Natural Language Engineering **22** (5): 655–686. © Cambridge University Press 2016
doi:10.1017/S1351324916000036

Open-domain extraction of future events from Twitter[†]

FLORIAN KUNNEMAN and
ANTAL VAN DEN BOSCH
Centre for Language Studies, Radboud University,
P.O. Box 9103, NL-6500, HD Nijmegen, the Netherlands
e-mails: a.vandenbosch@let.ru.nl, f.kunneman@let.ru.nl

(*Received 20 February 2015; revised 8 January 2016; accepted 8 January 2016;*
first published online 22 February 2016)

Abstract

Explicit references on Twitter to future events can be leveraged to feed a fully automatic monitoring system of real-world events. We describe a system that extracts open-domain future events from the Twitter stream. It detects future time expressions and entity mentions in tweets, clusters tweets together that overlap in these mentions above certain thresholds, and summarizes these clusters into event descriptions that can be presented to users of the system. Terms for the event description are selected in an unsupervised fashion.[1] We evaluated the system on a month of Dutch tweets, by showing the top-250 ranked events found in this month to human annotators. Eighty per cent of the candidate events were indeed assessed as being an event by at least three out of four human annotators, while all four annotators regarded sixty-three per cent as a real event. An added component to complement event descriptions with additional terms was not assessed better than the original system, due to the occasional addition of redundant terms. Comparing the found events to gold-standard events from maintained calendars on the Web mentioned in at least five tweets, the system yields a recall-at-250 of 0.20 and a recall based on all retrieved events of 0.40.

1 Introduction

A significant part of the messages posted on the social media platform of Twitter relate to future events. A system that can extract this information from Twitter and present an overview of upcoming popular events, such as sports matches, national holidays, and public demonstrations, is of potentially high value. This functionality may not only be relevant for people interested in attending an event or learning about an event; it may also be relevant in situations requiring decision support to activate others to handle upcoming events, possibly with a commercial, safety, or security goal. As an example of the latter category, *Project X Haren*,[2] a violent mass

[†] This research was funded by the Dutch national program COMMIT. We thank Erik Tjong Kim Sang for the development and support of the http://twiqs.nl service.

[1] A demo of this system is available at http://applejack.science.ru.nl/lamaevents/

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

riot on September 21, 2012, in Haren, the Netherlands, organized through social media, was abundantly announced on social media, with specific mentions of the date and place. A national advisory committee, installed after the event, was asked to make recommendations to handle similar future events. The committee stressed that decision-support alerting systems on social media need to be developed, 'where the focus should be on the detection of collective patterns that are remarkable and may require action' (Cohen *et al.* 2013, p. 31, our translation).

We describe a system that provides a real-time overview of open-domain future events of potential interest to any audience, by leveraging explicit references to the start time of upcoming events. Hence, the type of event that we focus on is an event that one can anticipate on, but that is typically not of a personal nature. When we refer to the word 'event' henceforth, we follow the definition of McMinn, Moshfeghi and Jose (2013, p. 411): 'An event is a significant thing that happens at some specific time and place', where 'significant' is defined as '[something that] may be discussed in the media'.

Many Twitter users choose and like to share their anticipations, as can be inferred from the frequent occurrence of the Dutch hashtag '#zinin' (#lookingforwardtoit) or of the term 'vanavond' (tonight). Queries for tweets with these terms yield, respectively, 677,156 tweets in 2011 and 2012 (Kunneman, Liebrecht and van den Bosch 2014) and about seven million tweets between August 2010 and Spring 2012 (Weerkamp and de Rijke 2012). Given an estimated average amount of four million Dutch tweets per day, the two terms comprise about 0.02 per cent and 0.29 per cent of all Dutch tweets in their respective periods. Thus, a system based on anticipatory references to future events starts with a wide selection. However, among this load of future event references a lot of events occur that are not of public importance, such as a person's holiday break or a family visit.

The challenge is then to distinguish events of public interest from personal events. We adopt the approach by Ritter *et al.* (2012), who look for the co-occurrence between the key descriptive entities of an event and an explicitly mentioned date of the event. Often, public events are referred to by different persons in combination with the same time reference. Ritter *et al.* (2012) show that ranking events based on this evidence indeed results in a large majority of socially anticipated events higher up the rankings. The tacit assumption here is that event significance is at least partly based on the number of people that post about the event. This assumption rules out the possibility of detecting significant events about which only few people post on Twitter.

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

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

- We extend the approach with a clustering stage to decrease duplicate output and a procedure to rank tweets that describe an event best by their informativeness;
- We conduct an extensive evaluation of the system, by presenting its output to a pool of human annotators who are unbiased towards the system. We also compare the system output to gold standard events from curated calendars on the Web, to assess the system's recall. Some components are evaluated in isolation.

2 Related work

Our system is an adaptation of the system proposed by Ritter *et al.* (2012), referred to by them as TWICAL. Explicit displays of knowledge of events in tweets are detected by scanning for the joint and frequent occurrence of a reference to a point in time, a so-called event phrase, and a named entity. The number of tweets in which an entity is mentioned with the same date is used as a signal to extract significant events as opposed to mundane or personal events. Events are ranked by the fit between the date and entity, leading to a precision at the ranked top-hundred events of ninety per cent and a precision at 500 of sixty-six per cent. An advantage of TWICAL is that it does not pose any restrictions on the type of event that is extracted, making it an open-domain approach: any event that people refer to with a future date can be found.

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

2.1 Event extraction

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

Both TWICAL and the approach of Weerkamp and de Rijke (2012) rely on the explicit mentioning of the time of future events. This general clue leads to the extraction of open-domain events of unknown types. Approaches that aim to find events of a known type focus on other clues in the short messages on Twitter,

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

such as marker words or hashtags. Sakaki, Okazaki and Matsuo (2010) aim to find earthquakes by harvesting tweets that mention (a variant of) the word ‘earthquake’ and relate the location at which they were posted to geological faults, enabling them to forecast the progression of the earthquake along the faults. Benson, Haghighi and Barzilay (2011) focus on the extraction of music events in the region of New York City, and scan tweets for mentions of artists and venues.

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

2.2 Event detection

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

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

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

A valuable clue for event detection other than bursty topics or terms is the influence of real-world events on emotions. Indirectly, emotion bursts (mood swings)

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

could indicate events. Ou *et al.* (2014) monitor emotion throughout Twitter communities, and look for bursty emotion states. Valkanas and Gunopulos (2013) aggregate emotions in tweets by location.

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

3 System outline

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

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

We describe and motivate the different components below. A separate evaluation of the most important components is presented in Section 6.2.

3.1 Tweet processing

The setting of our experiment is the Dutch Twitterverse. Taking a relatively lesser used language is illustrative of a situation in which we cannot rely on standard tools developed for English (Ritter *et al.* 2012). We used TwiNL, a database of Dutch tweet IDs harvested from December 2010 onward (Tjong Kim Sang and van den Bosch 2013), to simulate operating on the live Twitter stream (see Section 4.1 for more details). All tweets are tokenized[3] and turned to lower case. Each tweet is then scanned for time expressions. Tweets that contain a time expression are fed

[3] Ucto was applied for tokenization: `http://ilk.uvt.nl/ucto/`

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

to the second component in this stage, concept extraction. All other tweets are discarded.

3.1.1 Extraction of time expressions

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

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

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

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

The 'Weekday' rules match a mention of a weekday, optionally preceded by the phrase 'volgende week' (next week) or followed by 'ochtend' (morning), 'middag' (afternoon), 'avond' (evening), or 'nacht' (night). The weekday is translated into a date by computing the number of days to the forthcoming occurrence of the weekday after the time of the tweet post, and adding seven days if the weekday is preceded by 'volgende week'. To exclude tweets that refer to the previous occurrence of the weekday, and thus to a past event, we scanned the tweets that match a weekday for verbs in the past tense by applying automatic part-of-speech tagging with Frog (Van den Bosch *et al.* 2007), a Dutch morpho-syntactic tagger and parser. Tweets containing a verb in the simple past or past perfect were discarded.

Our system gives preference to the most specific TIMEX if more than one TIMEX is seen in a tweet. A TIMEX matching an exact rule is preferred over a TIMEX matching a weekday, and an exact rule matching TIMEX is overruled

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

by a TIMEX matching a specific date. If a tweet contains more than one future time reference from the same rule type, both future dates are related to the tweet.

3.1.2 Extraction of concepts

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

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

Rather than developing a Part-of-Speech tagger for Dutch tweets ourselves, we chose to apply the *commonness* metric, as formulated by Meij, Weerkamp and de Rijke (2012). They match the word *n*-grams in a tweet with equally named Wikipedia articles, and assign a score to such *n*-grams based on their commonness in Wikipedia. By leveraging the crowd-sourced platform of Wikipedia, on which many entities are described and added, we expected to extract a wide range of event types. We compared this approach to the performance of an off-the-shelf system for Named Entity Detection in Dutch, and found that the former yields a significantly better performance. See Section 6.2.2. for a description of this experiment.

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

$$Commonness(c,q) = \frac{|L_{q,c}|}{\sum_{c'} |L_{q,c'}|}, \qquad (1)$$

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

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

We downloaded the Dutch Wikipedia dump of November 14, 2013 from http://dumps.wikimedia.org/nlwiki/nlwiki-20131114-pages-articles.xml.bz2, and parsed it with the Annotated-WikiExtractor[4]. Then, we used Colibri Core[5] to calculate the commonness of any concept that has its own Wikipedia article, and is used as an anchor text on other Wikipedia pages at least once. These statistics are used to extract concepts from a tweet. Tweets that matched a future time reference in the first stage are stripped of this time reference, and *n*-grams with *n* up to five are extracted. Any *n*-gram that is found to have a commonness score which is above 0.05 (for any of the *n*-grams possible anchored concepts) is extracted as a concept.

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

3.2 Event extraction

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

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

3.2.1 Event ranking

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

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

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

[4] https://github.com/jodaiber/Annotated-WikiExtractor

[5] http://proycon.github.io/colibri-core/

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

means of the G_2 log likelihood ratio statistic:

$$G_2 = \sum_{z \in \{e, \neg e\}, y \in \{d, \neg d\}} O_{z,y} \times \ln \left(\frac{O_{z,y}}{E_{z,y}} \right). \tag{2}$$

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

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

$$G_2u = \left(\frac{u}{t} \right) * G_2. \tag{3}$$

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

The calculation of G_2u for each pair results in a ranked list of date–term pairs. To reduce subsequent computational costs, we discarded all pairs with a rank number below 2,500. In other words, at any point, we are computing the top-2,500 most significant date–term pairs.

3.2.2 Event clustering

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

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

As a preparation for clustering, each set of tweets in which the same date–term pair occurs is aggregated into one big document. Subsequently, the documents are

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

converted into a feature vector with $tf * idf$ weighting (Day and Edelsbrunner 1984). The idf value is based on all aggregated documents in the set of 2,500 date–term pairs.

A useful constraint for date–term pairs to be clustered is that the dates of the two pairs be equal. Instead of generating a similarity matrix of all 2,500 date–term pairs, a similarity matrix is generated for each set of date–term pairs labeled with the same date. The cosine similarity is calculated between each date–term pair in such a set, based on their feature vector, and the similarity pairs are ranked from most similar to least similar. Each date–term pair forms an initial cluster with only one event term. Starting from the two clusters that are most similar, they are merged if their similarity is above threshold x. This process was repeated until the highest ranked similarity was below x. We chose to apply single-link clustering rather than calculating a centroid after each merge, so as to reduce computational costs. Hence, only the initial similarity table is used, and one combination of event terms with an above-threshold similarity suffices to merge for example two large clusters.

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

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

The threshold x was empirically set to 0.7, by testing on the first two days in the tweet set described in Section 4.1. Arguably, event clusters that comprise multiple actual events are more unwanted as output than duplicate events. We therefore preferred a precision-oriented clustering, with a minimum amount of false positives. An evaluation of clustering performance is presented in Section 6.2.3.

3.2.3 Event filtering

Although we try to discard references to a past weekday falsely identified as a coming weekday, by scanning the tweets for verbs in the past tense, some references might still surpass this filter. For example, the (translated) tweet ‘State police takes over Ferguson safety – Thursday, Missouri’s state police has . . . http://t.co/zdujcsylnz’ clearly refers to a past event, while it does not contain a verb in the past tense. As such news reports are often repetitively forwarded (retweeted) in unaltered form, we add another filter by discarding any event with a type–token ratio below 0.4.

The type–token ratio is calculated from the tweets of an event by dividing the number of different words in the tweets by the total number of word tokens. A low type–token ratio indicates repetition; a high type–token ratio indicates a high variance of words, and may represent an event that is referred to from different angles. With a threshold of 0.4, we aim to filter the events that are described with the most repetitive tweets, while minimizing the chance to discard any event with a more diverse vocabulary.

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

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

- repeated trouble after Thursday Meppel day ' #police arrests couple after violence http://t.co/eubrb72zvh
- repeated trouble after Thursday Meppel day ' #police arrests couple after violence http://t.co/hfxiazav6u
- repeated trouble after Thursday Meppel day ' #police arrests couple after violence http://t.co/wmsvuttlf5

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

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

3.3 Event presentation

3.3.1 Resolving overlap of concepts

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

First, we rank the event terms by their commonness score. A list of 'clean' event terms is initiated, which at first consists only of the event term with the highest rank. Starting from the second-ranked event term, the term is compared to the list of clean event terms and added to this new list when there is no overlap with any of the terms in this list. An overlap occurs if two event terms have overlapping word tokens. Thus, event terms are only added if they contain completely new information. Hashtags are seen as a unigram for this comparison, and are stripped from their hashtag symbol (#). This way, a redundant presentation that would concatenate for example 'pukkelpop' and '#pukkelpop' or 'mario kart' and '#mario', is avoided. As hashtags are not linked to a commonness score, they are at the bottom of the list, so that only non-overlapping hashtags are added to the resulting list.

3.3.2 Enriching the event description

Terms that represent an event should ideally provide a sufficient summarization of the event, similar to a news headline. We added a method to our framework to enrich the existing event terms with additional terms. The method is unsupervised

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

and bases the addition of terms on the set of tweets that announce the event. The procedure is described below:

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

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

3.3.3 Ordering of event terms

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

3.3.4 Ranking of tweets

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

- The tweets that describe an event are sorted by their importance to the event. The importance of a tweet is scored by the summed $tf * idf$ values of the words in the tweet. These values are in line with the ones that were generated in the term addition procedure (Section 3.3.2). The intuition is that words with a high $tf * idf$ are more specific than words with a low $tf * idf$, and are likely to describe key aspects of the event. By summing up the $tf * idf$ values of a tweet, its descriptiveness can be scored heuristically.
- Any tweet that has a word overlap of eighty per cent or higher with one of the tweets at a higher rank is transferred to the bottom of the list. This procedure runs until every tweet has been seen. The result is a re-ordered list of tweets.

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

4 Experimental set-up

4.1 Data

We collected a large sample of Dutch tweets posted in August 2014 to evaluate our system. As mentioned earlier, we used TwiNL, a database of Dutch tweet IDs from December 2010 onward (Tjong Kim Sang and van den Bosch 2013), to collect the tweets. The sample of tweets in August totals 27,681,567 individual posts[6].

4.2 Precision evaluation

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

Commonness+ comprises the full system as described in Section 3. The Commonness system does not include the addition of event terms (described in Section 3.3.2). This is to evaluate the value of this component to the description of events. The Ngram system acts as a baseline. It has a different approach to concept extraction from tweets: rather than basing the extraction on an above-threshold commonness score, any *n*-gram with $n \leq 5$ qualifies as a concept. Accordingly, any *n*-gram that has a good fit with a date might be clustered with *n*-grams with similar tweets to form an event.

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

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

For the layout and distribution of the evaluation, we made use of survey tool Qualtrics[7]. Each annotator was sent a unique survey with fifty specifically assigned events. For each event, the annotator was presented with the five top-ranked tweets and was asked whether these tweets refer to the same event. At the start of the survey, the annotator was given a definition of what is an event: 'An event takes place at a specific point in time and has value for a larger group of people'. The annotator was also told that the five tweets might refer to different sub-events that relate to one overarching theme. In these cases, they should assess the overarching

[6] The tweet IDs can be found at http://www.ru.nl/lst/resources/

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

theme as event or no event. If the theme is an event itself, such as a football match, the tweets can be assessed as referring to the same event, whereas the name of a city (occurring in tweets referring to different events in that city) as overarching theme does not qualify as an event. The complete instructions that were shown to the annotators (translated from Dutch) is included in Appendix B.

Whenever the annotator assessed the output as an event, he was subsequently presented with the event terms. The task then was to assess, on a scale from one to three (corresponding to poor, moderate, and good), how well the terms relate to the event that was identified. If an annotator did not identify an event in the five tweets, he would directly move on to the next output.

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

4.3 Recall evaluation

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

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

- Matches in the top-level Dutch national football league, the *Eredivisie*. We extracted an overview of the matches in the 2014–2015 season, starting August 8, from `sport.infonu.nl`,[8] and selected all matches in August and September.
- Public events; local or national events that take place at a single location, such as expositions, carnivals, and parties. We scraped the overviews of August and September as listed on `www.evenementenkalender.nl`,[9] a calendar website to which anyone can submit events. Submitted events are checked by administrators before being placed on the calendar.
- Music Festivals. We extracted an overview from `http://www.festivalinfo.nl/`,[10] a popular festival website maintained by volunteers, that aims to

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

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

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 1. *Overview of the number of events that were collected as gold standard for recall evaluation, divided into the total of curated events, events mentioned at least once in the tweet set, and events mentioned at least five times. The percentage of the total is given between brackets*

	#Curated	#Mentioned	#Mentioned ≥ 5
Football matches	63	51 (81%)	40 (63%)
Public events	2,361	63 (3%)	30 (1%)
Music festivals	518	195 (38%)	98 (19%)
Movie premieres	50	29 (58%)	20 (40%)
Game releases	79	19 (24%)	14 (10%)
Stage performances	1,066	85 (8%)	29 (3%)
Total	4,137	442 (11%)	231 (6%)

provide an exhaustive overview of bigger and smaller music festivals in the Netherlands and Belgium.

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

We performed a subsequent filtering by removing gold standard events that are not mentioned in our tweet set. We compared the name of each event to each of the 27.7 million tweets and listed all tweets that refer to an event name. We subsequently inspected the list of matching tweets to see if they actually mention the event, which is not self-evident for event types such as movies. Any falsely selected tweet was discarded from the list. We performed a second filtering by imposing a minimum threshold of five tweets per event, which is equivalent to the threshold for event significance during the system component of event ranking (Section 3.2.1).

The numbers of reference events by type, before and after filtering, are given in Table 1. A surprisingly small part of the gold standard events are actually mentioned on Twitter (11%). Furthermore, only about half of these are mentioned five times or more (six per cent). The bulk of the gold standard events are stage performances or public events. However, a long tail of public events is either never or hardly ever

[11] `http://www.gamersnet.nl/gamereleases/201408/` and `http://www.gamersnet.nl/gamereleases/201409/`

[12] `http://www.filmvandaag.nl/bioscoop/08-2014` and `http://www.filmvandaag.nl/bioscoop/09-2014`

[13] `http://www.podiuminfo.nl/concertagenda/?input_zoek=&Date_Day=01&Date_Month=08&Date_Year=2014` and `http://www.podiuminfo.nl/concertagenda/?input_zoek=&Date_Day=01&Date_Month=09&Date_Year=2014`

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

mentioned in the tweets. The set of football matches are referred to for the largest part (eighty-one per cent), followed by movie premieres (fifty-eight per cent). The type of events mentioned the most are music festivals, with 195 events mentioned at least once and ninety-eight events mentioned five times or more.

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

5 Results

5.1 *Output*

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

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

To obtain insights into the range of dates at which the events take place, we plotted the number of extracted events per week within rank 250 in Figure 1. The events are more concentrated close to the tweet postings in August (week 31–35). The number drops below ten events from week 39 (September 22nd) onward, but never touches zero in any of the subsequent weeks. Hence, although the bulk of anticipations concerns events within a couple of weeks, our system also captures tweets that refer to events that take place months ahead.

5.2 *Precision*

The precision@250 of the Ngram baseline and the Commonness approach (which is the same for Commonness and Commonness+) is displayed in Table 3[14]. As the output of the Commonness approach was rated by four annotators, the precision can be scored with different degrees of strictness: labeling output as event only when all four annotators identify them as event, when at least three of the four see

[14] A full overview of the events, their assessment and IDs of the tweets that refer to them can be found at http://www.ru.nl/lst/resources/

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 2. *Top-10 ranked events from the commonness systems*

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

them as representing an event, and when half of the annotators do so. The results in the table show that almost two-thirds (sixty-three per cent) of the output of the Commonness approach is seen as event by all four annotators, while only forty-two per cent is scored as such for the N-gram approach. When taking a majority vote of three annotators, the percentage increases to eighty per cent, while a lax setting in which two or more of the annotators identify an event yields a precision of eighty-seven per cent.

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 3. *Precision@250 of output identified as event by human annotators at hundred per cent, seventy-five per cent, and fifty per cent agreement, and Cohen's Kappa and Mutual F-score between the annotators*

	Precision@250				
	100%	75%	50%	Cohen's Kappa	Mutual F-score
N-gram	0.42	–	0.52	0.80	0.89
Commonness	0.63	0.80	0.87	0.48	0.9

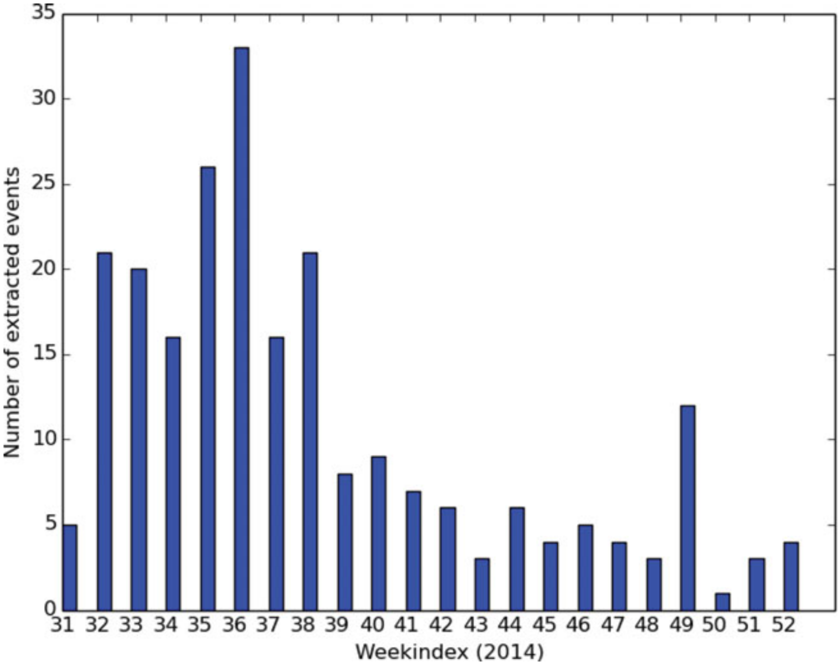

Fig. 1. (Colour online) Counts of the number of extracted events by week number in 2014, from the top 250 events extracted from the Twitter stream in August 2014 (weeks 31–35).

We scored the inter-annotator agreement by Cohen's Kappa and Mutual F-score. The latter provides an insight into the agreement for the positive (event) class. The Kappa score for the N-gram approach is substantial with 0.80, while the agreement for the Commonness events is only moderate with 0.48. However, the mutual F-score shows that the agreement for the positive event class in both approaches is quite accurate with 0.89 and 0.9, respectively.

We plot the precision-at from rank 1 to 250 for the two approaches in Figure 2. Surprisingly, the curves for the N-gram approach show an increasing performance lower down the ranking. It seems that the G_2 log likelihood ratio statistic by which the N-grams are ranked does not relate well to the likelihood that the n-grams signify an event. In contrast, higher rankings for the Commonness approach do relate to event probability. For all three degrees of strictness, a plateau is reached

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

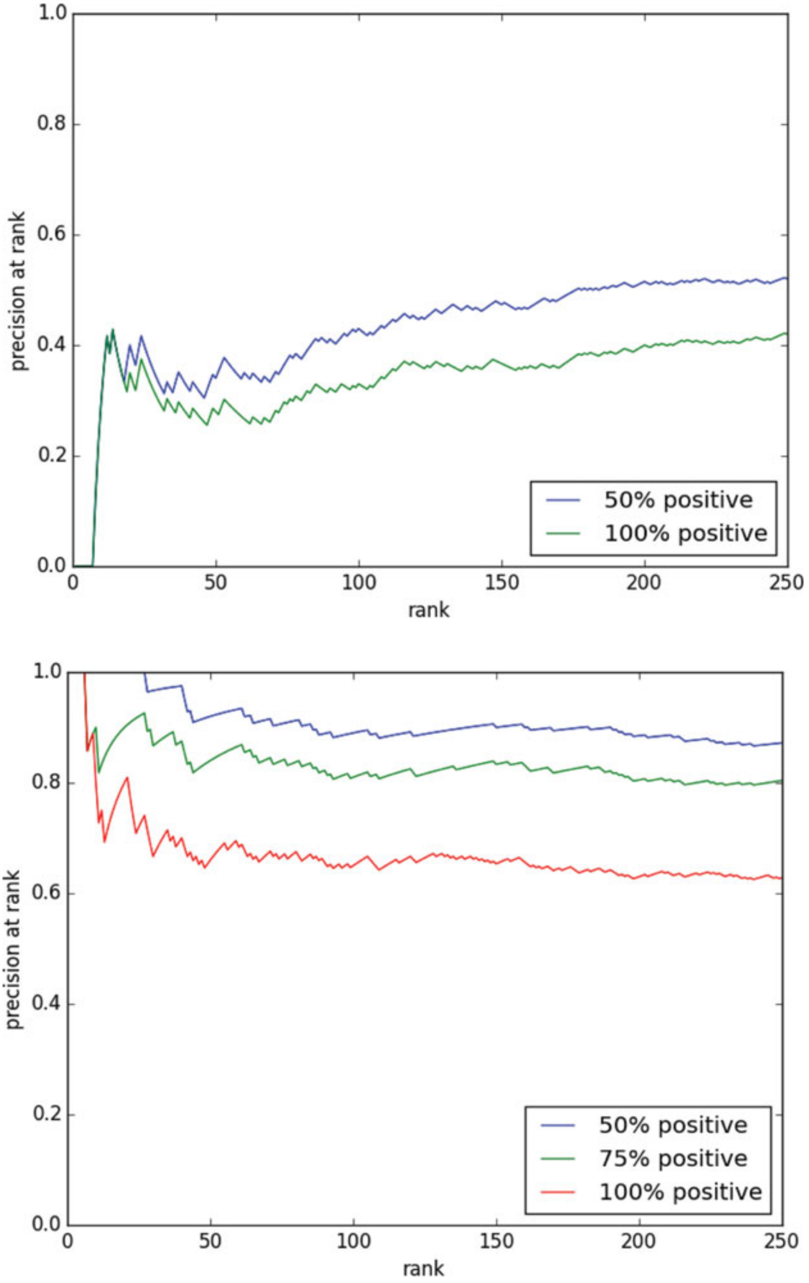

Fig. 2. (Colour online) Precision-at-curves for the N-gram and Commonness approach with different degrees of strictness. (a) Ngram approach. (b) Commonness approach.

after the rank number sixty. Any output up to a ranking of about fifty is seen by at least two annotators as an event.

For any output identified as event, annotators are asked to assess the quality of the event terms in relation to the event. The outcome of these assessments is presented in Table 4. The assessment was given on a scale from one to three, as a

Table 4. *Average assessment of terms for all three approaches. Assessment is on a scale of 1 (bad) to 3 (good)*

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

poor, moderate, or good representation. The terms representing an event could be either assessed by 0, 1 or both annotators, as the assessment was only presented if an annotator rated the tweets as representing an event. We calculated the average of all event assessments. When only one of two annotators gave an assessment, this value was adopted as event assessment. When both annotators gave an assessment, their average was taken as the event assessment. The agreement was scored with the Weighted Cohen's Kappa metric (Gwet 2001), in line with the ordinal annotation. Missing fields were taken into account in this metric.

The average assessment of terms does not show a large difference between the three approaches. Surprisingly, the Commonness+ approach, for which terms were added in a post-processing step, are assessed as a slightly worse representation than the terms for the Commonness approach, on average. The agreement is only slight or poor. This is in line with *post-hoc* remarks made by several annotators that it was hard to assess the quality of the terms.

5.3 Recall

To assess recall, we collected gold standard events that took place in August or September 2014, for six event types from curated websites (see Section 4.3). We compared the events that were extracted by the N-gram and Commonness system to the gold standard events that were tweeted about at least five times (the last column in Table 1). For both systems, we report a recall@250, which relates to the precision oriented evaluation in the previous Section, as well as a recall of all events (318 for N-gram and 966 for Commonness).

The results are given in Table 5. The Commonness approach outperforms the N-gram approach for each of the six event types. It yields the best performance in retrieving football matches and game releases. The N-gram approach fails to retrieve any event for some of the types. Overall, the Commonness approach scores a recall of 0.20 at rank 250, and a recall of 0.40 for all 967 events on these event types.

We apply the G_2u formula (Section 3.2.1) to rank events, which is an extension of the G_2 formula. As a comparison of the two formulas, we implemented the Commonness system with both and scored the recall at each rank by comparing the extracted events to the accumulated gold standard events of all event types (242 events in total). The recall at each rank is plotted in Figure 3. The shorter line of the G_2 rank is due to a larger number of discarded events. The G_2u rank has a comparable recall to G_2 up to rank 100, but retrieves increasingly more events

Table 5. *Recall performance by event type, based on a gold standard set of events that are mentioned in at least five tweets (see Table 1 for the exact numbers)*

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

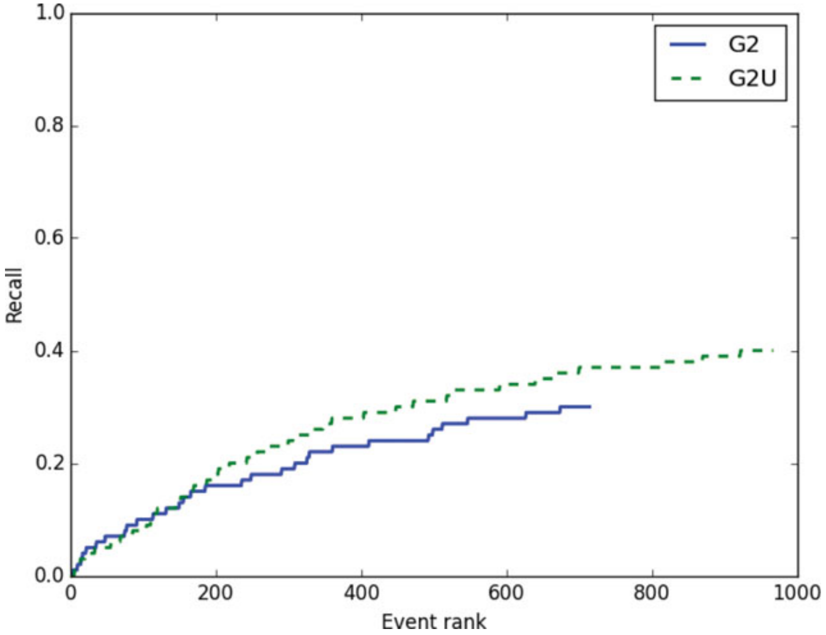

Fig. 3. (Colour online) Recall-at-curves of the Commonness system, ranked by either the G_2 formula or the G_2U formula.

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

6 Analysis

6.1 Event output

To obtain insight in the causes of non-event output, annotator disagreement, and the assessment of event terms, as well as the impact of the event term clustering

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

component, we analyzed the top-250 events from the Commonness approach in relation to the annotator assessments. We analyzed all 250 events on their five event tweets and the event terms of both the Commonness and Commonness+ approach.

6.1.1 Event annotation

Of the 250 annotated events, 157 are annotated by all four annotators as event, leaving ninety-three events that are deemed doubtful by at least one annotator. Of these ninety-three events, forty-four are still annotated by three of four annotators as event, seventeen by half of them, fourteen by only one annotator, and for eighteen entries all four annotators agree that they are not an event. We analyzed the five tweets that were shown to the annotator for these ninety-three events, and distinguished six causes for an annotator to doubt if all tweets represent the same event:

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

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

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

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

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

(6) Contest – The event tweets advertise about a product or participate in a contest. Example: *@afcajax because my friend is only free on Sunday and we would really*

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

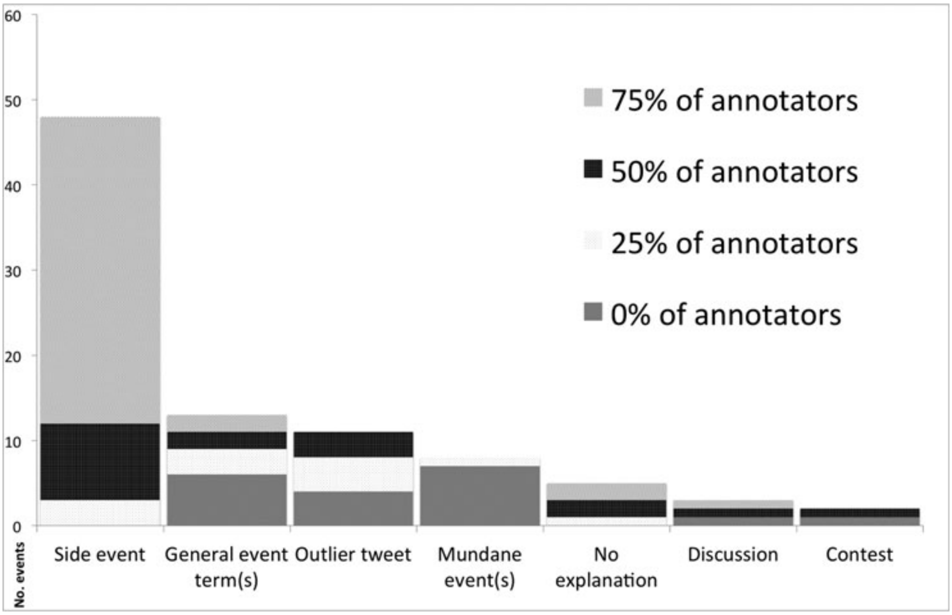

Fig. 4. Overview of output properties when they are not rated as event by at least one annotator, divided by the percentage of annotators who rated the output as event per property.

like to go to the match together #weareajax. This tweet joins a contest to win free tickets to a football match by stating a motivation.

We tallied the occurrences of each category and made a division by the percentage of annotators that nonetheless deemed the occurrence an event. The outcome is displayed in Figure 4. The bar chart shows that about half of the entries with negative annotations are due to side events being mentioned. In most cases, a majority (seventy-five per cent) of the annotators still judged the event cluster as a proper event. On the other hand, general event terms and mundane events are decisively not seen as event. Event tweets that include an outlier tweet (the third bar in Figure 4) might still be seen as event by some.

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

6.1.2 Assessment of event terms

We implemented a component in Commonness+ to add additional event terms and improve the event description. However, as is shown in Table 4, on average the annotators assess an event description better if no event terms were added. To analyze the cause of this outcome, we observed the terms and the assessment of

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 6. *Overview of possible combinations between the event terms of the Commonness and Commonness+ approaches and the assessment by annotators for any event output. Only the 214 events that were assessed for both approaches are included in the counts*

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

the Commonness and Commonness+ approaches. The four combinations that we found are displayed, along with their number of occurrences, in Table 6.

The Commonness+ approach does not always result in the addition of terms, but for ninety-three per cent of the events it does. In these cases, the addition of terms most frequently yields a similar assessment as the standard Commonness terms. Most striking, however, is that for thirty per cent of the events, the added terms include redundant information and the result is therefore valued worse than the standard terms. This percentage outweighs the number of times that the addition of terms is actually beneficial for the event description (twenty-four per cent).

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

This analysis shows that the approach to add event terms should be improved to minimize the output of redundant event terms. Apart from this, the design of the evaluation might have been of influence. We asked the annotators to assess event terms after having judged the tweet cluster to be an event. The quality of the event terms to describe the nature of the event without any prior knowledge is not assessed. In the example given above, ‘gwn, Ed Sheeran, concert’ might be valued better than ‘Ed Sheeran’ as clues on the type of event.

6.1.3 Characteristics of extracted events

In addition to analyzing the tweets and event terms in relation to annotator assessment, we checked the events for duplicates, if they took place in the future and if there were plannings of demonstrations in the output.

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

We kept a record of the number of duplicates and the number of clustered event terms in the top-250 output. This relates to the event clustering component (Section 3.2.2), which is aimed at diminishing the number of duplicates. In the output we found a total of seventeen duplicates (6.8 per cent of the total), while sixty-nine event terms were clustered with other event terms. This shows that the clustering module does combine many event terms. However, the part of the top-ranked output that is redundant shows that there is still room for improvement. A detailed evaluation of the clustering module is described in Section 6.2.3.

During the evaluation, the annotators were asked to assess whether the output was an event. It was not specified in the annotator guidelines that the tweets should refer to a future event, although this is an explicit goal of our system. For example, in Sections 3.1.1 and 3.2.3, we describe approaches to filter tweets and events that take place in the past. As a check, we analyzed the ‘futureness’ of the top-250 events and found that all output that was assessed as event by the annotator was indeed a future event at the time the last tweet in the set was posted.

In Section 1, we mention that the functionality of our system can assist security by extracting and displaying upcoming demonstrations that might form a security risk. In the top-250 output, we found two events of this type: a demonstration against ISIS in The Hague, organized by Pro Patria (a group on the right side of the political spectrum), and a demonstration during the opening of the academic year in Maastricht. As such demonstrations are sensitive to annulments, we inspected whether these two events actually took place. We found that the demonstration in The Hague, planned for September 20th, was canceled one day before.

6.2 Assessment of components

6.2.1 Rule based extraction of future referring time expressions

In Section 3.1.1, we mentioned that the Heideltime tagger (Strötgen and Gertz 2010) fails to detect some of the relevant time expressions in Dutch, so that it makes sense to work with manually formulated rules. To substantiate this statement, we applied the Heideltime tagger to the tweets that we used in our experiment, as described in Section 4.1, and compared the output of the two approaches.

We used Heideltime version 1.8.[15] We set the language to Dutch and the document type to ‘news’ (other options were narrative, colloquial, and scientific). In line with our rule-based system, we focused on time expressions that point to a future date. Hence, any time tag in the output of Heideltime that points to a duration or a date in the past was not taken into account. Also, ‘tomorrow’ was excluded as was deliberately done in our rule-based system.

The performance of the two approaches is displayed in Table 7. The rule-based approach outperforms Heideltime in terms of the number of extracted tweets with a future referring TIMEX. Of these 367,206 tweets, thirty-seven per cent is also extracted by Heideltime, leaving 231,641 additional tweets only found by our rules;

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 7. *Comparison between Heideltime and the rule-based approach to finding tweets with a TIMEX that points to a future date, from the August 2014 tweets that were used in the main experiment*

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

103,517 tweets are only found by the Heideltime tagger. If we regard the combined output of both approaches as gold standard and take the union of the tweets that are extracted by them (totaling 470,723), the rule-based approach has a recall of 0.78 and Heideltime has a recall of 0.51. This recall should be seen as an approximation, as we do not know which of the TIMEXs the two approaches failed to retrieve, or which of the extracted TIMEXs are correct.

An analysis of the exclusive tweets that are extracted by both approaches shows that the rule-based approach succeeds in extracting TIMEXs that explicitly mention a specified amount of days in the future, like ‘nog 12 nachtjes slapen’ (another twelve nights of sleep). Most tweets that were extracted exclusively by the Heideltime tagger contain TIMEXs like ‘volgende week’ (next week) and ‘dit weekend’ (this weekend). A disadvantage of such phrases is that it is hard to link them to a specific date, as they refer to spans of two or more days.

6.2.2 Extraction of entities

For the extraction of entities from tweets, we applied the commonness approach as described by Meij *et al.* (2012), which does not rely on common NED markers such as part-of-speech tags or capitalization. To obtain an impression of its performance, we annotated a sample of 1,000 tweets by their named entities and compared the performance of commonness to an existing NED system for Dutch, the NED component in Frog (Van den Bosch *et al.* 2007). We converted the output of both approaches for these sentences, as well as the annotated sentences, into the IOB-tagging format, and evaluated them with the CoNLL-2000 shared task evaluation script.[16] For commonness, possible overlap between output was resolved (Section 3.3.1).

We estimated significance in the differences between the commonness method and Frog’s NED by using bootstrap resampling (Noreen 1989). Per system, we selected 250 random samples of sentences. We assume that performance A is significantly different from performance B if A is not within the center ninety per cent of the distribution of B. Results are presented in Table 8.

The Frog NED system is outperformed by the commonness approach both in terms of recall and precision, with a resulting F1 score of 0.63 for commonness

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

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 8. *Significance estimates of commonness and Frog NED in retrieving entities from an annotated sample of 1,000 tweets. Scores are obtained after bootstrap resampling with 250 samples*

	Precision	Recall	F1
Commonness	0.50	0.87	0.63 ± 0.02
Frog NED	0.37	0.82	0.51 ± 0.01

against 0.51 for Frog NED. The difference is significant with small standard deviations of 0.02 and 0.01, respectively. This again shows that off-the-shelf tools are lacking generalization power when applied to non-standard language. Applying the commonness approach in such settings can be an effective replacement.

6.2.3 Event clustering

Event clustering is an important component in our system, aimed at reducing duplicate event output and enhancing event descriptions. In Section 6.1.3, we report on sixty-nine clusterings of event terms and seventeen duplicate events in the top-250 generated events. As these numbers do not give a complete overview of the clustering performance, we also evaluated all clusterings that were made on the initial set of 2,500 date-term pairs.

We manually made clusters of the 2,500 date–term pairs, by inspecting the tweets in which each event term is mentioned, and compared the automatic clusterings to this reference set. We evaluated performance by inspecting the *pairs* of event terms that were clustered (Halkidi, Batistakis and Vazirgiannis 2001). Any pair that was clustered in the reference set but was not clustered by the clustering component was added to the false negatives, while any pair that was clustered by the clustering component but should not have been clustered was added to the false positives. We used these numbers to assess precision, recall and F1. We also calculated the Rand Index (Rand 1971), an accuracy metric that not only takes into account objects that are classified in the same cluster, but also rewards objects that are rightfully not clustered together (the true negatives).

The cluster performance is presented in Table 9. The optimal clustering would result in 2,370 merges of event term pairs. The clustering component actually makes 822 correct merges, and incorrectly merges 156 pairs. This results in a precision of 0.84 and a recall of 0.35. The high value of the Rand Index, 0.97, is due to the large number of true negatives. These results show that the clustering component manages to merge part of the duplicates, at a fairly low rate of false positives. Nonetheless, this component leaves room for improvement, especially with regard to recall.

7 Conclusion and discussion

We propose a system for open-domain event extraction. An adaptation of the work by Ritter *et al.* (2012), it operates in a more unsupervised way and can be implemented relatively easily for any language, provided that a rule set is written

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table 9. *Performance of the clustering component. RI = Rand Index. #Total = the term pairs that should be merged according to a manual gold standard clustering. #Merged = the merges made by the clustering component. #Correct = the correctly merged pairs*

	Precision	Recall	F1	RI	#Total	#Merged	#Correct
Before clustering	0.00	0.00	0.00	0.95	2370	0	0
After clustering	0.84	0.35	0.49	0.97	2370	978	822

for detecting future time references. Central to the system is the extraction of *event terms*, for which we apply the commonness approach (Meij *et al.* 2012). Additional event terms are added based on the $tf * idf$ of words in the event tweets.

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

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

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

In future work, we aim to improve our system by working on the most common errors made by the current version of the system. Specifically, the addition of event terms includes too many redundant terms; some duplicate events pass through the clustering stage. Also, sometimes similar mundane events mentioned by several persons are merged into a single event, such as carpooling requests. Similarly, occasionally unrelated events occurring in the same location are merged. As another improvement, we aim to extend the set of time expression rules, in order to increase the recall of events.

Apart from improving the system, we plan to enrich the output with information on event type, which we aim to categorize automatically. Such a categorization would enable us, for example, to identify planned events, such as protests, and observe whether the event will actually take place and if it will escalate. Another strand of planned future research is to mine additional tweets related to an event in addition to those containing an explicit future time reference. These tweets may produce additional event terms, and may be used to provide more (and more abundant) information about an event, e.g. for automatically analyzing their degree

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

of sentiment and emotion. Such tweets can also reveal whether an event actually took place at the announced date.

References

Aggarwal, C., and Subbian, K. 2012. Event detection in social streams. In *Proceedings of SIAM International Conference on Data Mining,* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 624–35.

Becker, H., Iter, D., Naaman, M., and Gravano, L. 2012. Identifying content for planned events across social media sites. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining,* ACM, New York, NY, USA, pp. 533–42.

Benson, E., Haghighi, A., and Barzilay, R. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Stroudsburg, PA, USA, vol. 2, pp. 389–98.

Bosch, A. van den, Busser, B., Canisius, S., and Daelemans, W. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational Linguistics in the Netherlands: Selected Papers from the 17th CLIN Meeting*, LOT, Utrecht, pp. 99–114.

Cohen, M. J., van den Brink, G. J. M., Adang, O. M. J., van Dijk, J. A. G. M., and Boeschoten, T. 2013. *Twee werelden, You Only Live Once*.

Cordeiro, M. 2012. Twitter event detection: combining wavelet analysis and topic inference summarization. In *Doctoral Symposium on Informatics Engineering, DSIE*, Faculdade de Engenharia da Universidade do Porto, Porto.

Day, W. H., and Edelsbrunner, H. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification* **1**(1): 7–24.

Diao, Q., Jiang, J., Zhu, F., and Lim, E. P. 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 536–44.

Gwet, K. 2001. *Handbook of Inter-Rater Reliability*, Gaithersburg: Advanced Analytics, LLC.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* **17**(2): 107–45.

Jackoway, A., Samet, H., and Sankaranarayanan, J. 2011. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ACM, New York, NY, USA, pp. 25–32.

Kumar, S., Liu, H., Mehta, S., and Venkata Subramaniam, L. 2014. From tweets to events: exploring a scalable solution for twitter streams, arXiv preprint arXiv:1405.1392.

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

Li, C., Sun, A., and Datta, A. 2012. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, USA, pp. 155–64.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, ACM, New York, NY, USA, pp. 409–18.

Meij, E., Weerkamp, W., and de Rijke, M. 2012. Adding semantics to microblog posts. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, ACM, New York, NY, USA, pp. 563–72.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*, New Jersey: Wiley-Interscience.

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Ou, G., Chen, W., Wang, T., Wei, Z., Li, B., and Yang, D. 2014. Exploiting community emotion for microblog event detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1159–68.

Ozdikis, O., Senkul, P., and Oguztuzun, H. 2012. Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*, ACM, New York, NY, USA.

Petrović, S., Osborne, M., and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 181–9.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**(336): 846–50.

Reuter, T., and Cimiano, P. 2012. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, New York, NY, USA.

Ritter, A., Clark, S., and Etzioni, O. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1524–34.

Ritter, A., Mausam, Etzioni, O., and Clark, S. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, pp. 1104–12.

Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, ACM, New York, NY, USA, pp. 851–60.

Strötgen, J., and Gertz, M. 2010. HeidelTime: high quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 321–4.

Tao, K., Abel, F., Hauff, C., Houben, G. J., and Gadiraju, U. 2013. Groundhog day: near-duplicate detection on twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, ACM, New York, NY, USA, pp. 1273–84.

Tjong Kim Sang, E., and van den Bosch, A. 2013. Dealing with big data: the case of Twitter. In *Computational Linguistics in the Netherlands Journal* **3**: 121–34.

Valkanas, G., and Gunopulos, D. 2013. How the live web feels about events. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, ACM, New York, NY, USA, pp. 639–48.

Weerkamp, W., and Rijke, M. de 2012. Activity prediction: a twitter-based exploration. In *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access, TAIA-2012*, ACM, New York, NY, USA.

Weiler, A., Scholl, M. H., Wanner, F., and Rohrdantz, C. 2013. Event identification for local areas using social media streaming data. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*, ACM, New York, NY, USA, pp. 1–6.

Weng, J., and Lee, B. S. 2011. Event detection in twitter. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM-11)*, AAAI Press, Palo Alto, CA, USA, pp. 401–8.

Zhao, S., Zhong, L., Wickramasuriya, J., and Vasudevan, V. 2011. Human as real-time sensors of social and physical events: a case study of Twitter and sports games. Technical Report TR0620-2011, Houston, TX: Rice University and Motorola Labs.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., and Yan, H. 2011. Comparing twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. Jones,

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

W. Kraaij, H. Lee, and V. Murdock (eds.), *Advances in Information Retrieval*, pp. 338–49. Berlin: Springer Verlag.

Appendix A Rules for the extraction of time expressions

Table A1. *Date related rules for the extraction of time expressions. Values in the columns can be combined sequentially*

Day value	Hyphen (optional)	Month value	Hyphen (optional)	Year value (optional)
[1–31]	–	[1–12]	–	20[14–99]
een		januari		
twee		februari		
drie		maart		
vier		april		
vijf		mei		
...		juni		
zevenentwintig		juli		
achtentwintig		augustus		
negenentwintig		september		
dertig		oktober		
eenendertig		november		
		december		

Table A2. *Exact rules for the extraction of time expressions. Values in the columns can be combined sequentially*

indication of future moment	optional part	number of days	time unit	optional part	optional part
over	minimaal	[1–365]	dag(je)	(nog)te	tot
(met)nog	maximaal		dagen	-gaan	
	tenminste		daagjes	slapen	
	bijna		nacht(je)		
	ongeveer		nachtjes		
	maar		nachten		
	slechts		weken		
	pakweg		week(je)		
	ruim		weekjes		
	krap		maand(je)		
	(maar)een		maandjes		
	-kleine		maanden		
	(maar)iets				
	-(meer/minder)				
	-dan				

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press

Table A3. *Rules for the extraction of time expressions that contain a weekday. Values in the columns can be combined sequentially*

Time indication (optional)	Weekday	Part of day (optional)
volgende week	maandag	ochtend
	dinsdag	middag
	woensdag	avond
	donderdag	nacht
	vrijdag	
	zaterdag	
	zondag	

Appendix B The instruction letter for event evaluation (translated from Dutch)

We have developed a system that fully automatically detects events from the big stream of Dutch tweets. You will test the output of this system. You will judge fifty events in total. This will take about twenty minutes. You can close this survey at any moment and at a later time click the link to repeat the survey. As a start, read the instructions below thoroughly.

You will get to see five tweets each time. We ask you to indicate whether they all refer to the same event. To identify an event. you should make use of the following definition:

An event is something that happens at a specific time and is important to a larger group of people.

Sports matches and law amendments qualify as event in this definition, while a holiday to Turkey is too personal to qualify as event.

Warning: sometimes several events are described in a tweet, such as an initiative by the supporters of a football club during a match. If all five tweets indirectly refer to the same football match in this way, they do refer to the same overarching event. However, if five tweets describe different events in the city of Amsterdam, this does not qualify as the same event. These different events are not linked by a common event.

In case of a positive answer, a second question will appear. You will get to see one or more terms that describe the event, and are asked if these terms are a good, moderate or bad representation of the event.

Good luck!

https://doi.org/10.1017/S1351324916000036 Published online by Cambridge University Press