Anticipointment Detection in Event Tweets

Kunneman, Florian A.; van Mulken, Margot; van den Bosch, A.

published in
International Journal on Artificial Intelligence Tools
2020

DOI (link to publisher)
10.1142/S0218213020400011

document version
Peer reviewed version

International Journal on Artificial Intelligence Tools
Vol. 29, No. 2 (2020) 2040001 (28 pages)

DOI: 10.1142/S0218213020400011

World Scientific
www.worldscientific.com

Anticipointment Detection in Event Tweets

F. Kunneman

Faculty of Science, VU University
De Boelelaan 1111, NL-1081 HV Amsterdam, The Netherlands
f.a.kunneman@vu.nl

M. van Mulken

Centre for Language Studies, Radboud University
Erasmuslaan 1, NL-6500 HD Nijmegen, The Netherlands
m.vanmulken@let.ru.nl

A. van den Bosch

KNAW Meertens Institute and Faculty of Humanities, University of Amsterdam
Oudezijds Achterburgwal 185, NL-1012 DK Amsterdam, The Netherlands
antal.van.den.bosch@meertens.knaw.nl

Received 15 February 2019
Accepted 23 March 2020
Published 31 March 2020

We developed a system to detect 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 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 dataset of Dutch Twitter posts, and modeled classifiers to detect 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 dataset, we summarized the collective emotions for over 3000 events as the percentage of tweets classified as positive expectation (in anticipation), disappointment and satisfaction (in hindsight). Only a weak correlation of around 0.2 was found between positive expectation and disappointment, while a higher correlation of 0.6 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 favored sports team had lost. We conclude that senders of Twitter posts might be more inclined to share satisfaction than disappointment after a much anticipated event.

Keywords: Event detection; emotion detection; Twitter.

1. Introduction

'What's happening?' is the question that the social media platform of Twitter[a] 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. 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 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. Furthermore, disappointment may be more likely if expectations are high, while satisfaction is highest when expectations are exceeded.[1] The vernacular word 'anticipointment'[b] is used to refer to sensations where a considerable contrast holds between expectation and outcome. Although social events are oftentimes referred to on Twitter, both in anticipation and in hindsight, the manifestation of these interrelated emotions on Twitter has not been studied to date. By targeting a large number of Twitter posts, we propose a novel approach 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. This research is a continuation of existing work on automatic event extraction from Twitter and emotion detection.[2,3] We extract a large number of events from the Dutch Twitter verse and classify the emotion in the event-referring tweets. Based on the observed emotion in tweets before and after events, we measure the extent to which positive expectation, disappointment, and satisfaction are correlated and analyze the events that are most exemplary of anticipointment, or of positive expectation followed by satisfaction.[c]

2. Related Work

2.1. *Event-related emotions*

In describing emotions related to anticipation, Miceli and Castelfranchi[1] 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. Miceli and Castelfranchi categorize

[a]www.twitter.com
[b]http://nancyfriedman.typepad.com/away_with_words/2011/12/word-of-the-week-anticipointment.html
[c]The training models, emotion classifications and emotion patterns before and after events may be downloaded from https://hdl.handle.net/2066/174261

the emotion of positive expectation as IAR and define it 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.

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[4] 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. Van Boven and Ashworth[5] 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.[6] In addition, in a social platform such as Twitter, emotion might operate at different levels, such as individual, group or cultural.[7] Although tweets might not convey a clear collective emotion before or after an event, some events stir more emotional tweets than others. Our analysis of the extent to which post-event emotions on Twitter can be explained by the emotion beforehand is based on this premise.

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.[8–11] The second selects the most basic emotions as defined in psychological literature, such as the six basic emotions of Ekman[12–17] or the wheel of eight primary bipolar emotions of Plutchik.[18,19] The third targets

any emotion that might be present in the data that is studied.[20,21] 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.[14,15] 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 and deduce probable labelings from highly indicative features in the data.[22] 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.[3] 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,[13,19] collecting additional hashtags that convey the same emotion,[17,23] and classifying emotion in tweets that do not contain a hashtag.[10,16,3,20] 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,[8] lexicons such as WordNet Affect[20] and General Inquirer,[14] and patterns of Highly Frequent Words and Content Words.[10] 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 used in several studies.[13,19,17] 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.

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, Mao and Pepe[24] 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.*[25] 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 Faltings[26] 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[27] 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 *et al.*[28] apply a lexicon-based approach to identify eight different emotions in tweets reacting to 25 selected events. They analyze several events on the distribution of the detected emotions. Brooks, Robinson, Torkildson and Aragon[29] aim to facilitate collaborative visual analysis of event tweets and visualize 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[30] and Chen, Argueta and Chang[31] analyze tweets that refer to automatically detected events. Thelwall, Buckley and Paltoglou propose an approach to identify the top 30 words with the most 'bursty' time pattern as events from a month of English tweets.[30] The sentiment strength of these tweets was classified and analyzed for significant fluctuations in sentiment polarity. Chen, Argueta and Chang also apply burstiness-based event detection, and classify tweets that refer to the detected events into six emotion categories.[31]

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.* and Chen *et al.*[30,31] 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 analyze the relation between emotion in pre-event and post-event Twitter messages.

3. Dataset

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.

3.1. *Extracting open-domain events*

We extract events from the tweets archived in TwiNL, a corpus of Dutch tweet IDs posted from December 2010 onward,[32] following the approach described in

Kunneman and van den Bosch,[2] which in turn is based on Ritter, Mausam, Etzioni and Clark.[33] In summary, explicit forward-referring time expressions in tweets are leveraged to identify event terms that have a strong connection to a specific future date. Kunneman and van den Bosch[2] applied the approach to a new unseen month of Dutch tweets and rank events by the goodness-of-fit between the date and event terms, scoring a precision-at-250 of 0.80. Apart from the tweets that mention the event, an extracted event is characterized by the date at which it takes place and by terms and hashtags that describe the event. Terms can consist of the names of the event, proper names of persons, locations, or organizations, and also actions and other aspects or properties of the event.

We extracted events from all tweets available in the TwiNL database. Our approach to event extraction assumes a set of tweets within the duration of a month. With a time span of over five years in the TWiNL archive, references to recurring events may overlap. To keep a constrained time window we employ a daily sliding window that spans a month of tweets from which events are extracted. After each new window, the top-2500 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. We applied this approach to TwiNL tweets between 2011/01/01 and 2015/10/31, resulting in a set of 97 885 events in total.

3.2. *Harvesting additional event tweets*

By definition, our approach to event extraction identifies tweets posted before event time and containing a forward-pointing time expression. 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, Iter, Naaman and Gravano[34] 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 characterized 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 shows that a bursty text phrase in Twitter is strongly related to the occurrence of an event.[35–37] 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. Based on this burstiness calculation, events that are mentioned only a few times on the event date might yield a high burstiness score when they are hardly mentioned before and after this date. To ensure a reliable burstiness calculation that highlights the more significant bursty events, 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.

3.3. *Selecting pre-event and post-event tweets*

Our approach to event extraction does not indicate the hour and minute at which an event starts; it only detects the event date. We therefore separated the event tweets into pre-event and post-event tweets by comparing the date of each tweet to the date of the event. We excluded tweets posted on the date of the event to ensure 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.

We limited the dataset to tweets that were posted within three days before and three days after the date of an event. The majority of messages about an event are posted during this time frame. 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 event with the highest score assigned by our event extraction system, indicating the certainty that the event terms denote an event of significance, would be chosen over the other overlapping events. The resulting number of events and tweets are presented in Table 1. In total, the dataset comprises of 3338 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 dataset we extracted a random sample of 100 events and for each event randomly selected ten pre-event and ten post-event tweets. One of

Table 1. General statistics of the collected tweets posted within three days before and three days after an event.

	#Events	#Tweets	Maximum	Median	Mean	St. Dev.
Pre-event	3338	3901 431	128 747	269	1169	4782
Post-event	3338	2925 069	63 669	216	876	3080
Total	3338	6826 500	192 416	553	2045	7159

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 dataset have a proper connection to the event for which they were queried.

4. Emotion Classification

In this section we describe the procedure to train and test models of positive expectation, disappointment and satisfaction, to be applied on our dataset of pre-event and post-event tweets. We will motivate the selection of hashtags to model the emotions of interest, and provide a thorough evaluation of the quality of these models. We define them in the following way:

- Positive expectation — Anticipatory excitement for a future event. The sender is personally involved with the event: s/he will attend or follow it. S/he does not only announce the event, but also expresses in any way that s/he has positive expectations.
- Disappointment — The sender looks back to an event in a negative way. S/he indicates, implicitly or explicitly, that things have not turned out as hoped or expected.
- Satisfaction — The sender looks back at an event, indicating that s/he enjoyed it.

4.1. *Training models of emotion*

In Section 2.2 we motivated our approach described in Kunneman, Liebrecht and van den Bosch[3] 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. Earlier studies show that multiple hashtags that convey the same meaning or emotion can be successfully combined to train a classifier to recognize this emotion.[38,17] The advantage

Table 2. Overview of the hashtag and number of tweets on which the initial emotion models were trained.

Emotion	Seed Hashtag	Gloss	#Tweets	#Random Tweets
Positive expectation	#zinin	#excited	606 310	606 310
Disappointment	#teleurgesteld	#disappointed	11 138	11 138
Satisfaction	#tevreden	#satisfied	17 459	17 459

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,[38] or an initial selection can be expanded based on an empirical procedure.[17] We follow the bootstrapping approach by Qadir and Riloff[17] in order to identify hashtags with a good link to the target emotions.

For each of five basic emotions, Qadir and Riloff[17] 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. After stripping away the hashtag itself from the feature space, they 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.

In contrast to Qadir and Riloff[17] we started with only one seed hashtag, the best fitting one, for each of the three target emotions: '#zinin' (#excited) for positive expectation, '#teleurgesteld' (#disappointed) for disappointment, and '#tevreden' (#satisfied) for satisfaction. 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.[39] 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 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[d] to tokenize 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. As text representation, 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

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

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.

Classification was performed by the Balanced Winnow algorithm.[40] 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 (event) emotion hashtags from, we randomly selected 20% of the events and their tweets from the event dataset described in Table 1 (also tweets posted more than three days before or after the events). This resulted in 3416 096 pre-event messages and 2713 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 and Riloff[17] 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 3.

Table 3. Overview of the hashtags that were added as training label after classifying a pool of event tweets.

Emotion	Hashtag	Gloss	#Tweets (Filtered)	Total
Positive expectation	'#zinin'	#excited	606 310	610 576
	'#klaarvoor'	#readyforit	996	
	'#heelveelzinin'	#veryexcited	3270	
Disappointment	'#teleurgesteld'	#disappointed	11 138	283 399
	'#zonde'	#pity	23 846	
	'#balen'	#bummer	96 651	
	'#spijtig'	#deplorable	4630	
	'#jammer'	#shame	142 385	
	'#teleurstelling'	#disappointment	4749	
Satisfaction	'#tevreden'	'#satisfied'	17 459	301 420
	'#blij'	#happy	141 276	
	'#dankbaar'	#grateful	15 835	
	'#genieten'	#enjoy	126 850	

Especially the training sets for Disappointment and Satisfaction were expanded considerably based on this procedure, 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.

4.2. *Emotion model evaluation*

We applied the same procedure as described in Section 4.1 for training the emotion classifiers on the hashtags listed in Table 3. We again balanced the hashtag-labeled tweets with an equal amount of random tweets as training data. In line with Kunneman, Liebrecht and van den Bosch,[3] we chose to apply the classifiers on a large sample of unlabeled tweets, in order to evaluate their quality in a real-world setting. We use the pool of tweets described in Section 4.1 as test set, with 3416 096 pre-event messages and 2713 961 post-event messages.

As it is unfeasible to evaluate performance by manually labeling the emotion in each of the many test tweets, we follow the evaluation procedure in Kunneman, Liebrecht and van den Bosch.[3] They assess classifier performance by means of (a) a hashtag-based evaluation and (b) an evaluation of the top ranked tweets.

4.2.1. *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 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, same metric as Recall), False Positive Rate (FPR), and Area Under the Curve (AUC).[41] The classifier that was trained to recognize positive expectation manages to retrieve

Table 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).

Emotion	Test tweets	With Target Hashtag	Classified	Correct	TPR	FPR	AUC
Positive expectation	3416 096	4999	836 289	4407	0.88	0.24	0.82
Disappointment	2713 961	1207	893 036	1008	0.84	0.33	0.75
Satisfaction	2713 961	1764	900 032	1501	0.85	0.33	0.76

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, 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.

4.2.2. *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. The two 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 for every annotator pair.[42] 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 5. For Positive expectation, 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 Disappointment 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 Satisfaction is lowest, with 0.40 based on a 66% threshold and 0.25 with a 100% threshold. The Cohen's Kappa agreement of tweets classified as Positive expectation is substantial, at 0.64, the agreement on Disappointment is moderate, at 0.49 and the agreement on Satisfaction is substantial, at 0.75. The mutual F-score is highest for Positive expectation (0.85), and slightly lower for Disappointment and Satisfaction (both 0.76).

Table 5. Precision and inter-annotator agreement on the presence of the target emotion in the top ranked 250 tweets by classifier confidence.

	Precision@250		Cohen's	Mutual
Emotion	66%	100%	Kappa	F-score
Positive expectation	0.66	0.49	0.64	0.85
Disappointment	0.51	0.31	0.49	0.76
Satisfaction	0.40	0.25	0.75	0.76

Table 6. Overview of the division of feature types in the 400 most indicative features per emotion, based on the respective models trained by Balanced Winnow. The examples are translated from Dutch.

Category	Subcategory	% in PE	Examples	% in D	Examples	% in S	Examples
Emotion	Feeling	2.8	#cantwait #curious #nervous	2.8	#sad #frustration #broken	8.0	#hyped #relieved #proud
	Feeling about	10.5	#fun due for it #exciting	26.3	letdown #toobad a waste	14.3	#inspiring #super #deserved
	Exclamation	2.5	#whoopwhoop #goforit #yes	3.8	ohno #ohwell by golly	16.3	yesyesyes yeaaah whaaaa
Target	Topic	47.3	#snowboarding #ll12 #carnaval	13.0	federer #olympics14 #rain	24.5	#beach lasagne no school
	Outcome	6.0	subscribed #newjob reserved	25.0	lost misjudged sickish	32.0	#graduated found #won
Temporal	Temporal	23.3	#almostweekend tomorrow	1.3	02-feb 04-jan	0.3	#sunday
Other	Other	6.5	#tweetup #hashtag	10.0	there_goes backwards	4.8	follows me got twitter

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. Despite their mixed performance, we expect the classifiers to be sufficiently sensitive to marked differences in emotions voiced about different events, and thereby to be suitable for the task of anticipointment detection.

4.2.3. *Analysis of models*

The Balanced Winnow algorithm returns per-class weights that we used to analyze the models of the classifiers. For the three models of emotion, we inspected the 400 features with the highest weight and divided them into four main categories:

- Emotion — subdivided into:
  - Feeling: A word or hashtag that refers to a feeling of the sender (could be preceded by 'I feel'; e.g. '#curious').
  - Feeling about: A word or hashtag that refers to a feeling of the sender about the topic ('#wasteoftime').
  - Exclamation: A word or phrase to utter a felt emotion ('yeaahh').
- Target — subdivided into:
  - Topic: The name of an event, entity, activity or object the emotion applies to.
  - Outcome: The outcome of an event or action. Might have an inherent positive ('graduated') or negative ('lost') connotation.
- Temporal — A reference to a point in time.
- Other — A word that does not fit into any of the previous categories.

In Table 6, we present the percentages of occurrences of the feature categories for the three models. The distributions of feature types in all three classifier models show some marked differences. The 400 top-ranked features in the Positive Expectation model are prominently categorized with topical and temporal words when compared to the other two models. Emotion in this model is mostly manifested as feelings about an upcoming event. This emotion, in the form of feelings about a past event, is also prominent in the Disappointment model, alongside features that describe an (undesirable) outcome of the event. The Satisfaction model, finally, is characterized by a relatively high percentage of emotional words (with exclamation as most frequent subcategory), while topical and outcome words are equally prominent.

5. Event Emotion

In this section we discuss the insights obtained after applying emotion classifiers on the set of event tweets described in Section 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.

5.1. *General patterns*

To explore the relation between emotions before and after the same event, we calculated the correlation between all three emotion pairs (positive expectation and disappointment, positive expectation and satisfaction, and disappointment and satisfaction) for each event. The correlation between the post-event emotions of disappointment and satisfaction, which can be expected to be low, are 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 cannot say anything about the intensity of emotion. Rather, the classifications of a collective of tweets reveal the commonality of an emotion. We base this commonality for positive expectation, disappointment or satisfaction on 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 Satisfaction represents a higher degree of public satisfaction with the event than a set of event tweets of which only 25% is classified as such.

For each event we calculated the percentage of tweets that were classified with any of the three emotions, after which we calculated the Pearson correlation coefficient 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.

In Figure 1 we display the correlation scores between all emotion pairs, based on the proportion of classifications. All scores were found to be significant ($p < 0.05$). The largest positive correlation, of around 0.60, exists between positive expectation and satisfaction. In contrast, positive expectation and disappointment are very weakly correlated, with values around 0.20. A weak positive correlation also holds between disappointment and satisfaction. Apparently, some events evoke strong emotions on both sides. A small effect of the minimum threshold of tweets per event is seen: the correlation is higher when only taking into account the more popular events with over 2000 tweets before and after event time.

In Figures 2 and 3 we display scatterplots of events by their proportion of classifications for Positive expectation versus Disappointment, and Positive expectation versus Satisfaction, 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 positive expectation, this is as likely followed by a large as by a small number of tweets classified as Disappointment. Some events are characterized by a high degree of anticipointment, with a large proportion of both positive expectation and disappointment. Conversely, some highly disappointing events did not follow high positive expectation, while some events with little disappointment

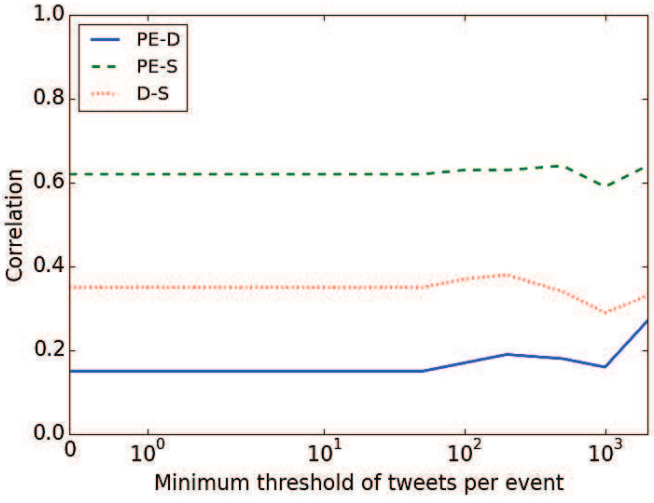


Fig. 1. Correlation between the emotion scores of an event when scoring emotion by the percentage of classifications. All reported correlations are significant (< 0.05); PE = Positive expectation, D = Disappointment, S = Satisfaction.

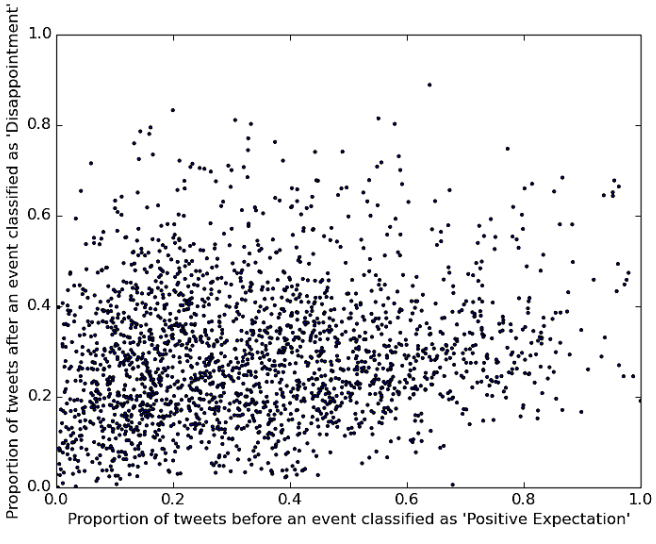


Fig. 2. Scatterplot of degree of positive expectation and disappointment by proportion of classifications for all events with over 100 tweets posted before and after event time.

did. The scatterplot of the percentage of Positive expectation and Satisfaction, in Figure 3, shows a more coherent relation between the two values in the point cloud. To some extent, the percentage of Satisfaction can be predicted from the percentage of Positive expectation.

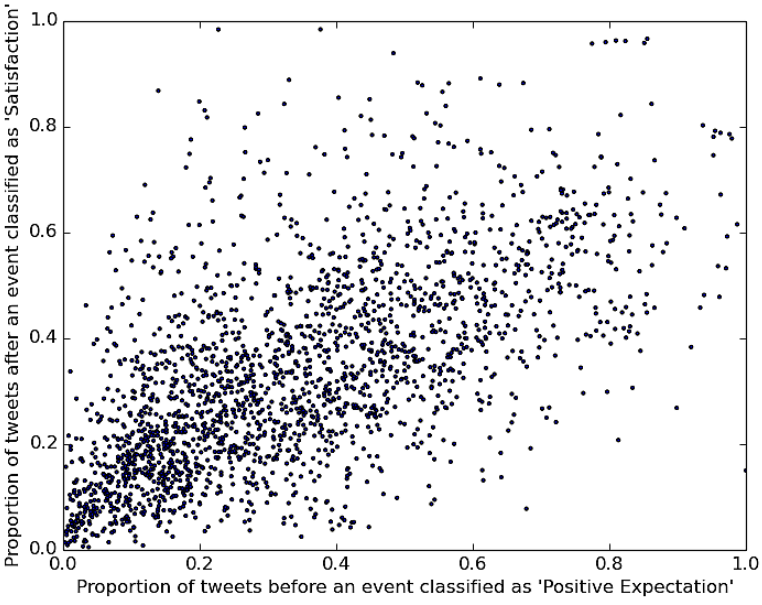


Fig. 3. Scatterplot of degree of Positive expectation and Satisfaction by proportion of classifications for all events with over 100 tweets posted before and after event time.

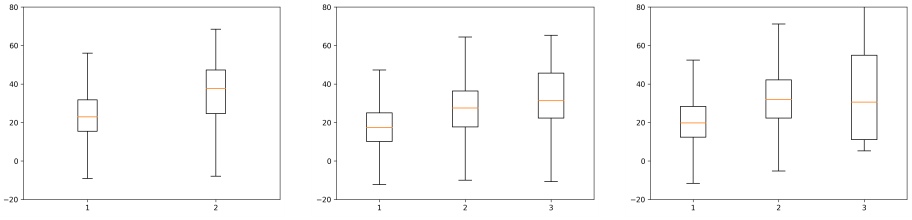
In sum, we find that positive expectation before an event does not make disappointment after the event more likely. Higher positive expectations are more often followed by satisfaction. In the following, we zoom in on individual events that are prototypical of anticipointment and positive expectation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

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 this section, we will highlight some of these patterns, and discuss the most exemplary events.

5.2.1. *Anticipointing events*

In order to score the degree of anticipointment for each event, we zoom in on the 90th percentile classifier score for positive expectations and disappointment. Tweets that are classified by the model at a higher score for an emotion tend to be more strongly linked to the emotion. The classifier score assigned by the Balanced Winnow algorithm to a tweet is inferred from its features as the sum of their positive and negative weights in the model. Hence, the score will be higher when more positively weighted emotion words are present. As a representation of this behavior, we matched the emotion words that we could identify in the top 400 features in each classifier model (see the analysis in Section 4.2.3) to the tweets that



(a) Positive Expectation (b) Disappointment (c) Satisfaction

Fig. 4. Box plots of classifier scores assigned to tweets by the number of emotion words occurring in the tweets (based on the top 400 features in the classifier models).

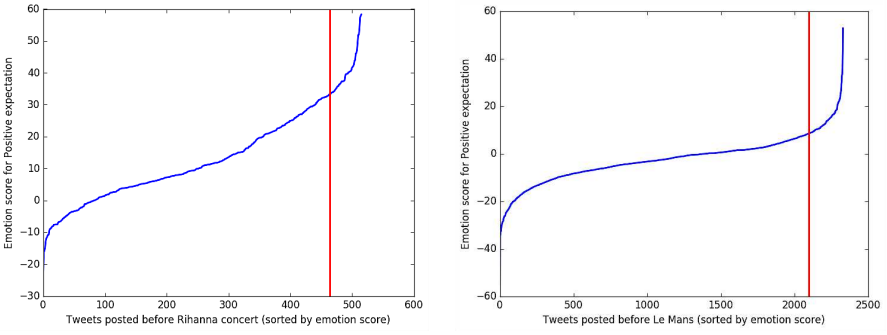


Fig. 5. Classifier score for Positive expectation of tweets posted before a Rihanna concert (left) and the 24-hours of Le Mans car race (right). The vertical red line indicates the 90th percentile.

were classified, relating the frequency of identified emotion words in the tweets to the classifier score assigned to them. In Figure 4 we display the box plots of these relations. When more of the known emotional features are present in the tweets, the average of emotion scores assigned to these tweets becomes significantly higher ($p < 0.01$, except for the difference between the Satisfaction model applied to tweets with two or three known emotional features). Thus, although the classifier score can only be interpreted as an approximation of emotion,[3] a higher density of emotional terms is generally reflected in a higher emotion score assigned by the classifier. Such information is neglected when we only focus on the binary classifications.

As an additional motivation to score anticipointment based on the 90th percentile classifier score of the tweets referring to an event, we noticed that the emotion scores assigned by the classifiers typically manifest extremes in both the higher scores and the lower scores for an event. Figure 5 presents examples of such curves, for tweets anticipating a concert by popstar Rihanna and tweets anticipating the annual 24-hour car race of Le Mans. The former tweets manifest a steeper curve on the right-hand side than on the left-hand side, while the curve of the latter tweets is more or less symmetric. The steeper curve arguably reflects a clearer excitement by those who look forward to the event, which would be ignored if we would summarize the emotion scores by means of average (respectively 13.4 and −2.4), median

Table 7. Top five anticipointing events.

Event	Date	Positive Expectation	Disappointment	Satisfaction	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

(10.7 and −1.9) or by focusing on the tweet with the highest score (58.3 and 52.7). Instead, we choose to describe the aggregate emotion by the 90th percentile: the confidence score of the tweet that is higher than 90% of the confidence scores of the other tweets (marked by the red line in Figure 5). This way, we take into account the positive classifications in the right-hand region of the curve, without a bias towards the extremes. In the context of our examples, the anticipation towards the events is scored as 33.3 for the Rihanna concert and 8.4 for the car race, which is a better reflection of their aggregate emotions.

Based on the 90th percentile classifier confidence for positive expectations and disappointment, we can quantify the anticipointment of each event and rank them accordingly. We use the formula as stated in Eq. (1) to calculate this score (PE = 90th percentile positive expectation, D = 90th percentile disappointment, S = 90th percentile satisfaction):

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

To avoid a strong influence of either the score for positive expectation or disappointment, we calculated the harmonic mean between them. The score for satisfaction is subtracted, in order to make sure that the highest scoring events are mostly disappointing and not also satisfactory.

The five most anticipointing events are shown in Table 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 rather 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 (Ajax versus Twente) and #feytwe (Feyenoord versus Twente) were matches where the team that is favorited 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.

Table 8. Top five events by degree of anticipated satisfaction.

Event	Date	Positive Expectation	Disappointment	Satisfaction	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

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 favorable 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 favorite team or a canceled event.

5.2.2. *Events with positive expectation followed by satisfaction*

We calculated the degree of positive expectation followed by satisfaction in the same way as anticipointment (see Eq. (2)), as the harmonic mean between the 90th percentile classifier confidence score of positive expectation and satisfaction, subtracted by the score for disappointment (PE = 90th percentile positive expectation, D = 90th percentile disappointment, S = 90th percentile satisfaction):

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

The top five of anticipated satisfaction is given in Table 8. In contrast to most of the anticipointing events, none of these events are characterized 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 characterized by a promotional element. Most tweets that expressed satisfaction after the Exact Live event were posted by organizations that had a booth during the event, and probably wanted to positively emphasize 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.

Table 9. Top five events on scoring high on all three emotions.

Event	Date	Positive Expectation	Disappointment	Satisfaction	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

5.2.3. *Events with overall high emotion*

The weak correlation that was found between disappointment and satisfaction (Figure 1) 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 (see Eq. (3); PE = 90th percentile positive expectation, D = 90th percentile disappointment, S = 90th percentile satisfaction):

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

The five events with the highest overall emotion are displayed in Table 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 on all three emotions.

5.3. *Case study*

While Sections 5.1 and 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 Positive expectation, Disappointment and Satisfaction, are shown in Figure 6. The Netherlands played three first round matches, against Spain, Australia, and Chile, followed by an eighth final match against Mexico, a quarter-final

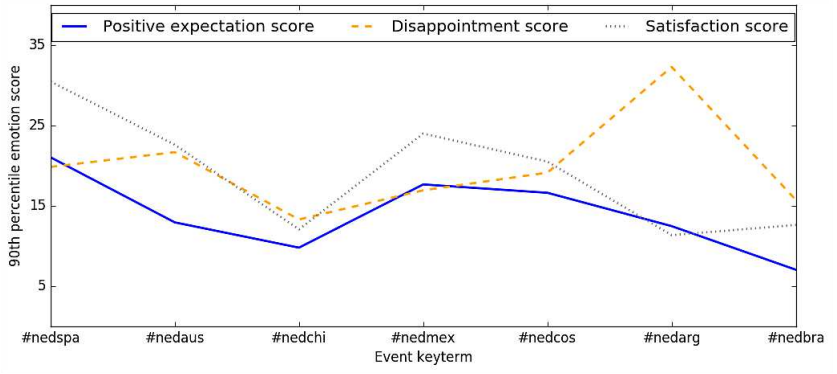


Fig. 6. Overview of 90th percentile emotion scores for all matches of the Dutch football team during the 2014 World Cup, in the order in which they were played; PE = Positive expectation, D = Disappointment, S = Satisfaction.

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 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

Table 10. Percentages of fully related tweets posted before and after the events in the case study, based on annotated samples of 100 tweets.

Event	Pre-Event Related Tweets	Post-Event Related Tweets
#nedspa	71%	60%
#nedaus	25%	47%
#nedchi	20%	18%
#nedmex	37%	47%
#nedcos	35%	48%
#nedarg	43%	70%
#nedbra	48%	14%

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.

To quantify the influence of tweets relating to side-events rather than the main event of analysis on the 90th percentile emotion scores for an event, we set out to highlight the fully related tweets. We drew a random sample of 100 pre-event tweets and 100 post-event tweets for each of the seven events, after which the first author annotated whether they referred to the football match. Fully related tweets shared expectations on or experiences with the match itself, while tweets were filtered as not fully related, for example, when they reported on the number of spectators, invited followers to participate in a prediction pool, or were mostly discussing another match. The numbers of fully related tweets per event are presented in Table 10. When strictly focusing on tweets in which the authors shared their thoughts on the football match, commonly more than half or even three quarters of the event tweets appeared unrelated. Only for the first match between The Netherlands and Spain a substantial number of tweets were fully related, as was the case for the tweets that referred to the semi-final match after it was won by Argentine.

The emotion plot based on the fully related tweets is presented in Figure 7. The lines for the three emotions largely follow the same pattern as before filtering, except for disappointment. After the second first-round match against Australia, where the performance on the pitch was at odds with the impressive first match, the 90th percentile disappointment score now decreases as the tournament develops. The original plot, in contrast, shows an increase in disappointment from the third match onward. The differences in satisfaction between the matches are less significant after filtering, hovering around a value of 25 before the lost semi-final match. Finally, the positive expectations for the first two matches in the knock-out phase appear higher after filtering, marking a more evident distinction with the disappointment scores.

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

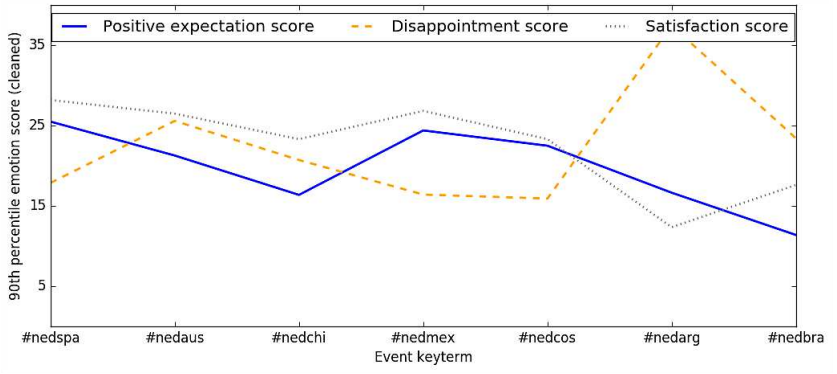


Fig. 7. Overview of 90th percentile emotion scores for all matches of the Dutch football team during the 2014 World Cup, after selection of directly related tweets; PE = Positive expectation, D = Disappointment, S = Satisfaction.

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. While such an overview can largely be yielded by an unfiltered set of event tweets that might partly relate to side-events, a more accurate overview of these patterns could be obtained when only focusing on tweets that fully relate to the event of analysis.

6. Conclusion and Discussion

We applied classifiers trained to recognize positive expectation, disappointment and satisfaction on a dataset that comprised over 3000 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 expectations followed from the insight that the felt disappointment is most likely when expectations are high.[1] 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.[43] After a disappointing experience, the discrepancy with the preliminary high expectations might be made consistent by underlining the good parts or trivializing 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 dataset the strongest anticipointment can be observed after a clear negative outcome, such as a defeat of the favorable 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 not as widespread. The correlation between positive expectation and satisfaction can also be explained by the Pollyanna hypothesis,[44] which states that there is a universal human tendency to use evaluatively positive words more frequently and diversely than evaluatively negative words in communicating. Following this hypothesis, a succession of positive expectation by satisfaction is by default more likely.

Our study is the first to explore the analysis of collective emotions in time on Twitter. Although we show that the task is feasible, much ground can be gained by improving on data collection and emotion classification. A limitation of the component that automatically extracts events from Twitter is that it only retrieves the date of an event, without knowledge of the exact start time. This prevented us from detecting emotion from tweets right before or after event time on the day of the event. In addition, events that span multiple days are now included as separate events per date, which led part of the tweets posted during the event to be included as pre-event and post-event tweets. In a future study, we will examine whether identical event terms on consecutive days reflect multi-day events rather than separate events, so they can be safely combined into a single 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.

Although the emotion 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. A possible reason for this is that we contrasted hashtag-labeled emotion tweets with random tweets during training. As a result, the classifier model will give considerable weight to the temporal aspect of these emotions: tweets with a hashtag that denotes excitement about the future (like '#lookingforwardtoit') will be much more likely to include markers that point to the future than random tweets. As a result, a lot of positive classifications were made when we applied the classifiers on tweets that refer to an event, which also tend to include time expressions. 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.

In future work we aim to solve the limitations described above. Another interesting avenue for further research is the addition of a component to distinguish event types. This can help to obtain insights into the influence of event type on anticipointment or satisfaction after positive expectation.

Acknowledgments

This research was supported by the Dutch national program COMMIT/ as part of the Infiniti project. We thank Erik Tjong Kim Sang for the development and

support of the http://twiqs.nl service, and thank Eric Sanders for his assistance as third annotator.

References

1. M. Miceli and C. Castelfranchi, *Expectancy and Emotion* (Oxford University Press, Oxford, UK, 2014).
2. F. Kunneman and A. van den Bosch, Open-domain extraction of future events from twitter, *Natural Language Engineering* (2016).
3. F. Kunneman, C. Liebrecht and A. van den Bosch, The (un)predictability of emotional hashtags in twitter, in *Proc. of the 5th Workshop on Language Analysis for Social Media* (*LASM*) (ACL, Gothenburg, Sweden, 2014), pp. 26–34.
4. W. W. van Dijk, M. Zeelenberg and J. van der Pligt, Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment, *Journal of Economic Psychology* **24**(4) (2003) 505–516.
5. L. van Boven and L. Ashworth, Looking forward, looking back: Anticipation is more evocative than retrospection, *Journal of Experimental Psychology*: *General* **136**(2) (2007) 289–300.
6. M. Thelwall and A. Kappas, The role of sentiment in the social web, in *Collective Emotions: Perspectives from Psychology, Philosophy, and Sociology*, eds. C. Von Scheve and M. Salmela (Oxford University Press, UK, 2014), pp. 375–388.
7. I. van der Löwe and B. Parkinson, Relational emotions and social networks, in *Collective Emotions*: *Perspectives from Psychology, Philosophy, and Sociology*, eds. C. Von Scheve and M. Salmela (Oxford University Press, UK, 2014), pp. 125–140.
8. A. Go, R. Bhayani and L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford (2009), pp. 1–12.
9. A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in *Proc. of the Seventh Int. Conf. on Language Resources and Evaluation* (*LREC'10*), eds. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, European Language Resources Association (ELRA), Valletta, Malta (2010).
10. D. Davidov, O. Tsur and A. Rappoport, Enhanced sentiment learning using Twitter hashtags and smileys, in *Proc. of the 23rd Int. Conf. on Computational Linguistics*: *Posters*, ed. C.-R. Huang (Tsinghua University Press, Beijing, China, 2010), pp. 241–249.
11. A. Montoyo, P. Martínez-Barco and A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments, *Decision Support Systems* **53**(4) (2012) 675–679.
12. P. Ekman, Universals and cultural differences in facial expressions of emotion, in *Nebraska Symp. on Motivation*, ed. J. K. Cole (University of Nebraska Press, Omaha, NE, USA, 1971).
13. M. Purver and S. Battersby, Experimenting with distant supervision for emotion classification, in *Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics* (ACL, Stroudsburg, PA, USA, 2012), pp. 482–491.
14. K. Roberts, M. A. Roach, J. Johnson, J. Guthrie and S. M. Harabagiu, EmpaTweet: Annotating and detecting emotions on twitter, in *Proc. of the Seventh Int. Conf. on Language Resources and Evaluation* (*LREC'12*), eds. N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (European Language Resources Association (ELRA), Istanbul, Turkey, 2012), pp. 3806–3813.

15. R. C. Balabantaray, M. Mohammad and N. Sharma, Multi-class Twitter emotion classification: A new approach, *International Journal of Applied Information Systems* **4**(1) (2012) 48–53.
16. S. M. Mohammad, #emotional tweets, in *Proc. of the First Joint Conf. on Lexical and Computational Semantics*, Vol. 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (ACL, Stroudsburg, PA, USA, 2012), pp. 246–255.
17. A. Qadir and E. Riloff, Bootstrapped learning of emotion hashtags #hashtags4you, in *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (ACL, Stroudsburg, PA, USA, 2013), pp. 2–11.
18. R. Plutchik, *Emotion: A psychoevolutionary synthesis* (Harper & Row, New York, NY, USA, 1980).
19. J. Suttles and N. Ide, Distant supervision for emotion classification with discrete binary values, in *Computational Linguistics and Intelligent Text Processing*, ed. A. Gelbukh (Springer-Verlag, Berlin, Germany, 2013), pp. 121–136.
20. S. M. Mohammad and S. Kiritchenko, Using hashtags to capture fine emotion categories from tweets, *Computational Intelligence* **31**(2) (2015) 301–326.
21. J. S. Y. Liew, Discovering emotions in the wild: An inductive method to identify fine-grained emotion categories in tweets, in *Proc. of the Twenty-Eighth Int. Florida Artificial Intelligence Research Society Conf.*, eds. I. Russell and W. Eberle (The AAAI Press, Menlo Park, CA, USA, 2015), pp. 317–323.
22. R. Snow, D. Jurafsky and A. Y. Ng, Learning syntactic patterns for automatic hypernym discovery, in *Advances in Neural Information Processing Systems 17* (*NIPS 2004*), eds. L. Saul, Y. Weiss and L. Bottou (MIT Press, Cambridge, MA, USA, 2005), pp. 1297–1304.
23. A. Fraisse and P. Paroubek, Twitter as a comparable corpus to build multilingual affective lexicons, in *The 7th Workshop on Building and Using Comparable Corpora*, eds. P. Zweigenbaum, S. Sharoff, R. Rapp, A. Aker and S. Vogel (2014), pp. 26–31.
24. J. Bollen, H. Mao and A. Pepe, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in *Proc. of the Fifth Int. Conf. on Weblogs and Social Media* (The AAAI Press, Menlo Park, CA, USA, 2011), pp. 450–453.
25. M. Larsen, T. Boonstra, P. Batterham, B. O'Dea, C. Paris and H. Christensen, We feel: Mapping emotion on Twitter, *IEEE Journal of Biomedical and Health Informatics* **19**(4) (2015) 1246–1252.
26. V. Sintsova, C.-C. Musat and P. Pu Faltings, Fine-grained emotion recognition in olympic tweets based on human computation, in *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (ACL, Stroudsburg, PA, USA, 2013), pp. 12–20.
27. M. K. Torkildson, K. Starbird and C. Aragon, Analysis and visualization of sentiment and emotion on crisis tweets, in *Int. Conf. on Cooperative Design, Visualization and Engineering* (Springer International Publishing, 2014), pp. 64–67.
28. M. D. Sykora, T. Jackson, A. O'Brien, S. Elayan and A. Von Lunen, Twitter based analysis of public, fine-grained emotional reactions to significant events, in *The Proc. of the European Conf. on Social Media* (University of Brighton, Brighton, UK, 2014), pp. 540–548.
29. M. Brooks, J. J. Robinson, M. K. Torkildson and C. R. Aragon, Collaborative visual analysis of sentiment in Twitter events, in *Proc. of the 11th Int. Conf. on Cooperative Design, Visualization, and Engineering*, ed. Y. Luo (Springer-Verlag, Berlin, Germany, 2014), pp. 1–8.

30. M. Thelwall, K. Buckley and G. Paltoglou, Sentiment in Twitter events, *Journal of the American Society for Information Science and Technology* **62**(2) (2011) 406–418.
31. Y.-S. Chen, C. Argueta and C.-H. Chang, EmoTrend: Emotion trends for events, in *Database Systems for Advanced Applications*, eds. M. Renz, C. Shahabi, X. Zhou and M. A. Cheema (Springer, Heidelberg, Germany, 2015), pp. 522–525.
32. E. Tjong Kim Sang and A. van den Bosch, Dealing with big data: The case of Twitter, *Computational Linguistics in the Netherlands Journal* **3** (2013) 121–134.
33. A. Ritter, Mausam, O. Etzioni and S. Clark, Open domain event extraction from Twitter, in *Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (*KDD '12*) (ACM, New York, NY, USA, 2012), pp. 1104–1112.
34. H. Becker, D. Iter, M. Naaman and L. Gravano, Identifying content for planned events across social media sites, in *Proc. of the Fifth ACM Int. Conf. on Web Search and Data Mining* (*WSDM '12*) (ACM, New York, NY, USA, 2012), pp. 533–542.
35. J. Weng and B.-S. Lee, Event detection in Twitter, in *Proc. of the Fifth Int. Conf. on Weblogs and Social Media* (The AAAI Press, Menlo Park, CA, USA, 2011), pp. 401–408.
36. C. Li, A. Sun and A. Datta, Twevent: Segment-based event detection from Tweets, in *Proc. of the 21st ACM Int. Conf. on Information and Knowledge Management* (ACM, New York, NY, USA, 2012), pp. 155–164.
37. Y. Qin, Y. Zhang, M. Zhang and D. Zheng, Feature-rich segment-based news event detection on Twitter, in *Proc. of the Sixth Int. Joint Conf. on Natural Language Processing* (AFNLP, 2013), pp. 302–310.
38. F. Kunneman, C. Liebrecht, M. van Mulken and A. van den Bosch, Signaling sarcasm: From hyperbole to hashtag, *Information Processing & Management* **51**(4) (2015).
39. R. González-Ibáñez, S. Muresan and N. Wacholder, Identifying sarcasm in Twitter: A closer look, in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2011), pp. 581–586.
40. N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, *Machine Learning* **2** (1988) 285–318.
41. T. Fawcett, ROC graphs: Notes and practical considerations for researchers, Tech. Rep. HPL-2003-4, HP Laboratories (Palo Alto, CA, USA, 2004).
42. J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1) (1960) 37–46.
43. L. Festinger, *A Theory of Cognitive Dissonance* (Stanford University Press, Redwood City, CA, USA, 1962).
44. J. Boucher and C. E. Osgood, The pollyanna hypothesis, *Journal of Verbal Learning and Verbal Behavior* **8**(1) (1969) 1–8.