Detecting Event-Related Links and Sentiments from Social Media Texts

Alexandra Balahur and **Hristo Tanev**
European Commission Joint Research Centre
Via E. Fermi 2749, T.P. 267
21027 Ispra (VA), Italy
{alexandra.balahur, hristo.tanev}@jrc.ec.europa.eu

Abstract

Nowadays, the importance of Social Media is constantly growing, as people often use such platforms to share mainstream media news and comment on the events that they relate to. As such, people no loger remain mere spectators to the events that happen in the world, but become part of them, commenting on their developments and the entities involved, sharing their opinions and distributing related content. This paper describes a system that links the main events detected from clusters of newspaper articles to tweets related to them, detects complementary information sources from the links they contain and subsequently applies sentiment analysis to classify them into positive, negative and neutral. In this manner, readers can follow the main events happening in the world, both from the perspective of mainstream as well as social media and the public's perception on them.

This system will be part of the EMM media monitoring framework working live and it will be demonstrated using Google Earth.

1 Introduction

In the context of the Web 2.0, the importance of Social Media has been constantly growing in the past years. People use Twitter, Facebook, LinkedIn, Pinterest, blogs and Web forums to give and get advice, share information on products, opinions and real-time information about ongoing and future events. In particular Twitter, with its half a billion active members, was used during disasters, protests, elections, and other events to share updates, opinions, comments and post links to online resources (e.g. news, videos, pictures, blog posts, etc.). As such, Twitter can be used as a complementary source of information, from which we can retrieve additional facts, but also learn about the attitude of the people towards certain events. On the one hand, news from the traditional media focus on the factual side of events, important for the society or at least large groups of people. On the other hand, social media reflects subjective interpretations of facts, with different levels of relevance (societal or only individual). Therefore, the events reported in online news can be considered a point of intersection for both types of media, which are able to offer complementary views on these events.

In this context, we describe a system that we developed as an additonal component to the EMM (Europe Media Monitor)[1] news monitoring framework, linking mainstream news to related texts from social media and detecting the opinion (sentiment) users express on these topics.

In the EMM news monitoring system, the different news sites are monitored and new articles are scraped from them, with a refresh rate of 10 minutes. Subsequently, news items are clustered and the most important ones are displayed (top 10). These are called "stories". Our system subsequently links these stories to messages from Twitter (tweets) and extracts the related URLs they contain. Finally, it analyzes the sentiments expressed in the tweets by using a hybrid knowledge-based and statistical sentiment detection module. The overview of the system is depicted in Figure

[1] http://emm.jrc.it/NewsBrief/clusteredition/en/latest.html

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 25–30,
Sofia, Bulgaria, August 4-9 2013.

1.

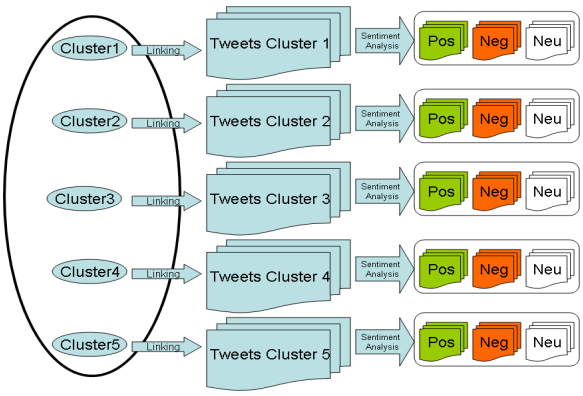

Figure 1: Overview of the news clusters-Twitter linking and sentiment analysis system.

The system will be demonstrated using the Google Earth interface (Figure 2), presenting the characteristics of the event described in the story (type, date, location, the first words in the article that is the centroid of the news cluster for that story). In addition, we present new information that we extract from Twitter - links (URLs) that we find from the tweets we retrieved linked to the story and positive, negative and neutral sentiment, respectively, as a proportion of the total number of tweets retrieved.

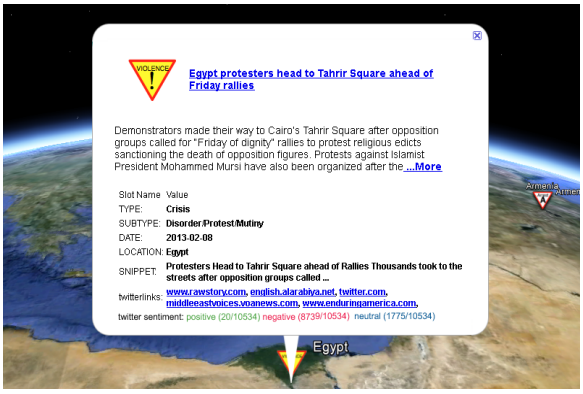

Figure 2: Demo interface for the event-Twitter linking and sentiment analysis.

2 Related Work and Contribution

The work presented herein is mostly related to the linking of events with social media texts and sentiment analysis from Twitter.

Although Twitter was used as an information source in the context of different crisis events, relatively little work focused on linking and extracting content about events which are known *a priori*, e.g., Becker et al. [2011].

In this context, the main challenge is to determine relevant keywords to search for event-related tweets and rank them according to their relevance. Related approaches (e.g., Verma et al. [2011]) report on the use of semantic features (e.g., objectivity, impersonality, formality, etc.) for detecting tweets with content relevant to situational awareness during mass emergencies. Other approaches elaborate on machine learning-based techniques for Named Entity Recognition (NER) from tweets, which are subsequently employed as search query terms (Ritter et al. [2011], Liu et al. [2011]).

Related research on sentiment analysis from Twitter was done by Alec Go and Huang [2009], Pak and Paroubek [2010] and Agarwal et al. [2011]. Alec Go and Huang [2009] and Pak and Paroubek [2010] exploit the presence of emoticons that represent positive or negative feelings to build a training set of tweets with sentiment labels, using which they build models based on n-gram features and part-of-speech tags. Agarwal et al. [2011] employ emoticons dictionaries and replace certain elements such as URLs and topics with predefinded labels. They employ syntactic features and specialized tree kernels and obtain around 75% to 80% accuracy for the sentiment classification.

The main contributions of our system reside in the linking of mainstream news to the complementary content found in social media (tweets and, through them, to the links to additional information sources like blogs, flickr, youtube, etc.) and the analysis of sentiment on these important news. For events such as “The Arab Spring”, protests, financial news (e.g. the fluctuations of the Euro, the bailout of different European countries, the rise in unemployment rate, etc.), it was seen that the sentiment expressed in social media has a high impact on the subsequent development of the story[2] (Saif et al. [2012], Bollen et al. [2011]). The impact of sentiment expressed in social media is also visible for topics which apparently have an apriori valence (e.g. disasters, crisis, etc.). Nevertheless, in these cases, people communicate using the social media platforms not only to express their negative feelings, but also their will to help, their situation, their messages of encouragement, their gratefulness for the help and so on.

[2]http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf

Secondly, the methods employed in our system are simple, work fast and efficient and can be easily adapted to other languages.

Finally, the methods presented take into account the specificity of social media languages, applying methods to normalize the language and adapting the features considered for the supervised learning process.

3 Linking News Clusters to Twitter

The first step in our system involves linking the news stories detected by EMM to related tweets. The linking system employs the Twitter Search API[3]. For each news story, our application detects relevant URLs by finding tweets that are lexically similar to the news story, represented by a cluster of news, and are mentioned frequently in Twitter. In Figure 3, we provide an example of the top six stories on the afternoon of April 2nd, 2013.

Figure 3: Top six clusters of news in the afternoon of April 2nd, 2013.

In order to detect lexically similar tweets, we use vector similarity: We build a term vector for both the news story and the tweet and then we consider as a similarity measure the projection of the tweet vector on the story vector. We do not calculate cosine similarity, since this would give an advantage to short tweets. We experimentally set a similarity threshold above which the tweets with URL are accepted. To define the similarity threshold and the coefficients in the URL ranking formula, we used a development set of about 100 randomly selected English-language news clusters, downloaded during a week. The threshold and the coefficients were derived empirically. We consider experimenting with SVM and other machine-learning approaches to define these parameters in a more consistent way.

Once the tweets that relate to the news story are retrieved, we evaluate each URL taking into account the following parameters:

- Number of mentions, which we will designate as $Mentions$.
- Number of retweets, designated $Retweet$.
- Number of mentions in conversations, designated $InConv$.
- Number of times the URL was favortited, designated $Favorited$.
- Number of tweets which replied to tweets, mentioning the URL, designated $ReplyTo$.

The score of the URL is calculated using the following empirically derived formula. The coefficients were defined based on the empirical analysis described above.

$$score(URL) = ((Mentions-1)+Retweets.1,3$$
$$+Favorited*4).(InConv+2*ReplyTo+1)$$

In this formula we give slight preference to the retweets with respect to the mentions. We made this choice, since retweets happen inside Twitter and reflect the dynamics of the information spread inside this social media. On the other hand, multiple mentions of news-related tweets (which are not retweeted) are due to clicking the "Share in Twitter" button, which nowadays is present on most of the news sites. In this way, news from visited web sites appear more often in Twitter. This phenomena is to be further explored. It should also be noted that our formula boosts significantly URLs, which are mentioned inside a conversation thread and even more the ones, to which there were "reply to" tweets. Conversations tend to be centered around topics which are of interest to Twitter users and in this way they are a good indicator of how interesting an URL is. Replying to a tweet requires more time and attention than just pressing the "Retweet" button, therefore conversations show more interest to an URL, with respect to retweeting. Examples of tweets extracted that complement information from mainstream media are presented in Figure 4.

[3]https://dev.twitter.com/docs/api/1/get/search

Figure 4: Examples of tweets extracted on the North Korea crisis (anonimized).

4 Sentiment Analysis on Tweets Related to Events Reported in News

After extracting the tweets related to the main news clusters detected by the media monitoring system, we pass them onto the sentiment analysis system, where they are classified according to their polarity (into positive, negative and neutral).

In order to classify the tweet's sentiment, we employ a hybrid approach based on supervised learning with a Support Vector Machines Sequential Minimal Optimization (SVM SMO - Platt [1998]) linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. We do not employ any specific language analysis software. The aim is to be able to apply, in a straightforward manner, the same approach to as many languages as possible. The approach can be extended to other languages by using similar dictionaries that have been created in our team.

The sentiment analysis process contains two stages: preprocessing and sentiment classification.

4.1 Tweet Preprocessing

The language employed in Social Media sites is different from the one found in mainstream media and the form of the words employed is sometimes not the one we may find in a dictionary. Further on, users of Social Media platforms employ a special "slang" (i.e. informal language, with special expressions, such as "lol", "omg"), emoticons, and often emphasize words by repeating some of their letters. Additionally, the language employed in Twitter has specific characteristics, such as the markup of tweets that were reposted by other users with "RT", the markup of topics using the "#" (hash sign) and of the users using the "@" sign.

All these aspects must be considered at the time of processing tweets. As such, before applying supervised learning to classify the sentiment of the tweets, we preprocess them, to normalize the language they contain. The preprocessing stage contains the following steps:

- ***Repeated punctuation sign normalization.*** In the first step of the preprocessing, we detect repetitions of punctuation signs (".", "!" and "?"). Multiple consecutive punctuation signs are replaced with the labels "multistop", for the fullstops, "multiexclamation" in the case of exclamation sign and "multiquestion" for the question mark and spaces before and after.

- ***Emoticon replacement.*** In the second step of the preprocessing, we employ the annotated list of emoticons from SentiStrength[4] and match the content of the tweets against this list. The emoticons found are replaced with their polarity ("positive" or "negative") and the "neutral" ones are deleted.

- ***Lower casing and tokenization.*** Subsequently, the tweets are lower cased and split into tokens, based on spaces and punctuation signs.

- ***Slang replacement.*** The next step involves the normalization of the language employed. In order to be able to include the semantics of the expressions frequently used in Social Media, we employed the list of slang from a specialized site [5].

- ***Word normalization.*** At this stage, the tokens are compared to entries in Roget's Thesaurus. If no match is found, repeated letters are sequentially reduced to two or one until a match is found in the dictionary (e.g. "perrrrrrrrrrrrrrrrrfeeect" becomes "perrfeect", "perfeect", "perrfect" and subsequently "perfect"). The words used in this form are maked as "stressed".

- ***Affect word matching.*** Further on, the tokens in the tweet are matched against three different sentiment lexicons: General Inquirer, LIWC and MicroWNOp, which were previously split into four different categories

[4] http://sentistrength.wlv.ac.uk/

[5] http://www.chatslang.com/terms/social_media

("positive", "high positive", "negative" and "high negative"). Matched words are replaced with their sentiment label - i.e. "positive", "negative", "hpositive" and "hnegative".

- ***Modifier word matching.*** Similar to the previous step, we employ a list of expressions that negate, intensify or diminish the intensity of the sentiment expressed to detect such words in the tweets. If such a word is matched, it is replaced with "negator", "intensifier" or "diminisher", respectively.

- ***User and topic labeling.*** Finally, the users mentioned in the tweet, which are marked with "@", are replaced with "PERSON" and the topics which the tweet refers to (marked with "#") are replaced with "TOPIC".

4.2 Sentiment Classification of Tweets

Once the tweets are preprocessed, they are passed on to the sentiment classification module. We employed supervised learning using SVM SMO with a linear kernel, employing boolean features - the presence or absence of unigrams and bigrams determined from the training data (tweets that were previousely preprocessed as described above) that appeared at least twice. Bigrams are used especially to spot the influence of modifiers (negations, intensifiers, diminishers) on the polarity of the sentiment-bearing words. We tested the approach on different datasets and dataset splits, using the Weka data mining software [6]. The training models are built on a cluster of computers (4 cores, 5000MB of memory each).

5 Evaluation and Discussion

5.1 Evaluation of the News-Twitter Linking Component

The algorithm employed to retrieve tweets similar to news clusters was evaluated by Tanev et al. [2012]. The precision attained was 75%. Recall cannot be computed, as the use of the Twitter API allows only the retrieval of a subset of tweets.

In order to evaluate the link extraction component, we randomly chose 68 URLs, extracted from 10 different news stories. For each URL, we evaluated its relevance to the news story in the following way: A URL is considered relevant only if it reports about the same news story or talks about facts, like effects, post developments and motivations, directly related to this news story. It turned out that 66 out of the 68 were relevant, which gives accuracy of 97%.

5.2 Evaluation of the Sentiment Analysis System

In order to evaluate the sentiment analysis system on external resources, we employed the data provided for training in the SemEval 2013 Task 2 "Sentiment Analysis from Twitter" [7]. The initial training data has been provided in two stages: 1) sample datasets for the first task and the second task and 2) additional training data for the two tasks. We employ the joint sample datasets as test data (denoted as $t*$) and the data released subsequently as training data (denoted as $T*$). We employ the union of these two datasets to perform cross-validation experiments (the joint dataset is denoted as $T * +t*$. The characteristics of the dataset are described in Table 1. On the last column, we also include the baseline in terms of accuracy, which is computed as the number of examples of the majoritary class over the total number of examples. The results of the experiments

Data	#Tweet	#Pos	#Neg	#Neu	B%
T*	19241	4779	2343	12119	62
t*	2597	700	393	1504	57
T*+t*	21838	5479	2736	13623	62

Table 1: Characteristics of the training (T*), testing (t*) and joint training and testing datasets.

are presented in Table 2. Given the difficulty of

Measure	Train(T*) & test(t*)	10-fold CV
Acc.	0.74	0.93
P_{pos}	0.66	0.91
R_{pos}	0.88	0.69
P_{neg}	0.94	0.62
R_{neg}	0.81	0.49
P_{neu}	0.93	0.80
R_{neg}	0.97	0.82

Table 2: Results in terms of accuracy and precision and recall per polarity class on training and test sets evaluation and 10-fold cross-validation.

language in social media, the results are good and

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] http://www.cs.york.ac.uk/semeval-2013/task2/

useful in the context of our application (Figure 2).

6 Conclusions and Future Work

In this demo paper, we presented a system that links mainstream media stories to tweets that comment on the events covered. The system retrieves relevant tweets, extracts the links they contain and subsequently performs sentiment analysis. The system works at a good level, giving an accurate picture of the social media reaction to the mainstream media stories.

As future work, we would like to extend the system to more languages and analyze and include new features that are particular to social media to improve the performance of both the retrieval and sentiment analysis components.

Acknowledgements

We would like to thank the EMM team of the OPTIMA action at the European Commission Joint Research Centre for the technical support.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of LSM 2011*, LSM '11, pages 30–38, 2011.

Richa Bhayani Alec Go and Lei Huang. Twitter sentiment classication using distant supervision. Technical report, Technical report, Stanford University, 2009.

Hila Becker, Feiyang Chen, Dan Iter, Mor Naaman, and Luis Gravano. Automatic identification and presentation of twitter content for planned events. In *Proceedings of ICWSM 2011*, 2011.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing Named Entities in Tweets. In *Proceedings of ACL 2011*, pages 359–367, Stroudsburg, PA, USA, 2011.

Alexander Pak and Patrick Paroubek. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of SemEval 2010*, SemEval '10, pages 436–439, 2010.

John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in Kernel Methods - Support Vector Learning, 1998.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of EMNLP 2011*, pages 1524–1534, Edinburgh, Scotland, UK., 2011.

Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. In *Making Sense of Microposts (#MSM2012)*, pages 2–9, 2012.

Hristo Tanev, Maud Ehrmann, Jakub Piskorski, and Vanni Zavarella. Enhancing event descriptions through twitter mining. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. Natural Language Processing to the Rescue? Extracting "Situational Awareness"' Tweets During Mass Emergency. In *Proceedings of ICWSM 2011*, pages 385–392. AAAI, 2011.