Towards Open Domain Event Extraction from Twitter: REVEALing Entity Relations

G. Katsios[1], S. Vakulenko[2], A. Krithara[1], G. Paliouras[1]

[1] Institute of Informatics and Telecommunications, NCSR Demokritos, Greece
[2] MODUL University Vienna, Austria

Abstract. In the past years social media services received content contributions from millions of users, making them a fruitful source for data analysis. In this paper we present a novel approach for mining Twitter data in order to extract factual information concerning trending events. Our approach is based on relation extraction between named entities, such as people, organizations and locations. The experiments and the obtained results suggest that relation extraction can help in extracting events in social media, when combined with pre and post-processing steps.

Keywords: Event extraction, social media analysis, relation extraction, Twitter

1 Introduction

Social media attracts millions of users, and has evolved to become a source of various kinds of information. In Twitter for example, more than 255 million active users publish over 500 million 140-character "tweets" every day[3]. Evidently it has become an important communication medium. More and more people use social media to communicate their ideas and thoughts, as well as to spread important news. Given the enormous size of information exchange happening every day, it is a rather challenging task to process these data and filter out the important and relevant information.

Twitter data is part of the Big Data paradigm and is characterized by high Velocity, Veracity and Volume ("the 3 Vs") [12]. The topics on Twitter span across multiple domains from private issues to important public events in the society. Therefore, filtering out the important or relevant to the user information poses the first challenge for automated processing of tweets.

Twitter provides user-generated content in real time. The data is stored in a form of short text messages called tweets. Each tweet has a body that contains text of the message itself, but also a variety of metadata associated with it, e.g. date of creation, author, user mentions, location etc. However, what makes Twitter texts unique is its word count limitation which causes extensive usage

[3] `https://about.twitter.com/company`

of acronyms and other abbreviations. Moreover, users often use colloquial words and phrases in tweets, which require context for interpretation.

The goal of this research is to develop tools that extract and efficiently summarize trending events, the so-called “breaking news“, mined from social media, e.g. Twitter. This task is especially relevant for the professional journalists helping them to utilize social media as an information source helping to cope with the information overload.

This research was conducted in the context of two European 7th Framework projects, REVEAL and DecarboNet. The projects aim at developing new tools and approaches to automatically process digital media content, extracting important information and summarizing it.

This paper is reporting on the results of the initial round of experiments, where we combined the current state-of-the-art methods and tools available, and further evaluated them for the task of event extraction from social media. We also enhanced the pipeline with pre- and post-processing procedures in order to adopt it to the specific requirements stemming from the nature of social media data, e.g. spam detection, mention disambiguation and relation selection. These initial investigation and prototyping results aim to reveal the pitfalls and shortcomings of the current state-of-the-art approaches and suggest directions for the future work.

The definition of an event itself might appear rather blurry and controversial from the first sight. We adopt the wide definition of an 'event', which goes beyond scheduled events, like a music concert, conference or a football match. In general, we consider any action, which can be observed in the physical world, to constitute an event [21].

Events are often communicated through social media, e.g. *"Chelsea won today"*, *"We are going to a bar"*. Due to the abundance of such event reports on social media we define the notion of an 'important event', i.e. an event, information about which is of a potential value to a user of the system. For example, information about an international political summit involving famous politicians may be considered as important for the journalist, while the content of a lunch meal of an average twitter user is likely to be of no particular value.

In this work we focus on extracting the factual information about an event, e.g. its location, time and participants. It is important to separate the factual information from the content that expresses an opinion or an emotion related to the event, such as feelings and thoughts of an individual or a group. This can be a rather tricky task, because sentences that are lexically very similar can convey semantically opposite facts. For example: *"Chelsea won today"* versus *"I wish Chelsea won today"* versus *"I wish Chelsea wins today"*.

In order to extract event-related information from tweets we adopt and enhance existing state-of-the-art approaches to automated information extraction, taking into account the unique properties of social media data. We implement and apply the proposed approach to several datasets, evaluate and discuss the results, outlining further directions for the future work.

2 Related Work

Existing algorithms for news monitoring typically detect events by grouping together words with similar burst patterns (i.e. words or phrases showing burst in appearance count [24]). They rely on clustering or topic modeling techniques [3, 10, 13]. The draw-back of these approaches is that the resulting bag-of-words representation of the clusters/topics is often not descriptive enough.

More sophisticated and precise approach is information extraction on the level of events. Event extraction involves parsing of natural language text with the aim of extracting event-related information. The usual suspects for the event facets are the named entities that belong to actor/place/time classes in Simple Event Model (SEM) [21]. Therefore, many approaches to event extraction include entity recognition stage [5, 19]. In our work we also utilize the assumption that many events are centered around named entities as in [19]. Still open remains the question of how to connect the event-related entities, e.g. persons, locations, dates. Most of the approaches use NLP-methods involving a set of regular expressions to extract verbs that are assumed to constitute an event and feed it together with the related entities into the event model [1, 8, 18, 19, 22].

On the contrary, in our approach we utilize the state-of-the-art method for relation extraction [6], that has already been successfully applied to news articles. Relation extraction is the task of identifying relations that hold between entities in text data. Up to now relation extraction systems were only evaluated on news collections, but not on social media data. Therefore, the novelty of the proposed approach is testing the suitability of relation extraction methods for the task of event extraction on Twitter. We also make several modifications in order to adapt the relation extraction approach to the specific nature of social media data and further enhance it to extract event-related relations between the frequent named entities from tweets.

There have been a number of projects aiming at extracting events specifically from tweets [5, 20, 23]. Tweets are specific in nature and require special treatment, different from the news articles. Therefore, Twitter-oriented systems often include methods to detect spam, reduce noise and eliminate uninformative messages [5, 20].

Domain-specific event extraction, such as [5, 23], allow fine-tuned event detection, but require a set of keywords or event types to be manually predefined. In this work we focus on extracting trending events, i.e. events which are most popular among the users and are most frequently discussed. This approach also allows us to be domain-agnostic and catch previously unknown events.

In this respect, our approach is most similar to TwiCal [20]. However, instead of training classifier for event extraction on in-domain training data we utilize already trained extractor from ClausIE [6]. The goal of TwiCal is constructing a calendar of upcoming events. Therefore, it extracts only scheduled events accompanied with explicit date mention. We are primarily interested in information concerning recent or current events, where explicit date annotation is often omitted.

3 Our Approach

We adopt the state-of-the-art approach to relation extraction [6] and further enhance it for the task of event extraction from tweets. In our approach we consider any action, which can be observed in the physical world, to constitute an event [21]. We assume that events are indicated by nonstative (dynamic) verbs. Dynamic verbs describe an action, such as 'kick', 'meet', 'visit', as opposed to stative verbs, such as 'believe', 'like', 'consider', etc.

Relation extraction approach enables us to extract predicates from a sentence (corresponding to the verbs indicating events) together with their subjects and objects. For example, the sentence: *"The match starts on Sunday"* will result in the following relation: *The match (Subject) - starts (Predicate) - on Sunday (Object).*

Objects of the relations often contain event facets that uniquely characterize events in spatial, temporal and social dimensions (e.g. place, date, organizers, participants). Thus, this approach allows for more fine-grained event extraction as opposed to clustering or topic modeling-based approaches which operate with the bag-of-words model, which tend to blend together several lexically similar events.

We have extended the initial approach to relation extraction with a few pre-processing steps in order to clean the input data and annotate it with named entities. After the pre-processing we extract relations, link them to named entities and rank according to their frequencies. The resulting pipeline summarizing our approach is presented in Figure 1. In the rest of this section, the different modules of our approach are described in details.

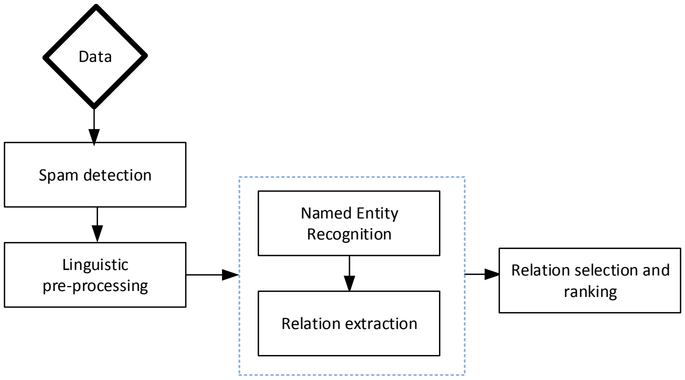

Fig. 1: System's pipeline

3.1 Spam detection

Here we define spam as useless uniformative or malformated messages, which are unlikely to provide us with any meaningful information. Our goal is to preprocess the raw data from Twitter and deliver to the end user only useful and relevant information. Therefore, we attempt to filter out meaningless and misleading messages already on the first stage of our pipeline.

In the first place, we use a freely distributed black-list of domain names [4] in order to exclude tweets containing links that point to the untrusted web sites. Next, we calculate a "spam score" for each of the remaining tweets and exclude the tweets that receive the score higher than the empirically learned threshold value. The **"spam score"** is calculated as the number of spam-associated tokens [4, 16] divided by the total number of tokens in the tweet:

$$spam\_score = \frac{|U| + |H| + |L| + |S| + |N|}{|T|} \tag{1}$$

where:

- $|U|$: number of user mentions (e.g. @themichaelowen);
- $|H|$: number of hashtags (e.g. #DavidGill);
- $|L|$: number of web links (e.g. `http://t.co/my55ZOoAko`);
- $|S|$: number of spam words (from the predefined list[5], e.g. dutyfree, poker, casino);
- $|N|$: number of non-word characters (e.g. %, !);
- $|T|$: total number of tokens in the tweet.

The bigger the value of the "spam score", the more likely that the tweet contains spam. We conducted an experiment spanning numerous trials to choose the optimal threshold value for the spam score and arrived at the value of 0.74. Further one, we identified 3% of the tweets in our datasets as spam and, therefore, excluded them from the next stages in our pipeline.

3.2 Linguistic Pre-processing

All the tweets that passed through the Spam Detection module, are further considered in the Linguistic Pre-processing module. The pre-processing steps include tokenization, user mentions resolution, further text cleaning and sentence splitting.

Tokenization is used to identify the tokens that will be replaced or removed from the text, such as URLs, user mentions, etc. First, we exploit tweet metadata to resolve user mentions to their canonical names. In particular, each tweet that contains user mentions, carries a list of the corresponding full user names from the Twitter database. Thus, we substitute the user mentions in the tweet text with the corresponding full names using the tweet metadata. For example, *@themichaelowen* is resolved to *Michael Owen*.

[4] `http://www.squidguard.org/blacklists.html`

[5] `http://notagrouch.com/wp-content/uploads/2009/12/wordpress-blacklist-words.txt`

3.3 Named Entity Recognition

In this module, we identify named entities mentioned in the text of the tweet, as well as their types. For example, the tweet containing the following snippet: "@DavidGill walks out of FIFA meeting in Sao Paulo", gets annotated with the named entities: *David Gill - Person*, *FIFA - Organization* and *Sao Paulo - Location.*

We used Stanford Named Entity Recognizer (Stanford NER) [15] for detecting named entities in tweets. According to the benchmark evaluation reported in [7], Stanford NER achieves highest average precision on all three datasets of tweets, when compared with other state-of-the-art Twitter-tailored algorithms.

Stanford NER detects the following types of named entities: Location, Person, Organization, Date, Money, etc.[6]. Due to our pre-processing procedure we also detect the entities "hidden" within the user mentions and hashtags (e.g. *@DavidGill*). This would not be feasible, when applying the Stanford NER on the original tweets.

3.4 Relation Extraction

The core of our approach is based on extracting relations from the pre-processed tweets. Relation is a triple that consist of subject, predicate and object. Subject and object are entities, predicate is the relation between these entities. For example, the sentence: *"The match starts on Sunday"* will result in the following relation: *The match (Subject) - starts (Predicate) - on Sunday (Object).*

We considered three state-of-the-art systems for the task of relation extraction: ReVerb [9], Ollie [14] and ClausIE [6]. ClausIE was reported to significantly outperform Ollie by the number of propositions extracted [6]. However, it has not been previously applied to social media data. Therefore, we ran our own experiments to compare the results returned by ReVerb and ClausIE. Subsequently, we chose ClausIE as the best-suited baseline system.

In ClausIE relation triples are extracted from clauses, parts of a sentence that express coherent pieces of information [6]. The clauses are identified based on the results from the dependency parser that helps to reveal the syntactic structure of an input sentence. In particular, ClausIE is using Stanford unlexicalized dependency parser [11].

Additionally, ClausIE has an option to return n-ary predicate by decomposing the object of the relation into several arguments. This option can be useful for extracting complex relations, that consist of several independent but overlapping parts, such as place and time relations. For example, the sentence: "The match starts on Sunday at Wembley" will result in the following relation: *The match (Subject) - starts (Predicate) - "on Sunday", "at Wembley" (Object).*

We made several modifications to the original implementation of ClausIE in order to adapt it to the task of extracting the relations describing events. Specifically, we enforce omitting the following types of clauses from the relation extraction process:

[6] `http://nlp.stanford.edu/software/CRF-NER.shtml`

- conditional clauses (If-clauses), e.g. “If @Chelsea wins I will celebrate till morning!!!!!!!!”
- clauses rooted in a stative verb, e.g. ”I believe @Chelsea is the actual winner!”

Conditional clauses are used to speculate about what might happen, what could have happened, and what we wish to happen. Stative verbs describe mental state of an agent, but do not signify any action. For example, the following verbs are stative: hate, love, believe, prefer, want, suppose, etc.

3.5 Relation Selection

We designed a post-processing step for selecting relations that will appear in the final results. For this we chose the Frequent Pattern Mining approach that helps us to reveal the recurrent information patterns following the assumption that input data from Twitter is often abundant and redundant. Additionally, we employ the following heuristic technique: for the relation to be selected it has to contain popular (frequently occuring) named entities. In this way we get rid of the trivial resuls, e.g. *”I - ate pizza - for breakfast”*, but retain the relations such as: *”President Obama - ate pizza - for breakfast”*, if they are reported by a considerable number of tweets.

Therefore, we combine the results from Relation Extraction (RE) and Named Entity Recognition (NER) modules produced on the previous stages. In particular, we select only those relations that contain named entities in subject and/or object of the relation. The intuition behind this approach enriching relations with NER annotations is that events in real-life are often associated with the corresponding named entities: dates, places and participants.

Hints about importance of the relations and named entities are given from their frequencies count. We assume that widely discussed news are more likely to be of importance and interest to the users of our system. Therefore, in order to link NER and RE results we identify frequent named entities and then select frequent relations, in which these entities occur. We use several approaches to select relations between the named entities described below.

Firstly, we detect the named entities that occur most frequently in the tweets (~ 10 entities for each of the datasets), e.g. Chelsea, Drogba, Ramires. We also identify the most frequently co-occurring pairs of named entities (~ 5 pairs per dataset), e.g. Chelsea and Liverpool, Putin and Ukraine. Then, we identify the following relations that hold between named entities:

1. Relations in which the most frequently occurring entities appear in subject or object of the relation;
2. Relations that hold between pairs of the most frequently co-occurring entities;
3. Relations for every combination of entity types pairs from the set: [Person, Organization, Location, Date], e.g. between Person and Organization, Person and Person, Location and Organization, Person and Date etc.

Finally, we calculate the support for each of the selected relations, i.e. number of tweets from which the same relation was extracted, and use it for ranking of the relations. The topmost relations are reported in the final results.

4 Experimental Evaluation

4.1 Datasets

We conducted experiments using three different Twitter datasets (see Table 1). All datasets are centered around one or several major events discussed on social media. We have deliberately selected the datasets containing event-related tweets for our evaluation with the goal to uncover the details surrounding these events using our approach.

The FACup dataset was created within the Social Sensor project[7] and covers the events during the last match of the Football Association Challenge Cup [2]. The SNOW dataset [17] is an attempt to capture the footprint in the social media regarding several important international events: uprising in Ukraine (#ukraine, #euromaidan), protests in Venezuela (#Venezuela), major Bitcoin exchange theft (#bitcoin), etc. The third dataset was collected in June 2014 and contains ∼ 270.000 tweets, that were extracted using the hashtag #WorldCup2014.

Dataset	**# Tweets**	**Hashtags**
FA Cup	∼ 20.000	#FACupFinal
SNOW	∼ 1.000.000	#ukraine, #euromaidan, #Venezuela, #bitcoin
World Cup	∼ 270.000	#WorldCup2014

Table 1: Datasets

4.2 Evaluation Method

We manually evaluated the results by annotating the relations returned on the last stage of our pipeline (section 3.5). Each of the annotators (3 in total) independently considered perceived correctness and usefulness (importance) of the relations by looking up the original text of a sample tweet, from which the relation was extracted by the system.

The relation was marked as *Correct*, if the information it provides naturally follows from the original text of the tweet and does not contradict the message conveyed in it. Negation handling is a good example for potential errors in the results returned by the system. If the original tweet reports, that Chelsea did not play better than Liverpool, the relation has to communicate the same fact and not the opposite. For example, *Chelsea - play better - than Liverpool* relation should be marked as *Incorrect* in this case.

Furthermore, all correct relations were further evaluated with respect to perceived importance for the end user of the system. The importance of a relation is harder to evaluate than its correctness, because of the complexity and subjectivity in the notion of importance with respect to an information piece. In general, a relation is considered *Important*, when it is perceived as being descriptive and

[7] `http://www.socialsensor.eu/`

potentially useful. Meaningless and uninformative relations are marked as *Not important*, respectively.

Collective discussion of the individual annotations resulted in a consensus and a single final evaluation table was constructed. Afterwards, we summarized our evaluation results by counting the number of relations for each of the classes: *Correct & Important*, *Correct & Not important* and *Incorrect* relations (see Table 2). We calculated the ratios and the total number of evaluated relations separately for each of the datasets. The last row of the evaluation table highlights the average precision values across the three datasets.

Dataset	Incorrect	Correct	
		Not important	Important
FA Cup	0.17 (8)	0.17 (8)	0.66 (32)
SNOW	0.1 (21)	0.14 (32)	0.76 (168)
World Cup	0.1 (18)	0.19 (35)	0.71 (134)
Average	**0.12 (47)**	**0.17 (75)**	**0.71 (334)**

Table 2: Precision of the evaluation results: fraction (total) of relations

4.3 Discussion and Future Directions

The average precision of our approach was estimated at 88% taking into account all correctly extracted relations. However, less that 3/4 of the relations returned by the system were considered as potentially valuable for the end users of the system (see *Correct & Important* in Table 2).

The most frequent relations that were selected using our approach from FA Cup dataset are listed in Table 3. These 5 relations provide a short summary of the event by revealing the names of the teams, the place where the game took place, the winner and the final score, as well as the player, who scored. The timestamps of the tweets can disambiguate the mentions "now", reveal date of the event and indicate the "hot spots" on the game timeline, such as the last relation in Table 3.

Relations extracted from the SNOW dataset are less homogeneous containing various political statements, business and sport announcements, as well as snapshots of historical events. Sample relations (with their support): *Ukraine's leaders - warn - "of Crimea separatism threat"* (106); *Chelsea fans - attending - "the Galatasaray match", "on 26 Feb"* (84).

World Cup dataset is another noisy collection containing many tweets not related to the football championship. Nevertheless, the three top-most relations reveal the major conflict in the football association: *director David Gill - walks out - "of FIFA meeting in Sao Paulo"* (902); *director David Gill - says - "Sepp Blatter should stand down"* (901); *FA Vice-Chairman David Gill - calls on - "Sepp Blatter not to stand for re-election as FIFA President"* (481).

In general, due to our broad definition of 'event' (as any kind of action reflected in a physical world) relations can be extracted from virtually any collection of tweets. However, in order to achieve comprehensive results the tweets need to be previously clustered according to the common topic, e.g. using a set of hashtags.

Subject	Predicate	Object	Count	Sample tweet
The Chelsea players	are throwing	”Robbie Di Matteo high in the air”	129	RT @chelseafc: What celebrations! The Chelsea players are throwing Robbie Di Matteo high in the air. And catching ...
Chelsea	have won	”17 major trophies”, ”now”	58	RT @chelseafc: Chelsea have now won 17 major trophies. We've caught Tottenham who are on the same total.
Liverpool	are out	”for the second half”	27	RT @chelseafc: Liverpool are out for the second half, and Chelsea are on the way. #CFCWembley #FACupFinal (SL)
Chelsea	beat	”Liverpool 2-1 to win the FA Cup at Wembley”	24	RT @premierleague: Chelsea beat Liverpool 2-1 to win the FA Cup at Wembley, their fourth win in six years in the competition. #cfc #lfc ...
Liverpool	is	”much”, ”pretty”, ”giving every Chelsea fan a heart attack right now”	21	RT @espn: Liverpool is pretty much giving every Chelsea fan a heart attack right now: http://t.co/MGxAkv94

Table 3: Results from FA Cup dataset

We performed only limited experimental evaluation for the proof-of-concept of our approach and can not quantatively compare our results with other approaches to event extraction. Moreover, the relation extraction algorithm is currently computationally rather expensive, which might prevent us from running the system on Twitter stream data in real time.

Nevertheless, our initial results provide further motivation and help to outline directions for the future work:

1. Linking relations that convey the same information. Disambiguating and clustering these relations will help to improve quality of the results by increasing support of the frequent relations and removing semantic duplicates. This can be achieved by:
   – grouping the predicates into semantic groups using existing lexical resources, such as FrameNet (e.g. verbs related to communication, cognition, perception: *say* = *tell* = *report*, *believe* = *think* = *consider*);
   – disambiguating and linking named entities contained in subjects and objects of the relations (e.g. *President Obama* = *Barack Obama*, *next month* = *June 2015*)
2. Linking relations that describe the same event. This can be achieved by building an event knowlege model, e.g. an event ontology, that will incorporate and meaningfully combine event facets extracted from different sources.

3. Linking events between each other. This task will help to reveal patterns within spatial/temporal/social dimensions by projecting the events on a timeline or a geographic map. This approach may help to learn the commonsense rules useful for reasoning and inference over the event data, such as the 'finish' event follows the 'start' event, but also reveal non-trivial patterns and the outliers.

5 Conclusion

We presented a novel approach to event extraction from Twitter, which builds upon current state-of-the-art relation extraction techniques. We manually evaluated the quality of extracted relations in terms of precision on three real-world datasets. Most of the results returned by the system are correct (88%) and contain descriptive and potentially useful event-related information (71%). However, recall and computational performance of the system was out of scope of this intial evaluation run.

Acknowledgments

This work was supported by REVEAL (http://revealproject.eu/) and DecarboNet (www.decarbonet.eu) projects, which have received funding by the European Unions 7th Framework Program for research, technology development and demonstration under the Grant Agreements No. FP7-610928 and 610829, respectively.

References

1. Puneet Agarwal, Rajgopal Vaithiyanathan, Saurabh Sharma, and Gautam Shroff. Catching the Long-Tail: Extracting Local News Events from Twitter. In *ICWSM*, 2012.
2. L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. In *Multimedia*, volume 15, 2013.
3. Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*, 2011.
4. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 2010.
5. Smitashree Choudhury and John G. Breslin. Extracting semantic entities and events from sports tweets. In *'Making Sense of Microposts': Big Things Come in Small Packages*, 2011.
6. L. Del Corro and R. Gemulla. Clausie: clause-based open information extraction. In *WWW*, 2013.
7. Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2), 2015.

8. Peter Exner and Pierre Nugues. Using semantic role labeling to extract events from Wikipedia. In *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE)*, 2011.
9. A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.
10. Yuheng Hu, Ajita John, Dore Duncan Seligmann, and Fei Wang. What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds. In *ICWSM*, 2012.
11. D. Klein and C. D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on ACL*, 2003.
12. D. Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, February 2001.
13. Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.
14. M., M. Schmitz, R. Bart, S. Soderland, and O. Etzioni. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.
15. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the ACL*, 2014.
16. M McCord and M Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and Trusted Computing*. Springer, 2011.
17. S. Papadopoulos, D. Corney, and L. M. Aiello. Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media. In *SNOW-DC@WWW*, 2014.
18. Thomas Ploeger, Maxine Kruijt, Lora Aroyo, Frank De Bakker, Iina Hellsten, and Antske Fokkens. Extractivism: Extracting activist events from news articles using existing NLP tools and services. In *The 12th International Semantic Web Conference (ISWC)*, 2013.
19. Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web*, 2011.
20. Alan Ritter, Oren Etzioni, Sam Clark, and others. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.
21. Willem Robert Van Hage, Vronique Malais, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011.
22. Willem Robert van Hage, Vronique Malais, Marieke Van Erp, and Guus Schreiber. Linked open piracy. In *Proceedings of the sixth international conference on Knowledge capture*, 2011.
23. Guido Van Oorschot, Marieke Van Erp, and Chris Dijkshoorn. Automatic extraction of soccer game events from twitter. In *Proc. of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*, 2012.
24. Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR*, 1998.