.3 Evaluation and Baselines

Our approach to TTE estimation makes use of all temporal expressions in our temporal expression list that are found to occur in the tweets. A

[9] For examples see Table 1 and Section 3.3.

[10] Note that *nog* does occur on the list as part of various multiword expressions. Examples are *nog twee dagen* 'another two days' and *nog 10 min* '10 more minutes'.

[11] Dates are presently not covered by our rules but will be added in future.

Temporal Expression	Gloss	Mean TTE	Median TTE	Rule
vandaag	*today*	5.63	3.09	today 15:00 - TT h
vanavond	*tonight*	8.40	4.78	today 20:00 - TT h
morgen	*tomorrow*	20.35	18.54	tomorrow 15:00 - TT h
zondag	*Sunday*	72.99	67.85	Sunday 15:00 - TT h
vandaag 12.30	*today 12.20*	2.90	2.75	today 12:30 - TT h
om 16.30	*at 16.30*	1.28	1.36	today 16:30 - TT h
over 2 uur	*in 2 hours*	6.78	1.97	2 h
nog minder dan 1 u	*within 1 h*	21.43	0.88	1 h
in het weekend	*during the weekend*	90.58	91.70	*No Rule*

Table 1: Examples of temporal expressions and their mean and median TTE estimation from training data. The final column lists the applicable rule, if any. Rules make use of the time of posting (Tweet Time, TT).

match may be for a single item in the list (e.g. *zondag* 'Sunday') or any combination of items (e.g. *zondagmiddag*, *om 14.30 uur*, 'Sunday afternoon', 'at 2.30 pm'). There can be other words in between these expressions. We consider the longest match, from left to right, in case we encounter any overlap.

The experiment adopts a leave-one-out cross-validation setup. Each iteration uses all tweets from 59 events as training data. All tweets from the single held-out event are used as test set.

In the FIN data set there are 42,396 tweets with at least one temporal expression, in the NFI data set this is the case for 27,610 tweets. The number of tweets per event ranges from 66 to 7,152 (median: 402.5; mean 706.6) for the FIN data set and from 41 to 3,936 (median 258; mean 460.1) for the NFI data set.

We calculate the TTE estimations for every tweet that contains at least one of the temporal expression or a combination in the test set. The estimations for the test set are obtained as follows:

1. For each match (a single temporal expression or a combination of temporal expressions) the mean or median value for TTE is used that was learned from the training set;
2. Temporal expressions that denote an exact amount of time are interpreted by means of rules that we henceforth refer to as Exact rules. This applies for example to temporal expressions answering to patterns such as *over N* {*minuut* | *minuten* | *kwartier* | *uur* | *uren* | *dag* | *dagen* | *week*} 'in N {minute | minutes | quarter of an hour | hour | hours | day | days | week}'. Here the TTE is assumed to be the same as the N minutes, days or whatever is mentioned. The rules take precedence over the mean estimates learned from the training set;
3. A second set of rules, referred to as the Dynamic rules, is used to calculate the TTE dynamically, using the temporal expression and the tweet's time stamp. These rules apply to instances such as *zondagmiddag om 3 uur* 'Sunday afternoon at 3 p.m.'. Here we assume that this is a future time reference on the basis of the fact that the tweets were posted prior to the event. With temporal expressions that are underspecified in that they do not provide a specific point in time (hour), we postulate a particular time of day. For example, *vandaag* 'today' is understood as 'today at 3 p.m., *vanavond* 'this evening' as 'this evening at 8 p.m. and *morgenochtend* 'tomorrow morning' as 'tomorrow morning at 10 a.m.'. Again, as was the case with the first set of rules, these rules take precedence over the mean or median estimates learned from the training data.

The results for the estimated TTE are evaluated in terms of the absolute error, i.e. the absolute difference in hours between the estimated TTE and the actual remaining time to the event.

We established two naive baselines: the mean and median TTE measured over all tweets of FIN and NFI datasets. These baselines reflect a best guess when no information is available other than tweet count and TTE of each tweet. The mean TTE is 22.82 hours, and the median TTE is 3.63 hours before an event. The low values of the

baselines, especially the low median, reveal the skewedness of the data: most tweets referring to a soccer event are posted in the hours before the event.

4 Results

Table 2 lists the overall mean absolute error (in number of hours) for the different variants. The results are reported separately for each of the two data sets (FIN and NFI) and for both sets aggregated (FIN+NFI). For each of these three variants, the table lists the mean absolute error when only the basic data-driven TTE estimations are used ('Basic'), when the Exact rules are added ('+Ex.'), when the Dynamic rules are added ('+Dyn'), and when both types of rules are added. The coverage of the combination (i.e. the number of tweets that match the expressions and the rules) is listed in the bottom row of the table.

A number of observations can be made. First, all training methods perform substantially better than the two baselines in all conditions. Second, the TTE training method using the median as estimation produces estimations that are about 1 hour more accurate than the mean-based estimations.

Third, adding Dynamic rules has a larger positive effect on prediction error than adding Exact rules. The bottom row in the table indicates that the rules do not increase the coverage of the method substantially. When taken together and added to the basic TTE estimation, the Dynamic and Exact rules do improve over the Basic estimation by two to three hours.

Finally, although the differences are small, Table 2 reveals that training on hashtag-final tweets (FIN) produces slightly better overall results (7.62 hours off at best) than training on hashtag-non-final tweets (8.50 hours off) or the combination (7.99 hours off), despite the fact that the training set is smaller than that of the combination.

In the remainder of this section we report on systems that use all expressions and Exact and Dynamic rules.

Whereas Table 2 displays the overall mean absolute errors of the different variants, Figure 1 displays the results in terms of mean absolute error at different points in time before the event, averaged over periods of one hour, for the two baselines and the FIN+NFI variant with the two training methods (i.e. taking the mean versus the median of the observed TTEs for a particular temporal expression). In contrast to Table 2, in which only a mild difference could be observed between the median and mean variants of training, the figure shows a substantial difference. The estimations of the median training variant are considerably more accurate than the mean variant up to 24 hours before the event, after which the mean variant scores better. By virtue of the fact that the data is skewed (most tweets are posted within a few hours before the event) the two methods attain a similar overall mean absolute error, but it is clear that the median variant produces considerably more accurate predictions when the event is still more than a day away.

While Figure 1 provides insight into the effect of median versus mean-based training with the combined FIN+NFI dataset, we do not know whether training on either of the two subsets is advantageous at different points in time. Table 3 shows the mean absolute error of systems trained with the median variant on the two subsets of tweets, FIN and NFI, as well as the combination FIN+NFI, split into nine time ranges. Interestingly, the combination does not produce the lowest errors close to the event. However, when the event is 24 hours away or more, both the FIN and NFI systems generate increasingly large errors, while the FIN+NFI system continues to make quite accurate predictions, remaining under 10 hours off even for the longest TTEs, confirming what we already observed in Figure 1.

TTE range (h)	FIN	NFI	FIN+NFI
0	**2.58**	3.07	8.51
1–4	**2.38**	2.64	8.71
5–8	**3.02**	3.08	8.94
9–12	**5.20**	5.47	6.57
13–24	5.63	**5.54**	6.09
25–48	13.14	15.59	**5.81**
49–96	17.20	20.72	**6.93**
97–144	30.38	41.18	**6.97**
> 144	55.45	70.08	**9.41**

Table 3: Mean Absolute Error for the FIN, NFI, and FIN+NFI systems in different TTE ranges.

5 Analysis

One of the results observed in Table 2 was the relatively limited role of Exact rules, which were intended to deal with exact temporal expressions such as *nog 5 minuten* '5 more minutes' and *over*

System	FIN				NFI				FIN+NFI			
	Basic	+Ex.	+Dyn.	+Both	Basic	+Ex.	+Dyn.	+Both	Basic	+Ex.	+Dyn.	+Both
Baseline Median	21.09	21.07	21.16	21.14	18.67	18.72	18.79	18.84	20.20	20.20	20.27	20.27
Baseline Mean	27.29	27.29	27.31	27.31	25.49	25.50	25.53	25.55	26.61	26.60	26.63	26.62
Training Median	10.38	10.28	7.68	7.62	11.09	11.04	8.65	8.50	10.61	10.54	8.03	7.99
Training Mean	11.62	11.12	8.73	8.29	12.43	11.99	9.53	9.16	11.95	11.50	9.16	8.76
Coverage	31,221	31,723	32.240	32,740	18,848	19,176	19,734	20,061	52,186	52,919	53,887	54,617

Table 2: Overall Mean Absolute Error for each method: difference in hours between the estimated time to event and the actual time to event, computed separately for the FIN and NFI subsets, and for the combination. For all variants a count of the number of matches is listed in the bottom row.

een uur 'in one hour'. This can be explained by the fact that as long as the temporal expression is related to the event we are targeting, the point in time is denoted exactly by the temporal expression and the estimation obtained from the training data (the 'Basic' performance) will already be accurate, leaving no room for the rules to improve on this. The rules that deal with dynamic temporal expressions, on the other hand, have quite some impact.

As was explained in Section 3.2 our list of temporal expressions was a gross list, including items that were unlikely to occur in our present data. In all we observed 770 of the 53,000 items listed, 955 clock time rule matches, and 764 time expressions which contain number of days, hours, minutes etc. The temporal expressions observed most frequently in our data are:[12] *vandaag* 'today' (10,037), *zondag* 'Sunday' (6,840), *vanavond* 'tonight' (5167), *straks* 'later on' (5,108), *vanmiddag* 'this afternoon' (4,331), *matchday* 'match day' (2,803), *volgende week* 'next week' (1,480) and *zometeen* 'in a minute' (1,405).

Given the skewed distribution of tweets over the eight days prior to the event, it is not surprising to find that nearly all of the most frequent items refer to points in time within close range of the event. Apart from *nu* 'now', all of these are somewhat vague about the exact point in time. There are, however, numerous items such as *om 12:30 uur* 'at half past one' and *over ongeveer 45 minuten* 'in about 45 minutes') which are very specific and therefore tend to appear with middle to low frequencies.[13] And while it is possible to state an exact point in time even when the event is in the more distant future, we find that there is a clear tendency to use underspecified temporal expressions as the event is still some time away. Thus, rather than *volgende week zondag om 14.30 uur* 'next week Sunday at 2.30 p.m.' just *volgende week* is used, which makes it harder to estimate the time to event.

Closer inspection of some of the temporal expressions which yielded large absolute errors suggests that these may be items that refer to subevents rather than the main event (i.e. the match) we are targeting. Examples are *eerst* 'first', *daarna* 'then', *vervolgens* 'next', and *voordat* 'before'.

6 Conclusions and Future Work

We have presented a method for the estimation of the TTE from single tweets referring to a future event. In a case study with Dutch soccer matches, we showed that estimations can be as accurate as about eight hours off, averaged over a time window of eight days. There is some variance in the 60 events on which we tested in a leave-one-out validation setup: errors ranged between 4 and 13 hours, plus one exceptionally badly predicted event with a 34-hour error.

The best system is able to stay within 10 hours of prediction error in the full eight-day window. This best system uses a large set of hand-designed temporal expressions that in a training phase have each been linked to a median TTE with which they occur in a training set. Together with these data-driven TTE estimates, the system uses a set of rules that match on exact and indirect time references. In a comparative experiment we showed that this combination worked better than only having the data-driven estimations.

We then tested whether it was more profitable to train on tweets that had the event hashtag at the end, as this is presumed to be more likely a meta-

[12] The observed frequencies can be found between brackets.

[13] While an expression such as *om 12:30 uur* has a frequency of 116, *nog maar 8 uur en 35 minuten* 'only 8 hours and 35 minutes from now' has a frequency of 1.

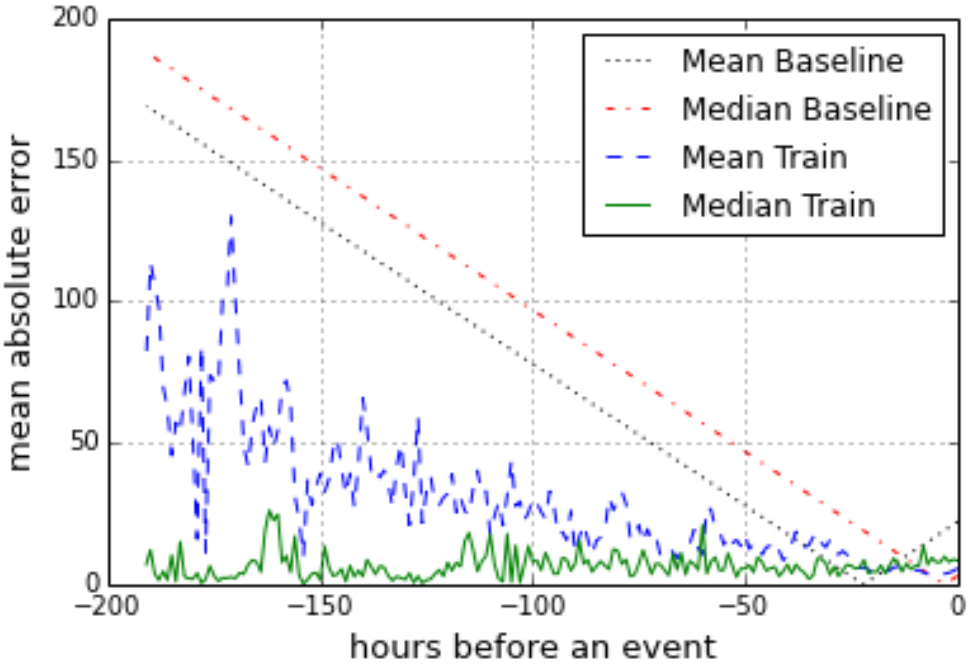

Figure 1: Curves showing the absolute error (in hours) in estimating the time to event over an 8-day period (-192 to 0 hours) prior to the event. The two baselines are compared to the TTE estimation methods using the mean and median variant.

tag, and thus a more reliable clue that the tweet is about the event than when the hashtag is not in final position. Indeed we find that the overall predictions are more accurate, but only in the final hours before the event (when most tweets are posted). 24 hours and earlier before the event it turns out to be better to train both on hashtag-final and hashtag-non-final tweets.

Finally, we observed that the two variants of our method of estimating TTEs for single temporal expressions, taking the mean or the median, leads to dramatically different results, especially when the event is still a few days away—when an accurate time to event is actually desirable. The median-based estimations, which are generally smaller than the mean-based estimations, lead to a system that largely stays under 10 hours of error.

Our study has a number of logical extensions into future research. First, our method is not bound to a single type of event, although we tested it in a controlled setting. With experiments on tweet streams related to different types of events the general applicability of the method could be tested: can we use the trained TTE estimations from our current study, or would we need to retrain per event type?

Second, we hardcoded a limited number of frequent spelling variations, where it would be a more generic solution to rely on a more systematic spelling normalization preprocessing step.

Third, so far we did not focus on determining the relevance of temporal expressions in case there are several time expressions in a single message; we treated all occurred temporal expressions as equally contributing to the estimation. Identifying which temporal expressions are relevant in a single message is studied by Kanhabua et al. (2012).

Finally, our method is limited to temporal expressions. For estimating the time to event on the basis of tweets that do not contain temporal expressions, we could benefit from term-based approaches that consider any word or word n-gram as potentially predictive (Hürriyetoglu et al., 2013).

Acknowledgment

This research was supported by the Dutch national programme COMMIT as part of the Infiniti project.

References

Ricardo Baeza Yates. 2005. Searching the future. In *In ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval (MF/IR 2005)*.

Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. 2012. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 533–542, New York, NY, USA. ACM.

Gaël Dias, Ricardo Campos, and Alípio Jorge. 2011. Future retrieval: What does the future talk about? In *In Proceedings SIGIR2011 Workshop on Enriching Information Retrieval (ENIR2011)*.

Ali Hürriyetoglu, Florian Kunneman, and Antal van den Bosch. 2013. Estimating the time between twitter messages and future events. In *DIR*, pages 20–23.

Adam Jatowt and Ching-man Au Yeung. 2011. Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1259–1264, New York, NY, USA. ACM.

Nattiya Kanhabua, Sara Romano, and Avaré Stewart. 2012. Identifying relevant temporal expressions for real-world events. In *Proceedings of The SIGIR 2012 Workshop on Time-aware Information Access, Portland, OR*.

Hideki Kawai, Adam Jatowt, Katsumi Tanaka, Kazuo Kunieda, and Keiji Yamada. 2010. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, pages 25:1–25:10, New York, NY, USA. ACM.

Florian A Kunneman and Antal van den Bosch. 2012. Leveraging unscheduled event prediction through mining scheduled event tweets. *BNAIC 2012 The 24th Benelux Conference on Artificial Intelligence*, page 147.

Taichi Noro, Takashi Inui, Hiroya Takamura, and Manabu Okumura. 2006. Time period identification of events in text. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1153–1160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 909–918, New York, NY, USA. ACM.

Alan Ritter, Oren Etzioni Mausam, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, KDD '12, pages 1104–1112, New York, NY, USA. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, Jun.

Hannah Tops, Antal van den Bosch, and Florian Kunneman. 2013. Predicting time-to-event from twitter messages. *BNAIC 2013 The 24th Benelux Conference on Artificial Intelligence*, pages 207–2014.

Wouter Weerkamp and Maarten De Rijke. 2012. Activity prediction: A twitter-based exploration. In *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access, TAIA-2012*, August.

Andrea Zielinski, Ulrich Bügel, L. Middleton, S. E. Middleton, L. Tokarchuk, K. Watson, and F. Chaves. 2012. Multilingual analysis of twitter news in support of mass emergency events. In A. Abbasi and N. Giesen, editors, *EGU General Assembly Conference Abstracts*, volume 14 of *EGU General Assembly Conference Abstracts*, pages 8085+, April.