The (Un)Predictability of Emotional Hashtags in Twitter

Florian Kunneman*, **Christine Liebrecht****, **and Antal van den Bosch***

*Centre for Language Studies
Radboud University Nijmegen
{f.kunneman,a.vandenbosch}@let.ru.nl

**Faculty of Social and Behavioral Sciences
University of Amsterdam
c.c.liebrecht@uva.nl

Abstract

Hashtags in Twitter posts may carry different semantic payloads. Their dual form (word and label) may serve to categorize the tweet, but may also add content to the message, or strengthen it. Some hashtags are related to emotions. In a study on emotional hashtags in Dutch Twitter posts we employ machine learning classifiers to test to what extent tweets that are stripped from their hashtag could be reassigned to this hashtag. About half of the 24 tested hashtags can be predicted with AUC scores of .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 .7 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 in the tweet.

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 Twitter messages (Ritter et al., 2010; Dann, 2010). Many Twitter messages (tweets) express an emotion of the sender: according to Roberts *et al.* (2012), 43 percent 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 et al., 2005; Aman and Szpakowicz, 2007). This approach is known to result in high-precision annotated data, but is labor-intensive and time-consuming.

The second approach is to use the annotations that Twitter users themselves add to a tweet: hashtags. A hashtag (a word prefixed by the typographical hashmark #) 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 categorization, most of which are conversational (Huang et al., 2010). Hashtags expressing emotions are often used in tweets and are therefore potentially useful annotations for training data. Wang et al. (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.

Therefore, using hashtags as annotated training data may 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.

Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) @ EACL 2014, pages 26–34,
Gothenburg, Sweden, April 26-30 2014.

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 research lies in the fact that we offer 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.

2 Related research

Leveraging uncontrolled labeling to obtain large amounts of training data is referred to as *distant supervision* (Mintz et al., 2009). With its conventions for hashtags 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 specific 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, Tsur and Rappoport (2010) 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 and Ide, 2013; 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 and 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. (2010) 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

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

classification is performed, distinguishing the different target emotions and optionally an emotionally neutral class (Purver and Battersby, 2012; 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 and Riloff, 2013; Suttles and 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 ten-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 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.

3 Approach

Our approach is to train a machine learning classifier on tweets containing an emotion-bearing hashtag and an equal amount of random tweets as counter-examples, resulting in a balanced binary classifier for the hashtag (which 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 hidden from the classifier. Second, the tweets not containing 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 linked to the hashtag.

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 containing 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 to what extent 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 generalizable models of emotion in Twitter that are not restricted to the occurrence of a hashtag.

4 Experimental setup

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 `twiqs.nl`, a database of Dutch tweets from December 2010 onwards (Tjong Kim Sang and van den Bosch, 2013). 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 containing any of the hashtags were collected through `twiqs.nl` from the time frame of December 2010 up 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, and the rest was used as test data.

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 pre-processed by extracting word unigrams, bigrams, and trigrams as features. We maintained capitalization 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 usernames and URLs were normalized to dummy values. All features containing a target hashtag were removed.

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 three.

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.

5 Results

5.1 Hashtag predictability

The results of our classifiers labeling a large sample of tweets posted on February 1, 2013 are listed in Table 1. Each line with a target hashtag represents a separate experiment. The amount 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 have 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 overprediction, 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 .80 or more. The influence of the amount of training data on the AUC score seems peripheral. Furthermore, there is no clear difference in 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 should therefore 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 dissimilar score, while #leuk and #stom are rather

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 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

similar.

5.2 Emotion detection

The second evaluation is based on the manual annotation of the 250 tweets 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 analyze 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 86% and 89%, and AUC scores of 90% to 91%. The latter was included as a comparison, expecting 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 conveyes 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 uneagerness
- #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 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 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 1, provides further insight into the emotion detection quality in relation to the confidence ranks. Precisions at higher rank cutoffs tend to peak early (indicating that the first top-ranked tweets fit the hashtag best), and decrease slowly or reach a plateau.

	Precision		Cohen's Kappa	Mutual F-score
	(67% majority)	(100% majority)		
#zinin	.75	.35	.09	.67
#geenzin	.31	.21	.60	.73
#fml	.69	.46	.48	.81
#omg	.49	.25	.29	.67

Table 2: Precision of correct hashtag predictions of the top 250 'false positives' based on human annotations

The twofold 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 utilizable data for emotion detection. Tweets retrieved based on #geenzin seem to have a less overt relation to the emotion of uneagerness, although other cues (such as topical words indirectly related to the emotion) lead to a fairly correct recovery of tweets that had 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.

6 Discussion

6.1 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 analyze 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 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 3. 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 the person is looking forward to. Also, Dutch positive adjectives such as 'lekker' ('nice') and 'gezellig' (multiple translations[2]), 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.

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

6.2 Emotional cues in Twitter

In contrast to spoken or face-to-face communication, Twitter does not allow for the use of special intonation or facial expressions to mark a message. However, authors on Twitter have other cues at their disposal. Previous studies show, for example, that they might mark the irony or sarcasm in their message by using linguistic markers such

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

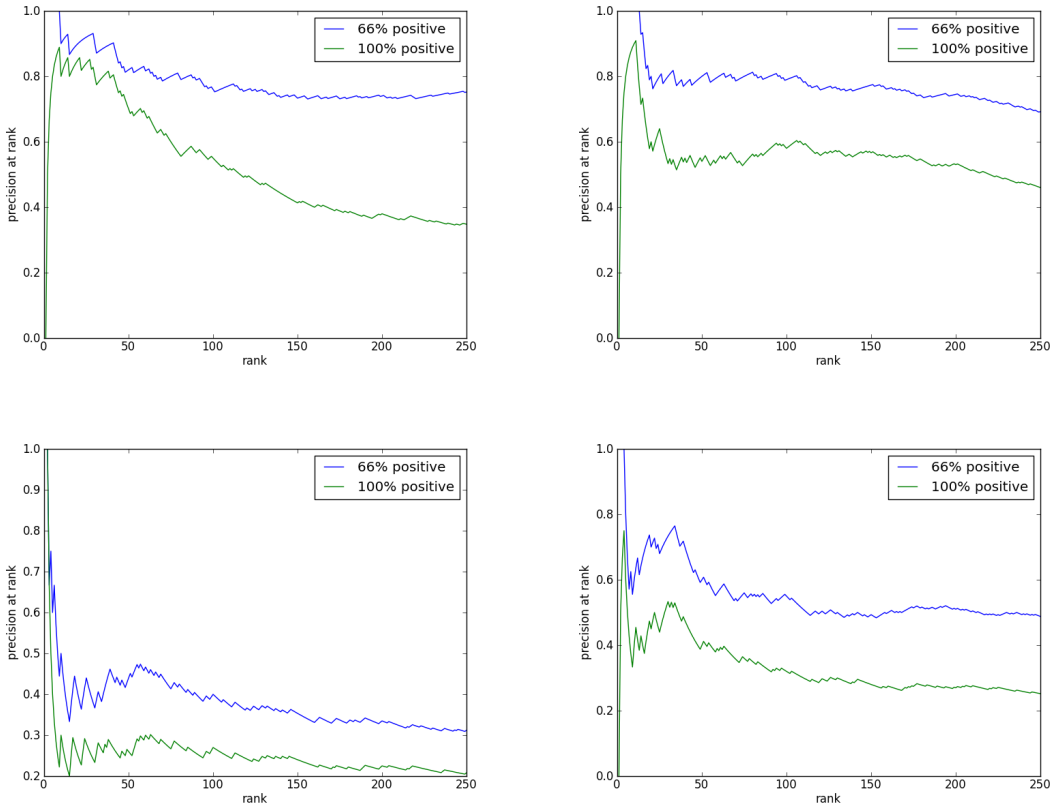

Figure 1: Precision at $\{1 \ldots 250\}$ on the classes #zinin (top left), #fml (top right), #geenzin (bottom left), and #omg (bottom right).

as hyperboles, exclamations and emoticons to help readers to correctly interpret the message (Burgers et al., 2012; Liebrecht et al., 2013). We argue that this is also the case for emotional messages.

Tweets are written messages with a strongly restricted length. Authors compensate the lack of non-verbal cues by adding emotion markers. This hypothesis is supported by research in the field of Computer-Mediated Communication (CMC), where many studies have been carried out on (the lack of) non-verbal emotional cues in (electronic) messages. Walther (1992) introduced the Social Information Processing Perspective: a theory that users can develop relationships via CMC if they have sufficient time and message exchanges and if communicative cues, such as non-verbal emotional cues, are available. He argues that humans easily switch between verbal and non-verbal cues. Based on previous studies, Walther distinguishes textual cues that express affection: relational icons (emoticons, see Asteroff, 1987), electronic paralanguage (such as intentional misspelling (*veeeery*), capitalization (*NICE*), repetition of exclamation marks (*good!!!!!*) and lexical surrogates for vocal segregates (*hmmm*) (Carey, 1980). Later he also recognizes emoticons as nonverbal emotion cues (Walther and D'Addario, 2001). Emoticons can serve many purposes, one of which is expressing emotions (Agarwal et al., 2011; Davidov et al., 2010).

7 Conclusion

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

	Example	Percentage in top 150 features			
		#zinin	#fml	#geenzin	#omg
emotion hashtag	'#foreveralone'	6.67%	10.00%	2.67%	18.67%
emoticon	':S'	0.00%	4.67%	0.00%	6.67%
exclamation	'noooo'	0.00%	2.67%	0.00%	8.67%
state of being	'curious'	3.33%	7.33%	3.33%	0.67%
temporal reference	'moment'	26.00%	7.33%	10.00%	1.33%
topic	'dentist'	52.67%	48.67%	69.33%	25.33%
other	'ready_to'	11.33%	19.33%	14.67%	38.67%

Table 3: Shares (in percentages) of seven categories in the top-150 highest-weighted features for four hashtags.

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, indicating 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 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 et al., 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 considerate 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. Whether the degree of consistency, along with an intensifying or emotion adding deployment, can be deduced from the inherent properties of an emotion hashtag is open for future research.

Acknowledgments

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

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer.

Christian Burgers, Margot van Mulken, and Peter Jan Schellens. 2012. Verbal irony differences in usage across written genres. *Journal of Language and Social Psychology*, 31(3):290–310.

John Carey. 1980. Paralanguage in computer mediated communication. In *Proceedings of the 18th annual meeting on Association for Computational Linguistics*, pages 67–69. Association for Computational Linguistics.

Stephen Dann. 2010. Twitter content classification. *First Monday*, 15(12).

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

T. Fawcett. 2004. ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, Hewlett Packard Labs.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586.

Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM.

Sepandar D Kamvar and Jonathan Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 117–126. ACM.

Suin Kim, JinYeong Bak, and Alice Haeyun Oh. 2012. Do you feel what i feel? social aspects of emotions in twitter conversations. In *ICWSM*.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 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*, pages 29–37, Atlanta, Georgia, June. Association for Computational Linguistics.

N. Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Saif M Mohammad. 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*, pages 246–255. Association for Computational Linguistics.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

Robert Plutchik. 1980. *Emotion: A psychoevolutionary synthesis*. Harper & Row New York.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Ashequl Qadir and Ellen Riloff. 2013. Bootstrapped learning of emotion hashtags# hashtags4you. In *WASSA 2013*, page 2.

Alan Ritter, Colin Cherry, and Bill Dolan. 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*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3806–3813.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer.

Erik Tjong Kim Sang and Antal van den Bosch. 2013. Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134, 12/2013.

Joseph B Walther and Kyle P D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3):324–347.

Joseph B Walther. 1992. Interpersonal effects in computer-mediated interaction a relational perspective. *Communication research*, 19(1):52–90.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.