Computational Linguistics in the Netherlands Journal 4 (2014) 39-52 Submitted 06/2014; Published 12/2014

Timely identification of event start dates from Twitter

Florian Kunneman* F.KUNNEMAN@LET.RU.NL
Ali Hürriyetoğlu* A.HURRIYETOGLU@LET.RU.NL
Nelleke Oostdijk* N.OOSTDIJK@LET.RU.NL
Antal van den Bosch* A.VANDENBOSCH@LET.RU.NL

* *Centre for Language Studies, Radboud University*
P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands

Abstract

We present a method for the identification of future event start dates from Twitter streams. Taking hashtags or event name expressions as query terms, the method gathers a certain number of tweets about an event and uses clues in these tweets to estimate at what date the event will start. Clues include temporal expressions with knowledge-based and automatically generated estimations, and other predictive words. The estimation is performed either with a machine-learning classifier or by taking a majority vote over the temporal expressions found in the set of tweets. Results show that temporal expressions are indeed strong predictors. The majority-based and machine-learning approaches attain equal performances when trained and tested on a single event type, soccer matches, with an average estimation error of 0.05 days; but when tested on a range of different events, the majority-voting approach shows to be more robust than machine learning for this task, yielding high performance on all events. Still, per-event differences hint at a context in which machine learning might be beneficial.

1. Introduction

A substantial number of posts on Twitter[1] refer to real-world events and report on what is currently happening. While most of these tweets are posted as the event unfolds, some refer to events that have not happened yet. We aim to leverage the latter type of posts in order to identify, in a continuous way and on unseen data, the dates at which future events take place.

A calendar of future events fed by the most recent information available on the social media may be of interest to various user groups. One such group may be journalists who would like to know about future events that emerge first on social media. The idea of publishing future calendars with potentially interesting events has been implemented before and is available through services such as Zapaday[2], Daybees[3], and Songkick[4]. However, these services do not mine social media automatically; to our knowledge, based on the public interfaces of these platforms, these services perform directed crawls of (structured) information sources, and identify exact date and time references in posts on these sources. They also manually curate event information, or collect this through crowdsourcing. Using Twitter instead as source of information is of particular interest as it might carry information about events that are not picked up by these services, especially when they are informal, regional, or relatively small-scale.

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

1. *Preparing the last few things for my beyoncé concert tomorrow*

1. http://twitter.com
2. http://www.zapaday.com
3. http://daybees.com/
4. https://www.songkick.com/

2. *86 days to #WC2014 - He's the only man born on 18-March to have scored at the World Cup. Guess who? pic.twitter.com/K9mGVYXKWQ*

3. *The Walking Dead Creator Eagerly Anticipates Mario Kart 8 http://wp.me/pLMz-htu*

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

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

The outline of the remainder of this paper is as follows. In Section 2 we discuss related work. In Section 3 we describe experimentation on a development set of tweets referring to Dutch soccer events, after which in Section 4 we describe the application of our methods on a selection of other event types. In Section 5 we formulate our conclusions and suggest points to be addressed in future research.

2. Related work

We study the early identification of an event start date from tweets referring to that event. Most of the previous studies on event mentions on Twitter focus on the tweets during or right after an event (Chakrabarti and Punera 2011, Jackoway et al. 2011, Becker et al. 2012, Quezada and Poblete 2013). In other research, the sudden bursts of tweets with common key terms ('trending topics') are leveraged to detect unknown events (Ozdikis et al. 2012, Qin et al. 2013, Weiler et al. 2013, Valkanas and Gunopulos 2013, Chunara et al. 2012, Zhou and Chen 2013).

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

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

3. Soccer events

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

3.1 Experimental set-up

3.1.1 Data

As data for our experiments we selected 60 soccer matches played in the Dutch premier league. We harvested tweets by means of `twiqs.nl`, a database of Dutch tweets from December 2010 onwards (Tjong Kim Sang and van den Bosch 2013). We selected the (average) top 6 teams of the league[5], and queried all matches played between these teams in 2011 and 2012. For each query, the conventional hashtag for a match was used with a restricted six-week search space, viz. three weeks before the match until three weeks after, so as to avoid overlap with another match between the same two teams. The queries resulted in tweet streams for a total of 60 matches; a total of 703,767 tweets. Of these, 269,753 are posted before event time. The number of tweets per event ranged from 321 to 35,464. Retweets were removed, as they only duplicate data and may pass on a previous tweet with a different TTE. The resulting set without retweets contains 411,784 tweets, of which 140,060 are posted before event time.

Every tweet in our data set has a time stamp of the moment it was posted. Moreover, for each soccer match we know when it took place. This information is used to calculate for each tweet the actual time that remains to the event in terms of the number of days and the error in estimating the TTE [6].

3.1.2 Features

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

For the extraction of TEXs we make use of the extensive list of Dutch TEXs that was compiled by (Hürriyetoğlu et al. 2014)[7]. Although Heideltime (Strötgen and Gertz 2010) has a module to extract Dutch time expressions, the list of TEXs compiled by Hürriyetoğlu et al. (2014) is more extensive and tuned to our research aims. In line with the approach of Hürriyetoğlu et al. (2014), we estimated the TTE of TEXs in two ways: trained and rule-based.

In the trained approach, the TTE estimation linked to a TEX is derived in a data-driven way, which we call 'Timelearn' (TL). By collecting all occurrences of a TEX and the observed TTE in days during training, we can calculate the median of this set, which we then take as the TTE estimate attributed to the TEX. We excluded TEXs of which the observed TTE had a standard deviation higher than 2 (i.e. two days). Examples of trained TTE estimations are given in Table 1. Specific TEXs ('only 2 weeks') are indeed linked to the specified TTE based on the training data. The added value of this approach is the estimation for less specific TEXs (such as 'only a couple of days').

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

Table 1: Examples of the median TTE in days for TEXS as calculated from training tweets

5. Ajax, Feyenoord, PSV, FC Twente, AZ Alkmaar and FC Utrecht
6. The tweet IDs for all 60 events, along with their calculated time to event, can be downloaded from `http://cls.ru.nl/~fkunneman/data_tte_estimation.zip`
7. The list has been made available through `http://www.ru.nl/lst/resources/`

In addition to a data driven estimation, the TTE from a TEX can be estimated manually by common-sense annotation. Hürriyetoğlu et al. (2014) distinguish two kinds of rules: dynamic and exact. Dynamic rules apply to TEXs of which the TTE is dependent on the point in time at which they are posted. A mention of a date should be linked to the date of posting to calculate the difference, and a mention of a weekday should be linked to the weekday at which it was mentioned. The example 'next Sunday afternoon', that is given in Table 1, would be calculated as the first Sunday from the day of tweeting (3 days if Thursday is the day of posting). Exact TEXs imply a TTE that can be estimated without information on the moment at which they are posted. Examples are 'tomorrow' and 'another 12 days before'.

Because Hürriyetoğlu et al. (2014) focus on estimating the TTE in terms of the number of hours within a frame of eight days, the list of TEXs does not include many expressions that relate to longer periods of time. Therefore, we complemented the list with a number of additional expressions with a longer range in time. For example, expressions like 'nog [1-8] dagen' ('only [1-8] days') were extended to a range of 21 days, and also expressions like 'over 3 weken' ('in three weeks') were included.

The third type of features, word n-grams, were extracted after preprocessing of the data. All characters in the tweets were lowercased and usernames and URLs were normalized into the dummy features 'USER' and 'URL'. The tweets were tokenized with ucto[8] and surrounded by beginning- and end-of-tweet markers. We extracted word unigrams, bigrams, and trigrams from the tweets.

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

3.1.3 Representing time windows

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

Different window sizes can be chosen. Long windows might lead to more accurate estimations, but sampling a long window takes longer: with a window size of 100 tweets, the first estimation can only be made after 100 tweets have been seen. In view of the scarcity of tweets that are posted many days before an event takes place, an estimation based on a large window might be made only right before or even after the event start time. From this perspective, smaller windows are favourable. To test for the optimal window setting we alternated 3 different window sizes: 50, 20 and 10 tweets.

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

For training and testing, the windows of tweets were given a label based on the actual TTE of the last tweet in the window. Windows of which the last tweet was posted during or after the event

8. `http://ilk.uvt.nl/ucto`

time were given the label 'during' and 'after' respectively[9]. Thus, the labels were any TTE in days from 21 until the day of the event, and during and after. The features in a window were derived by aggregating the features of the separate tweets in the window. In the case of rule-based features, the date that was derived from such features was translated into the TTE at the time of the last post of the window.

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

3.1.4 Prediction

We tested the different feature combinations and window settings on the 60 soccer events by means of ten-fold cross-validation, training on the windows in 54 events and testing on the 6 remaining events. For each fold, we trained and tested a Support Vector Machines (SVM) classifier based on libsvm (Hsu et al. 2003). During training, additional preprocessing was needed to facilitate the classifier. Given that the vast majority of tweets are posted right before, during and after an event, the number of instances per TTE label, or 'during' or 'after', is highly imbalanced. To avoid a classifier bias towards TTE close to event time, we balanced the number of instances per TTE label by a combination of undersampling and oversampling. This dual approach limits excessive duplication and removal of instances (Sappelli et al. 2013). After balancing the data set, we reduced the feature space by selecting the 10,000 most frequent features. We made use of error-correcting output codes (James and Hastie 1998) to obtain a single classification for each window. The different parameters of SVM were tuned by splitting the training data into 5 folds and performing classifications based on 10 random parameter combinations from a grid. The grid tested a linear, polynomial and RBF kernel, different values of C and *gamma* and different degrees.

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

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

3.1.5 Evaluation

We evaluate the variants of our method by taking the number of days that estimations are off on average, and by measuring the accuracy of predicting the correct date. Estimation errors are calculated with two evaluation metrics. The first one is the Mean Absolute Error (MAE), which sums for each estimate the absolute number of days that it is off, and takes the average for all window estimates. The second one is the Root Mean Squared Error (RMSE): a sum is made of the squared differences between the actual TTE and the estimated TTE; the square root is then taken to produce the RMSE of the prediction series. MAE can be interpreted as the average number of days a method is off. RMSE penalizes large errors more heavily. The overall MAE and RMSE for the 60 events is calculated as the mean of all event MAE and RMSE scores respectively.

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

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

Table 2: MAE (with standard deviations) averaged over the 9 methods applied on all 60 events per window and step combination, versus actual average TTE at first estimation, and average duration of the first estimate, per window.

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

Table 3: Performance of different settings on 60 soccer events for a window size of 50 and a step size of 10 (Rules = Dynamic and exact Rules, TL = Timelearn, ML = Machine Learning, n-grams= Word n-grams).

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

3.2 Results

To get an impression of the quality of the multiple window and step sizes, we calculated the average MAE per window and step size across the applied methods, shown in Table 2. Estimation errors are within one day for window sizes 20 and 50. The largest window size of 50 tweets produces the lowest errors, although the first estimation is only made a week before the start of the event rather than two weeks in the case of a window size of 10. Smaller step sizes lead to better estimates.

To compare the methods, we select the window and step combination that leads to the best performance for most methods: a window size of 50 and step size of 10. The results for this combination are presented in Table 3. The incorporation of history knowledge when the final estimation is made, shown in the right half of the table, shows to be rather beneficial for the performance of any method. Surprisingly, SVMs are nearly always outperformed by either of the more straightforward majority-voting methods, with a very high accuracy of 0.97 and a very low MAE of 0.05 in the history-sensitive variant. Apparently, the manually-set TTE estimations already provide sufficient information. The majority-voting method combining the trained and rule-based TTEs offers no improvement over the method that only uses the rule-based TTEs.

The best ML method is the one in which the three feature types are combined, leading to a performance roughly equal to the majority-voting methods. The RMSE, which is more sensitive to higher errors, is even better for this method. The two TEX feature types, TL and Rules, are

not effective by themselves when fed to the SVM classifiers, while word n-grams do lead to reliable estimations with a MAE of less than half a day.

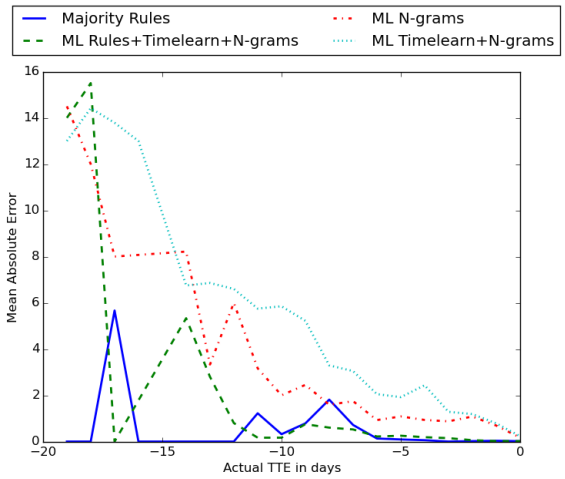


Figure 1: MAE per actual TTE in days averaged over the absolute errors for all 60 soccer events, for a window size of 50 and a step size of 10

To get an impression of the performance related to actual TTEs, we plotted the MAE per actual TTE averaged over all 60 soccer events for a selection of four methods in Figure 1. Interestingly, the majority-voting method obtains a flawless performance between an actual TTE of -17 and -13 days, while the ML method with all three feature types performs better between 13 and 7 days before event time. The two other methods lag behind for every TTE.

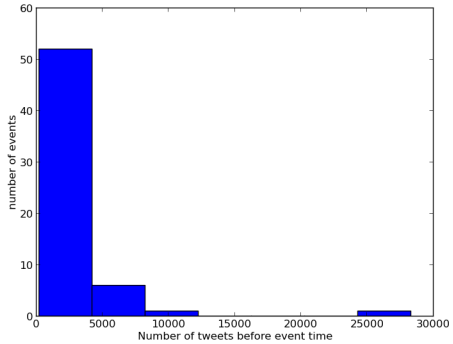
3.3 Error analyses

3.3.1 Quantitative Error Analysis

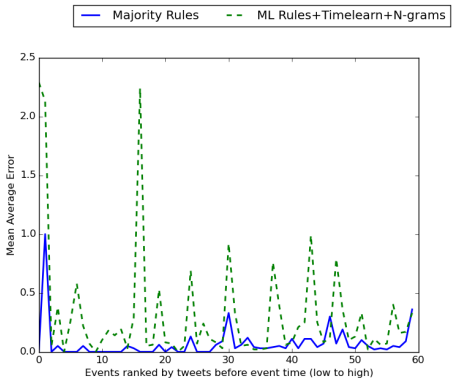
In section 3.2 we assessed the performance of our systems by averaging over the 60 events that were held out for testing. To acquire a sense of the performance at the event level, we visualize some characteristics of the events and the influence of these characteristics on performance in Figure 2: the number of tweets per event (Figure 2 (a) and 2 (b)), and the proportion of correct matching rule-based time expressions per event (that match the actual date of the event, Figure 2 (c) and 2 (d)). We focus on the performance of the best Majority method and the best ML method: Majority using Rules and ML using all three feature types, both using a window size of 50 and a step size of 10 tweets. The performance is scored by Mean Absolute Error. History information, that smoothes away some of the errors, is not included to highlight the actual errors.

Figure 2 (a) indicates that the bulk of the events in our set are referred to before event time in less than 5000 tweets. There is one outlier event that was tweeted about for over 25 thousand times. Figure 2 (b) displays the MAE by event size rank. The absolute number of tweets per event does not seem to have a big influence on the performance of both methods, that show peaks for both low and high ranked events. If anything, the Majority method seems to be suitable for smaller events, as it does not yield high peaks between rank 3 and 23.

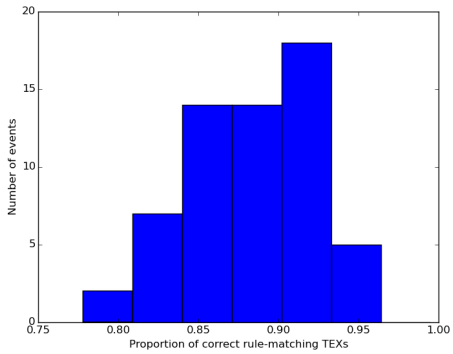
Another relevant characteristic is the proportion of correctly induced dates for all tweets per event, which ranges from 0.77 to 0.97 (Figure 2 (c)). The expectation that this proportion is directly correlated with the success of the Majority method is reflected in Figure 2 (d). When going higher up the ranking, most induced event dates are accurate and the error peaks get smaller. On



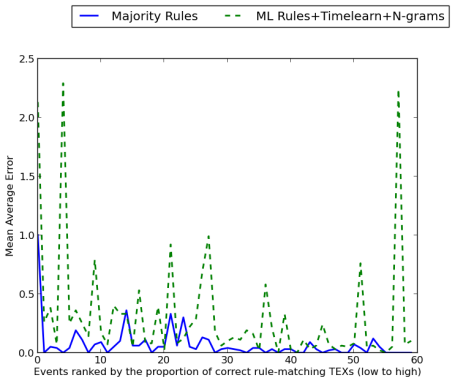
(a) Histogram of the number of tweets before event time, for all 60 football events



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



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



(d) performance by MAE on the events ranked by the proportion of correct Rule-induced TEXs from low to high

Figure 2: Overview of the characteristics of and performance on separate football events

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

3.3.2 Qualitative Error Analysis

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

The most important reason for the Majority system to err was when a number of tweets referred to a date that did not refer to the football match itself, but to a 'side event'. Two side events were most prominent: first the action of buying a ticket for the match, and second the anticipation of another football match played by the same team(s). Below, we present some examples of references to these side events that were seen in our data. We translated the tweets, Dutch by origin, to English:

1. *Gonna buy tickets for #psvfey Saturday*

2. *Going to feyenoord-kiev tuesday and to fc utrecht-feyenoord Sunday next week #feykie #utrfey #diehard*

3. *but first the Cup Final next week, still playing for the double :d #ajatwe*

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

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

4. Other types of events

As a direct test of the generalization performance of our method to other events, we tested all 9 methods trained on the soccer events with their best window and step settings on five public events of different types. The best window and step setting per method is specified in Table 5.

4.1 Experimental set-up

4.1.1 Data

We selected five recent events that occurred in the Netherlands that could be identified based on a common hashtag or key phrase. We made sure the events did not contain any big sub-event, which would be the case for a music festival that hosts a number of concerts. Such events pose extra difficulty which has to be dealt with in future research. Our test set contains two popular concerts in the Netherlands in 2013: a concert by Justin Bieber (identified by '#believetour') and a concert by Bruce Springsteen (identified by 'bruce springsteen'). Further, we included the national Queen's day celebration in 2013 ('#koninginnedag'), the national IQ test of 2013 ('nationale iq test', broadcasted on television) and a birthday celebration in 2012 of which the invitation was virally spread on social media, causing rioting and substantial damages ('project x haren'). For some of these events, the tweets were collected based on the most common hasthag, like for the football events, while for others the related tweets were collected based on string matching.

Like the soccer events, we collected the tweets that refer to the events by means of `twiqs.nl`. We specified a six-week search window around the known date of the events, to ensure a link with the TTE labels that ML trains for the soccer events[10]. We removed the retweets from the tweets that we obtained. The tweet counts of the resulting data sets are listed in Table 4[11].

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

Table 4: The selected events and tweet counts after removing retweets

10. Realistically, future references to an event should be recognized any period of time ahead of the event, but three weeks before event time would capture most of the tweets for a lot of events.
11. Again, the tweet IDs for these events and their TTE label are available from `http://cls.ru.nl/~fkunneman/data_tte_estimation.zip`

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

Table 5: MAE for the TTE identification of open domain events after training on 60 soccer events

4.1.2 Approach

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

4.2 Results

Table 5 lists the performance of the methods in terms of the MAE. The results show that methods based on the rule-based features obtain the best performances, with a flawless performance on all events by the majority vote on rules. The ML methods are especially troubled by the 'nationale iq test'. Overall, most methods are quite accurate in their estimations; also the ML based method trained on all features produces estimates with one day of error.

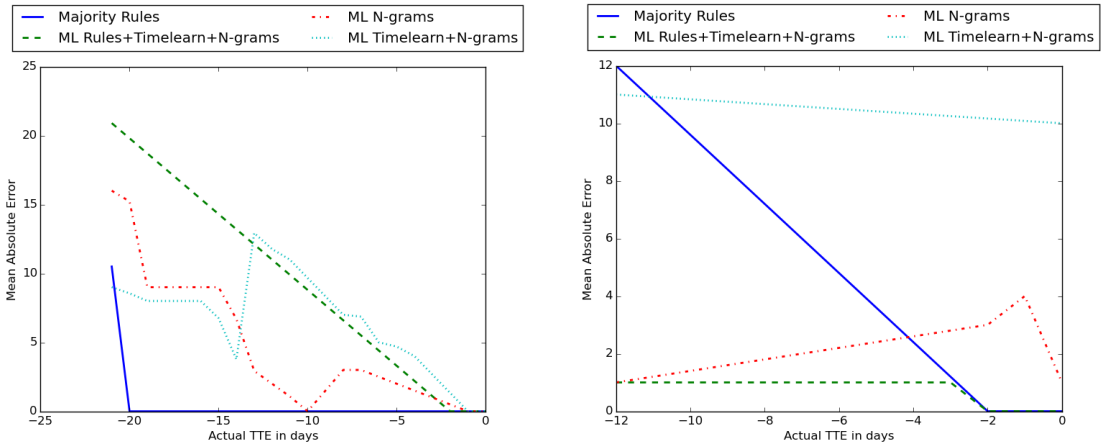
To obtain insights into the influence of event size, we plotted the MAE per actual TTE for the largest and smallest of the five events in terms of tweets posted before event start time: '#koninginnedag' (Figure 2 (a)) and 'nationale iq test' (Figure 2 (b)). As a window of 50 would only allow one estimation for the latter event, we applied four of the methods with a window size of 20 and their best step size as estimated on the soccer events.

Interestingly, the plots show changing performances of the majority vote method and the optimal ML method. Where the former generates flawless estimations from 20 days before '#koninginnedag' on to the beginning of the event, it is subtantially off in its estimation 12 days before the 'nationale iq test'. In contrast, the optimal ML method generates poor estimations for almost all windows when processing tweets referring to #koninginnedag, while its early estimation for the national IQ test is only 1 day off. The latter is an example where ML including word n-grams shows to improve TTE estimation in comparison to a more straightforward majority vote method based on TEXs.

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

5. Conclusion

In this study we experimented with different approaches to estimate the time-to-event of events mentioned on Twitter, in terms of the number of days. We compared machine-learning and majority-voting approaches using different feature combinations and window and step sizes. The methods



(a) #koninginnedag (b) nationale iq test

Figure 3: MAE per actual TTE in days, for a window size of 20 and the optimal step size per method (as scored during training and testing on the 60 soccer events)

were first applied to 60 soccer events, and then to five different types of events with their optimal window and step size. The results demonstrate that a machine-learning approach based on all types of features obtains an accurate performance equal to the majority vote approach based on rule-based TEX estimations, when trained and tested within a set of soccer events. However, the rule-based method obtains the most robust performance throughout different types of events. A qualitative error analysis indicated that time references in tweets sometimes refer to side events, so it is important to complement time features with other features in case of conflicting clues.

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

Although the approach based on majority voting over TEX estimations is often flawless in estimating the number of days to the event, for some points in time (between 13 and 7 days before soccer event start times), we found that machine learning based on all features was more accurate on average. This shows that different methods might complement each other in their estimation. We plan to experiment with an ensemble method in which the decisions of multiple methods are taken into account.

The most important strand of future research is to extend the set of open-domain events and develop an application that applies the optimal method to future events found in the open stream of tweets.

Acknowledgements

This research was made possible by the Dutch national program COMMIT. We thank Erik Tjong Kim Sang for the development and support of the `http://twiqs.nl` service.

References

Becker, Hila, Dan Iter, Mor Naaman, and Luis Gravano (2012), Identifying Content for Planned Events across Social Media Sites, *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, ACM, New York, NY, USA, pp. 533–542. http://doi.acm.org/10.1145/2124295.2124360.

Chakrabarti, Deepayan and Kunal Punera (2011), Event Summarization using Tweets., *ICWSM*.

Chunara, Rumi, Jason R Andrews, John S Brownstein, et al. (2012), Social and News Media Enable Estimation of Epidemiological Patterns early in the 2010 Haitian Cholera Outbreak, *American Journal of Tropical Medicine and Hygiene* **86** (1), pp. 39.

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin, et al. (2003), A Practical Guide to Support Vector Classification.

Hürriyetoğlu, Ali, Florian Kunneman, and Antal van den Bosch (2013), Estimating the Time between Twitter Messages and Future Events, *DIR*, pp. 20–23.

Hürriyetoğlu, Ali, Nelleke Oostdijk, and Antal van den Bosch (2014), Estimating Time to Event from Tweets Using Temporal Expressions, *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Association for Computational Linguistics, Gothenburg, Sweden, pp. 8–16. http://www.aclweb.org/anthology/W14-1302.

Jackoway, Alan, Hanan Samet, and Jagan Sankaranarayanan (2011), Identification of Live News Events using Twitter, *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ACM, pp. 25–32.

James, Gareth and Trevor Hastie (1998), The Error Coding Method and PICTs, *Journal of Computational and Graphical Statistics* **7** (3), pp. 377, JSTOR. http://dx.doi.org/10.2307/1390710.

Kunneman, Florian A and Antal van den Bosch (2012), Leveraging Unscheduled Event Prediction through mining Scheduled Event Tweets, *BNAIC 2012 The 24th Benelux Conference on Artificial Intelligence* p. 147.

Ozdikis, Ozer, Pinar Senkul, and Halit Oguztuzun (2012), Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter, *Proceedings of the 1st International Workshop on Online Social Systems*.

Qin, Yanxia, Yue Zhang, Min Zhang, and Dequan Zheng (2013), Feature-Rich Segment-Based News Event Detection on Twitter, *International Joint Conference on Natural Language Processing*, pp. 302–310.

Quezada, Mauricio and Barbara Poblete (2013), *Understanding Real-World Events via Multimedia Summaries Based on Social Indicators*, Springer-Verlag.

Ritter, Alan, Mausam, Oren Etzioni, and Sam Clark (2012), Open Domain Event Extraction from Twitter, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, ACM, New York, NY, USA, pp. 1104–1112. http://dx.doi.org/10.1145/2339530.2339704.

Sappelli, Maya, Suzan Verberne, and Wessel Kraaij (2013), Combining Textual and Non-Textual Features for E-mail Importance Estimation, *Proceedings of the 25th Benelux Conference on Artificial Intelligence*, Maastricht, The Netherlands, pp. 147–154.

Strötgen, Jannik and Michael Gertz (2010), Heideltime: High quality rule-based extraction and normalization of temporal expressions, *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 321–324.

Tjong Kim Sang, Erik and Antal van den Bosch (2013), Dealing with Big Data: The Case of Twitter, *Computational Linguistics in the Netherlands Journal* **3**, pp. 121–134.

Tops, Hannah, Antal van den Bosch, and Florian Kunneman (2013), Predicting Time-to-event from Twitter Messages, *BNAIC 2013 The 24th Benelux Conference on Artificial Intelligence* pp. 207–2014.

Valkanas, George and Dimitrios Gunopulos (2013), How the Live Web Feels about Events, *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM, pp. 639–648.

Weerkamp, Wouter and Maarten De Rijke (2012), Activity Prediction: A Twitter-based Exploration, *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access, TAIA-2012*.

Weiler, Andreas, Marc H Scholl, Franz Wanner, and Christian Rohrdantz (2013), Event Identification for Local Areas using Social Media Streaming Data, *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*, ACM, pp. 1–6.

Zhou, Xiangmin and Lei Chen (2013), Event Detection over Twitter Social Media Streams, *The VLDB Journal* pp. 1–20, Springer.