Catching Events in the Twitter Stream: A showcase of student projects

Tim Kreutz and Malvina Nissim
Center for Language and Cognition Groningen – Language Technology
Rijksuniversiteit Groningen, The Netherlands
t.j.kreutz@student.rug.nl, m.nissim@rug.nl

Abstract

A group of bachelor students in information science at the University of Groningen applied off-the-shelf tools to the detection of events on Twitter, focusing on Dutch. Systems were built in four socially relevant areas: sports, emergencies, local life, and news. We show that (i) real time event detection is a feasible and suitable way for students to learn and employ data mining and analysis techniques, while building end-to-end potentially useful applications; and (ii) even just using off-the-shelf resources for such applications can yield very promising results.

1. Introduction

The availability of a constant flow of information in the form of short texts makes it in theory possible to collect real time information about virtually all sorts of events that are being written about. The attractiveness of this is evident, and so is its potential social utility. ReDites, an event detection and visualisation system (Osborne et al. 2014), is a prime example of this, as it was developed in order to help information analysts identify security-related events.

However, how the twitter stream can be successfully exploited to this end isn't straightforward, neither conceptually nor practically. First, specific events must be detected, and related tweets clustered. This step must rely on a definition of what an event is, which is often application dependent. Second, the retrieved tweets must be filtered and processed to minimise noise, both in terms of pertinence as well as in terms of the noise typical to the nature of tweets. Third, in order to produce a meaningful tool, the output must be evaluated at a development stage, and made as usable as possible for the end user in its final form, for example providing customisation and visualisation features.

In this short paper, we report a series of efforts within a bachelor programme in information science, where a group of nine students developed different systems with different specific aims, all exploiting real time tweet-derived information, and all socially relevant. The systems work with Dutch tweets, but their architecture is virtually language-independent, as long as tweets and basic language processing tools are available. Suggesting novel methodologies or applications for event detection wasn't our primary concern when developing the systems and when writing this contribution. Instead, by describing this collection of different student projects in the area of real time event detection towards social utility, we have a twofold aim. First, we show that even leveraging off-the-shelf tools and basic preprocessing can yield interesting, promising, and even unexpected results, with a variety of applications. Second, as we have observed that event detection on Twitter has been a useful and suitable task for students, helping them gain familiarity with data mining and data analysis while putting together end-to-end systems, we hope to inspire others to embark on similar exercises.

2. Applications

The University of Groningen extracts Dutch tweets from the Twitter Firehose and provides access to the stream for students and employees (Tjong Kim Sang 2011). Students used the backlog of available tweets to select a specific snapshot or demonstrated their systems using the most recent tweets.

The definition of an event was very application dependent, but always informed by some secondary information, be that peaks in Twitter usage, shared time and location of tweets or overlap with news headlines. Practically and generally speaking, tweets about a single event overlap in some way, and finding sophisticated ways to detect this overlap was a core challenge for all projects.

For none of the developed applications, typical noise in tweets was of any particular concern, although hashtags and URLs were often removed or replaced by placeholders in preprocessing.

Visualization for end users was provided by a few of the projects (Kreutz 2015; de Kleer 2015; Pool 2015), where live demonstrations of their results consisted for example in a website listing relevant tweets per news headline, or maps where categorised events are clustered by location.

The students' systems that we describe showcase four different application areas: sports, emergencies, local events, and news.

2.1. Sports

Sports fans like to stay up to date with real time scores by accessing websites such as `livescore.com`, which provides live overviews of football matches for major Leagues.[1] The overview consists of tables including major events in the game, like goals and yellow/red cards. It takes a lot of time to manually input in-game events, which is why obtaining reliable real-time updates can be expensive and automatising this process leveraging real time Twitter data becomes attractive. Three of the projects were concerned with automatically reproducing such tables by using the stream of Twitter data to automatically detect and classify in-game events. We describe here one of the developed systems, where matches of the Dutch national team in the

[1] Companies such as Livescore.com buy real-time information for prices that vary according to the prestige of the League.

2014 World Cup of soccer are used as a case to demonstrate the approach (Kuiper 2015).

The selected five matches featured a total of 19 goals, 15 yellow cards and no red cards, so only the first two types of events were predicted. For each match, two hours of Dutch Twitter data from the first minute of the match was collected, resulting in a total of 4.376 relevant tweets.

Kuiper (2015) makes use of past sports event and automatic annotation to save on the effort that would go into annotating such a large set of tweets. Firstly, only tweets that contain hashtags referring to a certain match are considered. The convention of such a hashtag is using the first three characters of the involved teams (#SpaNed for Spain versus the Netherlands). Secondly, the timestamp of each tweet is compared to the timeline of the actual match. As such, all tweets posted up to three minutes after an actual event took place will be annotated as being about the event. In the training data, distribution of the classes no event, goal and yellow card were respectively 55%, 37% and 8%.

The data was then modified to allow for a more general application of the system. This involved replacing specific scores and players with placeholders. Occurrence of specific keywords that denote an event were used as features, along with the length of the tweet and the tf-idf term vector.

Beyond detecting events in single tweets, detecting events has to do with grouping relevant tweets in the right way. Detection of peaks in Twitter activity has been used to detect events (Corney, Martin, and Goker 2014; Van Oorschot, Van Erp, and Dijkshoorn 2012). Specifically, Chakrabarti and Punera (2011) demonstrate how tweet volume signifies important events in sport matches. However, in soccer it is more likely that two important events occur in close proximity which is problematic for peak detection, since the events will be grouped as one event. To more accurately distinguish between events, Kuiper (2015) implements a rule-based system that looks at tweet content during peeks. If at least fifteen% of tweets are classified as goal-tweets, the rule based system determines whether the mentioned score is logically probable (a match with score 1-1 logically progresses to either 1-2 or 2-1) and updates the score. This way, the score is updated before a potential second goal, allowing consecutive score updates to be detected.

Using the Multinomial Naive Bayes implementation in Scikit learn (Pedregosa et al. 2011), with the set of features in Table 1, classification of individual tweets yields an f-score of .843. For the matches of the Dutch national team, sixteen out of seventeen goals were detected in the right minute (f-score .940), but detection of yellow cards was harder (f-score .500).

Beyond achieving good results for a case of very specific matches in a very specific sport, Kuiper (2015) demonstrates the feasibility of automatic sub event detection in sports in general, specifically with regards to grouping and distinction of isolated events. However, since no two sports are the same, consideration has to be made of the volatility of events. Detecting subevents in the stream of twitter data may be a lot harder for the faster-paced basketball for example. More importantly, a substantial amount of tweets that discuss specific events are needed to reach the accuracies presented in Kuiper (2015).

Table 1: Features for Kuiper (2015)'s sports event detection system.

Feature	Type	Reason
TF-IDF	Float	Indicate the importance of a word in relation to the document in which the word has been found.
Length	Integer	Tweets posted after important events tend to be shorter.
Goal	Boolean (1 or 0)	Tweets containg the word "goal" or "scoort" or "doelpunt" could very well be tweets describing a goal.
Score	Boolean (1 or 0)	Tweets containg scores such as "2-1" or "3-2" could very well be tweets describing a goal if they are not predictions.
Predict	Boolean (1 or 0)	Tweets containg words such as "gok" or "prono" could very well be predictions instead of goals.
Yellow	Boolean (1 or 0)	Tweets containg words such as "geel or "gele" could very well be indicators of a given yellow card.

2.2. Emergencies

Twitter allows for detection of real-time sub-events in sports because relevant tweets follow these events almost instantly. The delay between a real-time occurrence and its social resonance are thus minimal. This adds to the social relevance of detection of events that are particularly time-sensitive, such as emergency situations. This section will look at two different emergency scenarios: earthquakes in the Dutch province of Groningen, and detection of context for events reported by Dutch emergency services.

2.2.1. Earthquakes

The detection of earthquakes on Twitter has been extensively documented (Sakaki, Okazaki, and Matsuo 2010) for Japan, where the tweet density is high and earthquakes occur relatively frequently. The research focuses on detection of earthquakes and extraction of the time that it occurred, along with the location. Earthquakes with a magnitude of 3.0 or higher on the Richter scale were successfully detected in 96% of the cases, and real-time detection led to notifying civilians faster than the Japan Meteorological Agency could, in most cases.

Detection of earthquakes in Groningen has only recently become relevant since gas extraction in the province led to a 200% increase in earthquakes over the past ten years (Kuipers 2015). This has lately sparked debate in politics and media and increased public involvement. Detection of earthquakes using Twitter can thus contribute to timely updates, but it may also map public sentiments.[2]

To develop his system for detecting earthquake events via Twitter, Kuipers (2015) used data from the Dutch Meteorological Institute (KNMI) from January 2014 until April 2015. The data contained 60 earthquakes with a magnitude of 1.2 or higher on the Richter scale, their timestamp and location of the epicenter. Weaker earthquakes are generally considered intangible for humans, and hence not useful for

[2]In the context of earthquakes in the Groningen area, this is interesting also in the context of NAM's compensation duties for earthquake-caused damage to local properties.

the research as there would be no tweets about them. A preselection of Twitter data was made by selecting tweets containing the words 'beving' or 'aardschok' that were tweeted up to four hours after the occurrence of an earthquake.

Using Weka (Hall et al. 2009), a Naive Bayes classifier was trained on the annotated tweets and tested via cross-validation. Results show that tweets are correctly classified as relevant or irrelevant to a given earthquake in over 91% of the cases. Among the most distinguishing features are the mention of a location in the Groningen or Drenthe province (boolean) which usually signals an actual earthquake, and the mention of political terms (boolean) which usually signals no actual earthquake. Further features used as potential indicators of relevant tweets are mentions of numbers, which can signal a specific time or magnitude, and certain signal words that are used to signal the sensation of experiencing an earthquake ('voel', 'tril', 'knal').

2.2.2. Emergency services

P2000[3] is a live repository of all emergency services active in a given area. All reports are publicly available and real-time updated communications of and between Dutch police, ambulance and fire department services are available. You can think of the first reports of a fire, the cars inbound to the location of the reports and the way the distress is handled.

The work described in (Louwaars 2015) is concerned with matching user tweets to reports from emergency services in the same area. The rationale behind this is that such matches could be used to diminish delay in notifying stakeholders, or adding context to official, quantitative reports. The real-time nature of Twitter makes it particularly suitable for detecting time-sensitive events like emergencies.

One month of emergency reports and tweets were downloaded from the P2000 website in April 2015 for the larger Groningen area. This resulted in 700 'matches' of reports to one or more tweets with a similar location and time. A Naive Bayes model was trained on 80% of the annotated data, using simply word occurrences as features, to classify tweets as 'relevant' or 'irrelevant'. Testing on the remainder 20% resulted in a significant improvement over the baseline (75% of tweets were annotated as irrelevant) with a final accuracy of 91%. The limited amount of data, and the skewedness between relevant and irrelevant tweets does not make this result generalizable to a real-time application for detecting emergency situations, but successful matches do add some context to otherwise formal reports.

The approach in which words are used to predict relevancy allows for an overview of the most indicative words (Table 2).

A critical reflection on these results can be that Louwaars (2015) observes that for some events the amount of tweets is too low to draw any solid conclusion. He further indicates that few of the relevant tweets comes from 'real' twitter users, with substantial data coming from automated emergency service accounts. This is also reflected in the list of most indicative words, which for a large part contains random tokens (apart from 'Brandweer'). As a solution, the system could be trained on emergencies that can draw from a larger pool of tweets, indicating more severe cases or emergencies that occur in more densely populated locations.

[3] http://www.p2000-online.net/groningenf.html

Table 2: Most indicative words for relevancy.

Token	Proportion (relevant to irrelevant)
Gaat	13.8 : 1
1	11.3 : 1
Weer	1 : 8.2
@	1: 7.8
!	1 : 7.5
Niet	1: 7.0
Brandweer	6.9 : 1
/	6.9 : 1
Maar	1 : 6.4
(	6.0 : 1

2.3. Local events

The meta-data attached to tweets can be useful for certain instances of event detection. Pool (2015) and de Kleer (2015) show that using the relatively low frequency of geo-tagged tweets, it is possible to cluster various sorts of events on the local scene, classify them and map where they occur in real-time.

Detecting events using geo-locations from Twitter has previously been done by Walther and Kaisser (2013) and applying a similar approach to Dutch tweets is plausible because the Netherlands has one of the highest twitter accounts to population ratio (Pool 2015).

All geo-tagged tweets from a month of Dutch Twitter data were used for training. This resulted in a total of 566.549 geo-tagged tweets. The geo-information was translated into a geoHash that denotes a specific area, and tweets with a similar geoHash and comparable timestamp were grouped and added to a list of event candidates. To handle the hard borders of the geohash area, candidates with matching timestamps in adjacent areas were then merged (Figure 1).

Figure 1: A border case in (de Kleer 2015)

Two judges annotated the event candidates in the training data and the test data with the following labels: *No event*, *Meeting*, *Entertainment*, *Incident*, *Sport* and *Other*, with the most frequent category being *No event* (triggering a 46% baseline). Inter-annotator agreement was measured via Cohen's Kappa (Cohen 1960). Features that were found

Figure 2: Visualisation of automatically classified events in The Netherlands, in April 2015 (de Kleer 2015).

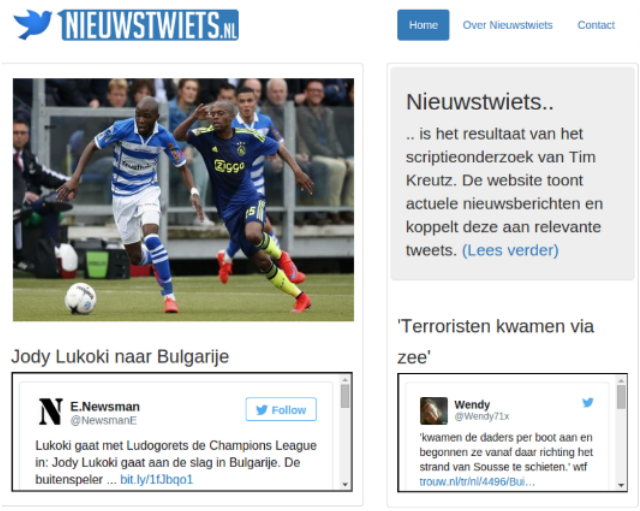

Figure 3: Visualisation of relevant tweet commentary on news events on `nieuwstwiets.nl` (Kreutz 2015).

to be most useful for this task were the most frequent words, the location using only the first five characters of the geoHash, the average word overlap between tweets in an event candidate and the average word overlap between different users in an event candidate.

Several models were built and tested on development data, with the final system being a Naive Bayes model which yielded an accuracy of 84% on test data. Especially considering the inter annotator agreement was measured at $K = 0.87$, this is a very good result. This research also shows that meta-information from tweets can successfully be used to detect events. The end-user output of the system is a map with classified events (Figure 2).

2.4. News

Twitter data has also been used to detect real time commentary on news events. News media websites often feature their own social plugins which allow readers to discuss news items. These discussions are relatively structured and easy to relate to the news article. However, when people post their commentary to Twitter, it becomes problematic to link the tweet back to the article and to group all relevant discussion together.

In (Kreutz 2015), RSS feeds of the three most popular Dutch news websites are used to detect similar content on Twitter. The RSS feeds give access to 41 headlines and related abstracts at the same time. Each of the news items is then compared to the last hour of Dutch Twitter data to extract reaction, opinions and other meta-commentary that users posted.

To deal with the computational effort involved in making this many comparisons (an hour of Dutch twitter data often contains more than 30,000 tweets), a first module of the system makes a pre-selection of candidates to be considered. The candidates are made up of the 25 tweets with the highest cosine similarity compared to the title of the news items. Generally, these 25 tweets contain tweets that are too similar to the title (a retweet for example), too dissimilar (are not about the article) and actual relevant tweets that add meta-commentary. It is the latter type that one would want to detect and use, while discarding the former two as non-relevant.

To distinguish between relevant and non-relevant tweets, four non linear machine learning algorithms were trained and tested. Four news articles from the 21st of May 2015 were selected to be compared to 24 hours of Dutch Twitter data from the same day. For training and testing, 250 candidate tweets were selected using the approach mentioned above. This cut off point was chosen because after annotation it became clear that for each of the articles, the number of relevant tweets that could be found after rank 250 was negligible.

In the 1,000 candidates, 593 were annotated as relevant to the articles which results in a baseline of 59.3%. The system was trained on six features: (1) the difference in timestamp between the publication of the article and the publication of the tweet, (2) the cosine similarity between the title of the article and the tweet, (3) the difference in length between the title of the article and the tweet, (4) the cosine similarity between the abstract of the article and the tweet, (5) the amount over overlapping Named Entities, (6) the cosine similarity between bigrams in the abstract of the article and the tweet.

Named entities were extracted by means of a Named Entity Recognizer trained on the CoNLL2002 Dutch corpus using NLTK (Bird, Klein, and Loper 2009). A Random Forest classifier performed the best on the test data with an F-score of 0.874. It also helped to determine the most important features in the task. The list of features mentioned above adheres to this order, timestamp difference being the most predictive feature.

The selection of viable candidates before automatic classification proves successful in reducing computational effort, while still keeping the detection of relevant commentary possible. This approach is therefore suitable for a real-time application of the system. demonstrates this by apply-

ing the system to a specifically dedicated website that updates its news articles and tweets hourly (`http://www.nieuwstwiets.nl`, Figure 3). Rather than representing the results of the research with an F-score, it provides insight by showing users the resulting tweets.

3. Discussion and Conclusions

In this overview we have reported efforts of bachelor students in the field of automatic event detection exploiting the (Dutch) Twitter stream.

Besides the differences in fields of application, this overview gives insight in the considerations that were made in dealing with the inherent challenges in event detection. For the emergency detection and the detection of local events, geo-information of tweets was used. Since this information is not always available, this sometimes resulted in very little data to work with. For the detection of subevents in soccer, peaks of tweets with certain key patterns were used. This worked well for important subevents (goals) and worse for minor subevents (yellow cards).

The students used similar ways to remove noise from tweets, by removing or replacing URLs and hashtags. Even when hashtags were crucially used to detect events, such as in (Kuiper 2015), they were then normalised at a second stage in order to make the approach general and portable. The local event detection and news event detection sought to extract as much possible information from tweets by normalizing the hashtags. In the data selection for the soccer events, hashtags had a leading role in selecting tweets only when they contained a hashtag that referred to a certain match.

Finally, evaluation showed good results in all the theses. Some students chose to apply their findings in a real time setting by visualizing them. This lead to demonstration of the systems by Pool (2015) and de Kleer (2015) on EventDetective and Kreutz (2015) on `nieuwstwiets.nl`.

With this exercise we observed that real time event detection on social media is a field that students can successfully experiment with. Although the aim was not to build the next generation event detection applications, the choices that the students made in the course of such a research reflect some of the core challenges and considerations central to this task and we believe are useful lessons for future endeavors, both from a research and a teaching perspective. It also reflects that there are no ready-made best practices when it comes to defining events and selecting data, and that each socially useful task will have its own needs and therefore strategies.

References

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media.

Chakrabarti, Deepayan and Kunal Punera (2011). "Event Summarization Using Tweets." In: *ICWSM* 11, pp. 66–73.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, p. 37.

Corney, David, Carlos Martin, and Ayse Goker (2014). "Spot the Ball: Detecting Sports Events on Twitter". In: *Proceedings of Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands*. Ed. by Maarten et al. Rijke. Springer, pp. 449–454.

de Kleer, David (2015). "EventDetective: detectie, verrijking en visualisatie van Twitter events". Bachelor Thesis in Information Science. University of Groningen.

Hall, Mark et al. (2009). "The WEKA Data Mining Software: An Update". In: *SIGKDD Explor. Newsl.* 11.1, pp. 10–18. ISSN: 1931-0145.

Kreutz, Tim (2015). "Detecting news event commentary on Twitter". Bachelor Thesis in Information Science. University of Groningen.

Kuiper, Jacco (2015). "Real-time automatic detection of soccer match events using Twitter". Bachelor Thesis in Information Science. University of Groningen.

Kuipers, Rolf (2015). "'En we schudden weer'". Bachelor Thesis in Information Science. University of Groningen.

Louwaars, Olivier (2015). "P2000 locatiedata als classifier voor tweets". Bachelor Thesis in Information Science. University of Groningen.

Osborne, Miles et al. (2014). "Real-time detection, tracking, and monitoring of automatically discovered events in social media". In: *Proceedings of ACL 2014: System Demostrations*. Association for Computational Linguistics, pp. 37–42.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12, pp. 2825–2830. ISSN: 1532-4435.

Pool, Chris (2015). "Detecting local events in the Twitter stream". Bachelor Thesis in Information Science. University of Groningen.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of WWW '10*. New York, NY, USA: ACM, pp. 851–860.

Tjong Kim Sang, Erik (2011). "Het gebruik van twitter voor taalkundig onderzoek". In: *TABU: Bulletin voor Taalwetenschap* 39.1/2, pp. 62–72.

Van Oorschot, Guido, Marieke Van Erp, and Chris Dijkshoorn (2012). "Automatic extraction of soccer game events from Twitter". In: *Proceedings of Detection, Representation and Exploitation of Events in the Semantic Web (DeRiVE 2012). Boston, MA, USA. (CEUR Proceedings, 902)*.

Walther, Maximilian and Michael Kaisser (2013). "Geospatial event detection in the twitter stream". In: *Advances in Information Retrieval*. Springer, pp. 356–367.