TOWARDS SOCIAL MEDIA MINING: TWITTEROBSERVATORY

Inna Novalija, Miha Papler, Dunja Mladenić
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144; fax: +386 1 4251038
e-mail: inna.koval@ijs.si

ABSTRACT

This paper presents an approach to social media mining based on a pipeline that implements observing, enriching, storing, modeling and presentation techniques. While social media mining solutions have been used for information extraction and sentiment identification about certain topics, applying social media for modeling and nowcasting is a promising new research direction. A novel tool TwitterObservatory that allows observing, searching, analyzing and presenting social media is introduced as a part of the research. Illustrative examples of using the proposed pipeline show how the TwitterObservatory implementing the pipeline can support the user in interaction with the social media data.

1 INTRODUCTION

Social media mining refers to data mining of content streams produced by people through interaction via Internet based applications. Social media mining is usually associated with noisy, distributed, unstructured and dynamic data, as well as with informal text processing.

In this paper we propose a pipeline for social media mining that includes observing, enrichment, storage, modeling and user interface components.

In this research we introduce a novel TwitterObservatory tool for observing, searching, analyzing and presenting information obtained from social media and in particular, from Twitter[1].

The paper is structured as follows: Section 2 contains the related work on social media mining; Section 3 describes the social media mining pipeline at a high level; Section 4 introduces the observing techniques for social media; Section 5 provides the insights into enriching and storing techniques for social media; Section 6 presents the user interface; Section 7 introduces modeling as a part of social media mining pipeline and finally, Section 8 concludes the paper.

2 RELATED WORK

The related work in the area of social media mining covers a number of interesting and relevant topics. In particular, researchers discussed summarization of tweets according to the given query [1], summarization of YouTube comments with sentiment detection and tag cloud [2], identification of the main headlines for the day with language modeling [3]. A number of approaches to classification of the informal text have been suggested by Irani et al. [4], Ramage et al. [5], Lambert et al. [6] and Sriram et al. [7]. And while some researchers have been dealing with spam versus non spam classifications [4], other clustered twitter stream into several topics [5] or classes, such as news, events, opinions [7]. Retrieval of the relevant tweets based on trained language model for each hash-tag on tweeter has been covered by [8]. Rupnik et al. [9] suggest a method for multilingual document retrieval through hub languages, which have alignments with many other languages. A special attention should be dedicated to the approaches dealing with sentiment detection in social media streams. Sentiment detection has been performed at different levels, starting with user sentiments about certain topics [10]. Štajner et al. [11] addressed the problem of sentiment analysis in an informal setting in different domains and two languages.

3 SOCIAL MEDIA MINING PIPELINE

In this paper we present a pipeline that implements a complete mechanism for mining social media. The pipeline (Figure 1), uses EventRegistry [12] mechanisms and includes observing, enrichment, storage, modeling and user interface components.

The pipeline finds its practical implementation as a TwitterObservatory[2] tool. TwitterObservatory uses data observation, enrichment and storage techniques for social media data presentation, search and analytics. In addition,

[1] ***twitter**.com*

[2] ***twitterobservatory**.net*

TwitterObservatory provides a suitable user interface that allows users to:

- observe upcoming tweets, search by keywords,
- search social media data by keywords, hashtags etc.
- analyze social media data by volume, sentiment etc.

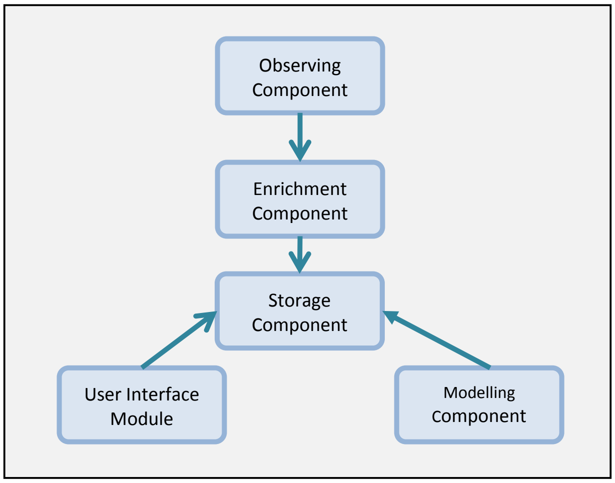


Figure 1: Social Media Mining Pipeline

4 OBSERVING SOCIAL MEDIA

The observing functionalities and mechanisms behind social media mining pipeline and TwitterObservatory are provided in subsections 4a Observing Tweets by Locations and 4b Observing Tweets by Keywords.

a. OBSERVING TWEETS BY LOCATION

This subsection provides description of approaches behind gathering social media data.

For obtaining social media data from Twitter we have used REST Twitter API[3].

The Twitter API allows observing tweets by geo coordinates.

Location can be considered as an important parameter of social media data, since modeling, analyzing and nowcasting is often a location based task.

In this research we have used geo coordinates from United Kingdom. Ten largest cities (by population) have been picked out, then according to Twitter API requirements we have formed a geo coordinates boxes around each city coordinate and set them as filters into Twitter API requests.

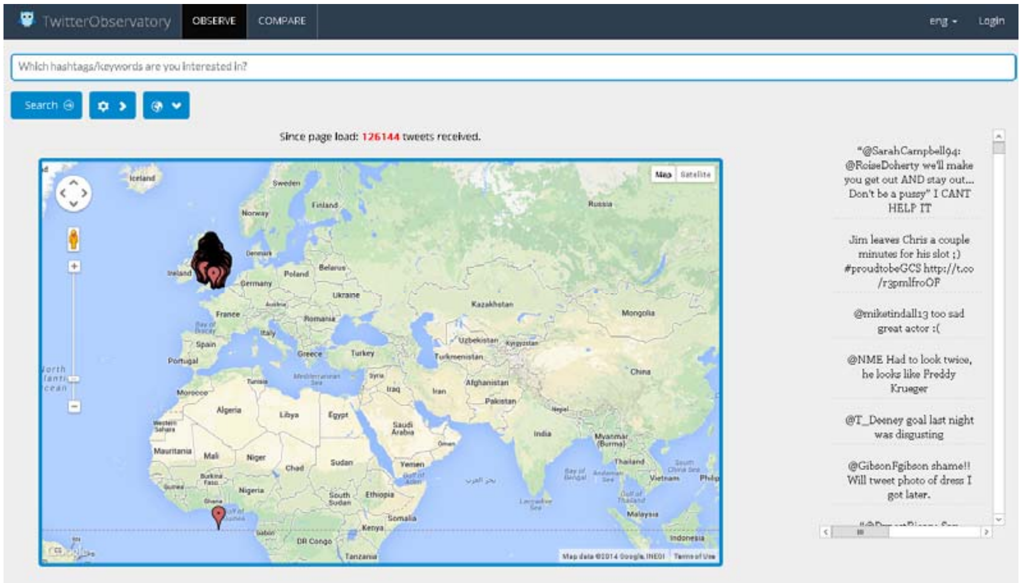
Figure 2 demonstrates upcoming tweets from UK – the location of each upcoming tweet is pinned up on the map and the textual content of the tweet is provided on the right panel.

Overall, we have obtained 31 GB of location based tweets from United Kingdom for a period from November 2013 until July 2014.

As possible to notice, the approach used for data gathering can be easily adapted for other geographical places.

b. OBSERVING TWEETS BY KEYWORDS

Another technique for obtaining tweeter data is to filter social media data by keywords. Up to 400 keywords can be used in one application that uses REST Twitter API. For implementation of this approach we have used the following procedure:



Since page load: 126144 tweets received.

"@SarahCampbell94: @RoiseDoherty we'll make you get out AND stay out... Don't be a pussy" I CANT HELP IT

Jim leaves Chris a couple minutes for his slot ;) #proudtobeGCS http://t.co/r3pmlfroOF

@miketindall13 too sad great actor :(

@NME Had to look twice, he looks like Freddy Krueger

@T_Deeney goal last night was disgusting

@GibsonFgibson shame!! Will tweet photo of dress I got later.

Figure 2: Observing Tweets by Locations (UK)

[3] *https://**dev.twitter**.com/docs/api/1.1*

- select most common words from Wikipedia,
- set up to 400 words in the filter.

An example of most common words in English, according to ranks, are the following: “the”, “be”, “to”, “of”, “and”, “a”, “in”, “that”, “have”, “I”.

Over 500 GB of tweets data have been observed using keyword filters.

5. ENRICHING AND STORING SOCIAL MEDIA DATA

In this section enrichment and storage components of social media data are briefly discussed.

In order to generate additional features that can be used for modeling and nowcasting, we perform enrichment of social media data. The most typical enrichment tasks include sentiment and cross-lingual topic identification of social media data. Enrycher[4] and XLing[5] tools are used for these purposes.

Storage and analytics of social media data is one of the main tasks of the social streams processing infrastructure. Storage component of social media mining pipeline is based on QMiner[6] tool functionalities. QMiner is a data analytics platform for processing real-time streams of structured or unstructured data.

6. USER INTERFACE

In order to give the users a possibility to observe social media data and perform simple analytics tasks based on their experience and intensions, social media mining pipeline contains a user interface module.

In particular, TwitterObservatory provides a suitable user interface that allows user to view upcoming social media data (tweets), search tweets by different queries and analyze the search results within different dimensions.

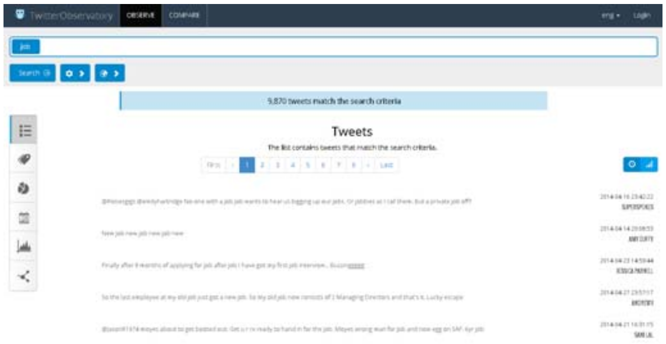
One of the TwitterObservatory functionalities demonstrated at Figure 3 is the possibility to filter the stored social media data by keywords. Keyword “job” is provided as a filter for our storage.

The users can view the text of tweets, the author of the tweet and the publishing date/time.

Figure 4 presents a possibility to obtain a tag cloud for tweets filtered by keyword “job”. The most relevant tags are: “good”, “today”, “time”.

Figure 5 shows a sentiment graph for tweets filtered by keyword “job”. Sentiment varies on a daily basis.

Figure 6 shows a timeline (or volume) for tweets filtered by keyword “job”. Volume of tweets varies on a daily basis.



9,870 tweets match the search criteria

Tweets

The list contains tweets that match the search criteria.

[illegible]

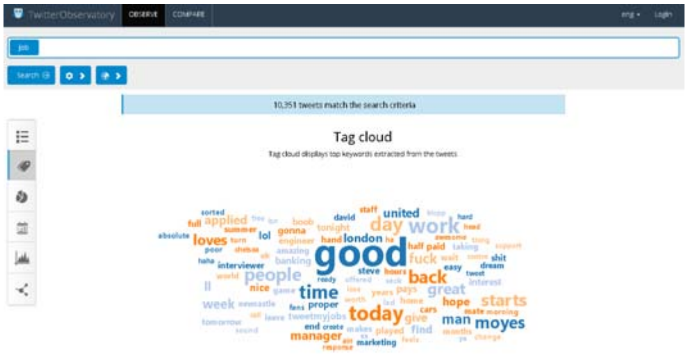
New job new job new job new

Finally after 9 months of applying for job after job I have got my first job interview... Buzzing

[illegible]

[illegible]

Figure 3: Observed Tweets with Details (Filter: “job”)

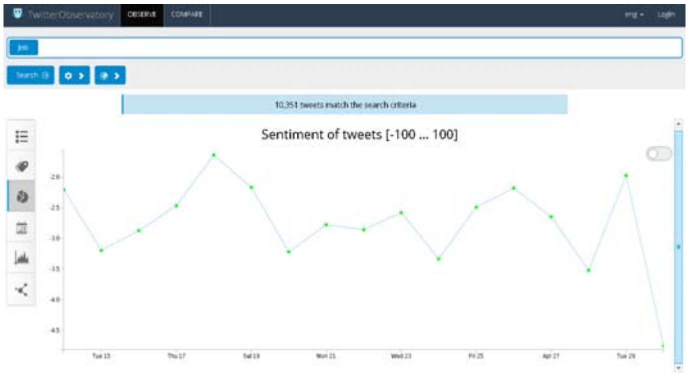


10,351 tweets match the search criteria

Tag cloud

Tag cloud displays top keywords extracted from the tweets

Figure 4: Tag Cloud for Tweets (Filter: ”job”)



10,351 tweets match the search criteria

Figure 5: Sentiment for Tweets (Filter: ”job”)

7. INTRODUCTION TO DATA MODELING

An important part of social media mining pipeline is a modeling component. Modeling and nowcasting functionalities are intended to connect social media with external datasets, such as macroeconomic data.

In particular, the goal of modeling and nowcasting is to relate micro-signals coming from social media (such as

[4] ***enrycher***_.ijs.si_

[5] ***xling***_.ijs.si_

[6] ***qminer***_.ijs.si_

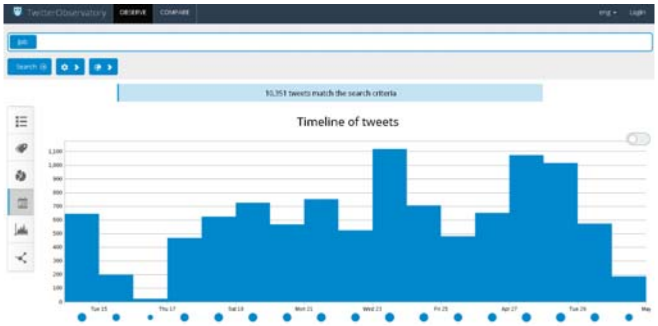


Figure 6: Tweets Timeline (Filter: ”job”)

micro-signals related to stocks, micro-signals related to labor, micro-signals related to consumers, micro-signals related to real estate and credit, micro-signals related to energy) with macro-economic variables.

In order to perform the modeling and nowcasting, a set of features extracted from social media (features, such as volume, sentiment, trending concepts or hashtags etc) should be applied.

First test will be done on data such as NTSF indices and other stock indices relevant to regional based crawling of tweets (most twitter data crawled so far is from UK). Also historic stock market data provides relatively unbiased, general, frequent and free to use data, which we think will be a good starting point for building models. What we are hoping to see in the initial steps when correlating twitter and stock market data is spikes in volume of published tweets with relevant keywords and hashtags. Through further analysis of different correlations we would like to find a map (maybe even a graph) of keywords that best responded to events in each macroeconomic data.

Later we will expand the model for a wide variety of macroeconomic data from different fields as mentioned before. Frequency is an important factor in what data will be used, and so is diversity. Ideally we want to make a model which will cover all aspects of economic environment, so we could study how events in different areas of economy influence public opinions which we hope to see mirrored in twitter data.

Combined features from social media should be correlated with macroeconomic time series, with a number of operators for time series analysis used (moving average (MA), exponential moving average (EMA), moving average convergence/divergence (MACD), moving norm, variance, moving variance, standard deviation, moving standard deviation, differential, derivative, skewness, kurtosis, volatility).

8. CONCLUSION AND FUTURE WORK

In this paper we presented an approach for social media mining based on a pipeline that implements observing, enriching, storing, modeling and presentation techniques. A novel tool TwitterObservatory that allows observing, searching, analyzing and presenting social media has been introduced.

The developed software components enable monitoring of social media stream including enrichment and storing of the data.

The future work will be based on implementing additional functionalities for social media mining pipeline and on developing extensive modeling and nowcasting functionalities for social media and external datasets.

9. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and SYMPHONY EU FP7 project (Contract no: 611875).

References

[1] Sharifi, B., Hutton, M.-A., & Kalita, J. (2010). Summarizing Microblogs Automatically. NAACL HLT 2010.

[2] Potthast, M., & Becker, S. (2010). Opinion Summarization of Web Comments. ECIR 2010.

[3] Lee, Y., Jung, H.-Y., Song, W., & Lee, J.-H. (2010). Mining the blogosphere for top news stories identification. SIGIR 2010.

[4] Irani, D., Webb, S., & Pu, C. (2010). Study of Static Classification of Social Spam Profiles in MySpace. ICWSM 2010.

[5] Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing Microblogs with Topic Models. ICWSM 2010.

[6] Lampert, A., Dale, R., & Paris, C. (2010). Detecting Emails Containing Requests for Action. NAACL HLT 2010.

[7] Sriram et al, B. (2010). Short text classification in twitter to improve information filtering. SIGIR 2010.

[8] Efron, M. (2010). Hashtag retrieval in a microblogging environment. SIGIR 2010.

[9] Rupnik, J., Muhič, A., & Škraba, P. (2012). Multilingual Document Retrieval through Hub Languages. SiKDD 2012.

[10] O'Connor et al, B. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM 2010.

[11] Štajner, T., Novalija, I., & Mladenić, D. (2012). A service oriented framework for natural language text enrichment. Informatica Journal, 34:3, 307-313.

[12] Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. Event Registry – learning about world events from news, In Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion.