Social Event Detection on Twitter

Elena Ilina[1], Claudia Hauff[1], Ilknur Celik[2], Fabian Abel[1], and Geert-Jan Houben[1]

[1] Web Information Systems, Delft University of Technology
{e.a.ilina,c.hauff,f.abel,g.j.p.m.houben}@tudelft.nl
[2] Middle East Technical University Northern Cyprus Campus
cilknur@metu.edu.tr

Abstract. Various applications are developed today on top of microblogging services like Twitter. In order to engineer Web applications which operate on microblogging data, there is a need for appropriate filtering techniques to identify messages. In this paper, we focus on detecting Twitter messages (tweets) that report on social events. We introduce a filtering pipeline that exploits textual features and n-grams to classify messages into event related and non-event related tweets. We analyze the impact of preprocessing techniques, achieving accuracies higher than 80%. Further, we present a strategy to automate labeling of training data, since our proposed filtering pipeline requires training data. When testing on our dataset, this semi-automated method achieves an accuracy of 79% and results comparable to the manual labeling approach.

Keywords: microblogging, Twitter, event detection, classification, semi-automatic training.

1 Introduction

Twitter is a popular micro-blogging web application serving millions of users. Twitter users chat and share information on news, work-related issues and community matters [1]. Despite the noise in Twitter blogs [2], Web applications can exploit the blogs' content as a source of information to identify natural disasters [3], news [2], or social events [4]. Since the existing Twitter search is cumbersome in finding event-related information [5], a targeted search aiming at finding tweets specifically related to real-life events might be useful. The capability of searching real-life events would be of great benefit for personalization purpose in search.

Accordingly, our motivation is to separate event-related content from the rest of micro-posts. For this, the large volume of non-event-related messages is one of the paramount challenges to be solved. Our approach could be used as a first filtering step before applying other techniques for finding event-related content. The goal is to identify tweets related to real-life events, social events such as music concerts and festivals. Based on Twitter content published by event broadcasters, we train our classification model to distinguish social events from other tweets. The proposed approach is based on a text classification technique, which enables to classify content into two mutually exclusive groups.

M. Brambilla, T. Tokuda, and R. Tolksdorf (Eds.): ICWE 2012, LNCS 7387, pp. 169–176, 2012.

We employ the Naive Bayes classification algorithm utilizing features extracted from the Twitter messages. For training our classifier, we assume, that the selected event broadcasters publish only event-related tweets. Tweets published by other users are initially assigned to the *Other* class. We apply heuristic rules defining the presence of event-related aspects such as time, persons involved and locations. In the presence of these three aspects, the *Events* class is assigned, otherwise the class *Other*. This semi-automatic training approach enables automatic identification of event-related Twitter content and helps to achieve comparable results with supervised learning approach, while reducing efforts of manually labeling training datasets.

Our main contributions include: a semi-automatic training approach for training a classification model which assists in determining event-related tweets, the application of a Naive Bayes classification to identify tweets related to social events based on the proposed semi-automated approach, text preprocessing strategies to improve tweet classification outcomes, an evaluation of the semi-automated learning approach and classifier.

2 Related Work

Twitter received much attention in recent years. Twitter data is openly available, motivating research in social interactions on the Web, micro-blogging and data mining. At the same time, Twitter differs from other blogging software due to its shorter messages, facilitating up-to-date publishing [1]. Researchers have been motivated to analyze Twitter as a source of sensory information provided by Twitter users, reacting on real-life events such as social events [6,7] or natural occurrences as earthquakes [3].

The most prominent works investigating event detection on Twitter are based on statistical [3] and machine learning techniques [8,7]. Sakaki et al. [3] applied classification and particle filtering methods to event detection from Twitter messages, reporting a significant accuracy in detecting earthquakes. Chakrabarti and Punera [9] proposed an approach to group event-related tweets in real-time applying Hidden Markov Models. Their approach can be used for well-structured events, but requires prior knowledge on events, participating athletes and defined event-related hashtags. [8] applied an online clustering approach, grouping tweets with similar content together. After manually labeling clusters as event-related and not event-related, they trained a Naive Bayes text classifier for identifying event-related tweets. This approach, however, requires calculating pairwise similarities before actually identifying tweets as related to events. Popescu et al. [7] used named entity recognition and decision trees, calculating the quantity of found named entities in time.

Other works identify event content using other information sources besides Twitter. Benson et al. [4] align tweets with particular events mentioned in a city guide, employing a distant-supervision approach for training their event classification model. Social blogging content as a source of information for user opinions on events was investigated by [6], which created software that retrieves tweets related to events published on the *Upcoming* web site.

Due to the inherent lack of structure in micro-posts, Twitter is a challenging media platform to work particularly when it comes to identifying relevant tweets [5]. Sankaranarayanan et al. [2] stated that the noise of Twitter messages leads to large volumes of unrelated tweets, introducing an increased complexity of events identification. None of the aforementioned works, however, investigate in-depth the problem of identifying tweets related to social events based solely on the tweets' content. We close this gap by focusing on social event detection, applying a semi-automated learning classification technique similar to [8]. In contrast, our semi-automated classification approach is based on tweets content and does not require additional data sources, clustering or named entity recognition steps. Our approach can be used for filtering event-related content on microblogs.

3 Social Events on Twitter

Twitter microblog posts may include any free-text, special tags or links to other Web resources and are limited to 140 characters. This is why tweets' content often include abbreviations, shortened words or phrases, as well as shortened Uniform Resource Locators (URLs). Forwarding services such as *bit.ly* or *oil.ly* are used to decode the shortened links. Given these limitations, Twitter users try to convey their ideas in a very concise form and make use of special labels, so-called *hashtags*, for tagging the topics of their tweets. For referring to other Twitter users or replying to them, the "@"-symbol is used.

Twitter user profiles are usually linked with profiles of other users, called followers and friends. Twitter can be used for communicating with networking partners, organizations, music bands and even famous people. Twitter assists in marketing and promotion and is therefore widely used by advertising agencies and social media broadcasters to inform on social happenings such as touring artists or upcoming concerts.

In previous works such as [3], events are typically defined using the time and location dimensions. Since social occasions such as music concerts involve musicians and music bands, social events can also be defined by the personalities and/or organizations involved. Therefore, we choose to describe a social event such as a music festival by three main dimensions: "agents involved", "time" and "location". When a particular tweet does not mention all three dimensions, missing dimensions have to be inferred from its content. For instance, we have observed that time references were included in less than 30% of 333 tweets randomly selected from our initial dataset. Only 10% of the 333 tweets mentioned all the three event dimensions, of which 9% were event-related tweets and 1% of tweets were not related to events. This implies that the majority of event-related tweets contain references to these three dimensions, while most of the tweets referring only to one dimension are likely not to be related to social events. Overall, the largest discrepancy was detected for the combinations of event dimensions "Location+Artist+Time" and "Location+Time", which were identified 9 and 6 more times respectively for event related tweets compared to non-event

related tweets. This means that time and location dimensions are paramount for finding event-related tweets, whilst adding the artist dimension increases the identification of event-related tweets.

4 Classification Approach

For the implementation, we adapt the standard Naive Bayes classification [10] approach and employ n-gram features. Kanaris et al. [11] argue that character sequence n-gram classification models are relatively resilient towards spelling errors, do not require stemming procedures and can help in decreasing a feature set when compared with word level n-grams. The reason for this is, that there are more n-gram word combinations compared to the number of character n-gram combinations defined by the number of characters used in a particular vocabulary. The lexical benefits of the character n-grams was a motivation for us to create the character n-gram classifier for working with Twitter data. Based on our goal of determining if a particular tweet is related to an event or not, we formulate the following binary classification problem:

Tweet Classification Problem: *Given a tweet $t \in T$, the classification algorithm is used to label the tweet as event related or non-event related by approximating the function $F : T \longrightarrow C$ mapping tweets to their respective classes $C = \{Events, Other\}$*

Based on the Bayes theorem [10], we can calculate the probability $P(C|t)$ of a tweet t belonging to the class C using:

$$P(C|t) = P(t|C) * P(C)/P(t) \ , \tag{1}$$

with $P(t|C)$ the conditional probability of observing tweet t in class C, $P(C)$ the unconditional probability of observing class C, and $P(t)$ the probability of observing tweet t. Next, each tweet we break into a set of n-grams, called $g_1, \ldots, g_m$. For calculating the likelihood that an n-gram appears in the class C, we calculate the product of probabilities of all n-grams based on the Naive Bayes assumption that n-grams appear independently from each other:

$$P(t|C) \backsimeq P(g_1|C) * P(g_2|C) * \ldots * P(g_m|C) \ . \tag{2}$$

For calculating the probability of a particular n-gram g belonging to the class C, we divide the number of times n-gram g appears in the class C by the total number of n-grams in the class C. The likelihood of class C is computed by dividing the total number of n-grams of the class C by the total number of n-grams in both categories, *Events* and *Other*. Finally, we identify the largest $P(C|t)$, which will be the classification class (*Events* or *Other*) assigned to the tweet t, while ignoring $P(t)$, which is the same for both classes.

For training our classification models we consider manual and semi-automatic labeling. In both cases, we apply several heuristic rules rather than selecting training instances randomly. The reasoning behind this choice is that in our

dataset, the ratio between "event" and "not event" tweets from the tweets sample of 333 tweets mentioned above was 0.25. Our aim was to increase the number of training instances while ensuring a satisfactory classification performance. For manual labeling, we follow shortened URLs as generated by shortening services such as *bit.ly* or *oil.ly* and considered only tweets including the sub-strings: "/event/", "/artists/" or "/venue/". Interestingly, only roughly two out of three tweets having such URLs are event-related.

For semi-automated labeling, we include the tweets of the selected event broadcasters into our training dataset of positive instances (*Events* class), when they include the mention of time concepts, references to other users, and words starting with capital letters. For identifying time dimensions, we consider date and time mentions, or words and phrases such as "today", "this evening" or "this summer". In order to relate tweet content elements to the "involved agents" dimension, we consider not only accurately spelled artist names, but also their twitter names. This way we avoid a named entity recognition step for detecting artist and location names. Tweets that do not satisfy heuristic rules of the positive class are assigned to the negative training set (*Other* class).

Hovold [12] demonstrated that the removal of stopwords improves classification accuracy in the context of spam detection and that punctuation marks can have a negative effect on classification. We experiment with removing stopwords, punctuation marks, shortened URLs, hashtags and user mentions for selecting our text-preprocessing strategy applied to Twitter content.

5 Evaluation

In this section we evaluate our Twitter content classification approach. We identify which of the proposed text preprocessing strategies and n-gram sizes are best suited for manual evaluation. The selected n-gram size and text preprocessing strategy are further applied to compare supervised and semi-automated learning approaches.

For running our classification experiments, we created six datasets[1]. The datasets have quite different proportions of "event-related" and "not event-related" tweets (which we denote as R_e ratio), due to their different origin. Datasets TEST_{MIT} (total number of instances N=334, R_e=1) and TRAIN_{MIT} (N=2400, R_e=1) were created by selecting tweets which content overlaps with strings from the New York city guide and provided by [4]. The rest of datasets were published by event broadcasters having more than 1000 of followers and being included in at least ten public lists. For each of the selected 30 broadcasters we followed their 1000 random followers, which posted at least 200 tweets each. Datasets $TRAIN_{auto_1}$ (N=13615, R_e=0.36) and $TRAIN_{auto_2}$ (N=267938, R_e=0.12) were automatically labeled as described in the previous section. Datasets TRAIN_{TUD} (N=2400, R_e=1) and TEST_{TUD} (N=333, R_e=0.33) were labeled manually.

First, based on the manually labeled datasets, TEST_{TUD} and TRAIN_{TUD}, we have found that removal of URLs, hashtags, user mentions and punctuation marks

[1] See: `http://www.wis.ewi.tudelft.nl/people/elena/elenaprojects/events/`

has a positive influence on classification performance, increasing F_1-measure in 17% and accuracy from 81% to 84%. Removal of stopwords had a negative impact on all performance metrics. Second, after removing hashtags, URLs, user mentions and punctuation marks from tweets, we identify the best performing n-gram size of 4, resulting in a precision of 96% of events detection for the manually labeled dataset. Therefore, in the next experiments we employed 4-grams, we left stopwords and removed other syntactic elements mentioned above.

Table 1 summarizes the tests we performed with cross-validation of testing and training datasets. The first two tests were performed on manually labeled training sets and achieved an above baseline accuracy of 50%. However, in the second test using the TEST_{MIT} and the TRAIN_{TUD} datasets, we achieved a lower performance for all metrics. Test 4 using TEST_{TUD} testing set and TRAIN_{MIT} training set did not achieve an accuracy of baseline accuracy value. We explain this by the different features used for creating the classification models. Both training sets have different historic data, while our tweets selection strategy differs considerably.

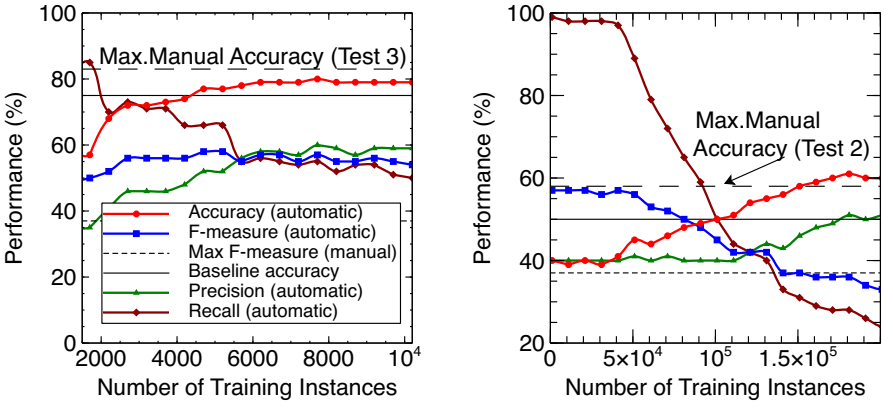
Table 1. Performance on Different Testing Datasets (percentages), where $A_{baseline}$ and $A_{achieved}$ are respective accuracies

Test	Testing	Training	$A_{baseline}$	$A_{achieved}$	Precision	Recall	F_1
1	TEST_{MIT}	TRAIN_{MIT}	50	71	66	83	74
2	TEST_{MIT}	TRAIN_{TUD}	50	58	88	18	30
3	TEST_{TUD}	TRAIN_{TUD}	75	83	96	32	48
4	TEST_{TUD}	TRAIN_{MIT}	75	43	25	67	36
5	TEST_{TUD}	TRAIN_{auto_1}	75	79	63	41	50
6	TEST_{MIT}	TRAIN_{auto_2}	50	60	52	20	29

Figure 1 (a) shows that accuracy of classification using the semi-automatic training improves with a growing number of training instances. After reaching about 5000 training instances, the classification accuracy is above the baseline classification[2] accuracy of 75% when tested on the $TEST_{TUD}$ dataset. The F_1-measure stays above the F_1-measure of the manually-trained classifier. In test 5, we employ TEST_{TUD} and achieve comparable results with the test 3 performed on manually labeled dataset. We observe a drop in precision from 96% to 63%, while, for recall and F_1-measure, we have a slight improvement for the semi-automatic training approach.

In test 6 performed on the TEST_{MIT} dataset, we increase the number of training instances up to 267938. As shown in the Figure 1 (b), we achieve an accuracy of 60%, which is comparable with the accuracy achieved when using the manual labeling approach in test 2. We achieve very similar performance values for tests 2 and 6; however, in test 6 we observe decreased precision.

[2] In our case the baseline classifier is a default classifier predicting a majority class of non-events.



(a) Test 5 (b) Test 6

Fig. 1. Semi-automatic Classification Performance

6 Conclusion and Future Work

In the foregoing, we propose a semi-automatic approach for detecting event-related tweets. This will allow to exploit large volumes of micro-blogging content for providing information on social events. The aim is eventually to use for instance Twitter content in web applications listing concerts, taking into account factors like a specific time or date, location or performers. For this, we use a classification approach based on Naive Bayes and n-gram features extracted from Twitter content of event broadcasters and their followers. The training and testing datasets are built up on a classifier of manually labeled tweets, with which we achieve high precision and accuracy. Training the classifier in a semi-automatic way using content of pre-selected broadcasters would allow to reduce manual labeling efforts. With a growing number of training instances, the prediction accuracy of the classifier using the proposed semi-automatic training approach is comparable to the classifier created on a manually labeled training set. Future work will include using the classifier with different and larger scale datasets derived from Twitter content, developing a classifier that could outperform one requiring manual labeling.

Acknowledgments. This work is partially sponsored by the ImREAL project (http://imreal-project.eu).

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th Workshop on Web Mining and Social Network Analysis (WebKDD), pp. 56–65. ACM (2007)

2. Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., Sperling, J.: Twitterstand: news in tweets. In: Proceedings of the 17th International Conference on Advances in Geographic Information Systems (SIGSPATIAL), pp. 42–51. ACM (2009)
3. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web (WWW), pp. 851–860. ACM (2010)
4. Benson, E., Haghighi, A., Barzilay, R.: Event discovery in social media feeds. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2011), pp. 389–398. Association for Computational Linguistics (2011)
5. Abel, F., Celik, I., Houben, G.-J., Siehndel, P.: Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 1–17. Springer, Heidelberg (2011)
6. Becker, H., Chen, F., Iter, D., Naaman, M., Gravano, L.: Automatic identification and presentation of twitter content for planned events. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011), pp. 655–656. AAAI Press (2011)
7. Popescu, A., Pennacchiotti, M., Paranjpe, D.: Extracting events and event descriptions from Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 105–106. ACM (2011)
8. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), North America, pp. 438–441. AAAI Press (July 2011)
9. Chakrabarti, D., Punera, K.: Event summarization using tweets. In: Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM), pp. 66–73. AAAI Press (2011)
10. Lewis, D.: Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 4–15. Springer, Heidelberg (1998)
11. Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E.: Words vs. character n-grams for anti-spam filtering. International Journal on Artificial Intelligence Tools 16(6), 1047–1067 (2007)
12. Hovold, J.: Naive bayes spam filtering using word-position-based attributes. In: Proceedings of the Second Conference on Email and Anti-Spam (CEAS 2005), Stanford University, California, USA, pp. 1–8 (2005)