A Rule-Based System for Monitoring of Microblogging Disease Reports

Wojciech Lukasiewicz, Kia Teymourian and Adrian Paschke

Freie Universität Berlin, Institute for Computer Science, AG Corporate Semantic Web
{wojlukas,kia,paschke}@inf.fu-berlin.de

Abstract. Real-time microblogging messages are an interesting data source for the realization of early warning systems that track the outbreaks of epidemic diseases like seasonal or pandemic influenza. Microblogging monitoring systems might be able to detect disease outbreaks in communities faster than the traditional public health services. The realization of such systems requires a message classification approach that can distinguish the user messages which concern diseases from other unrelated messages. The existing machine learning classification approaches have some difficulties due to the lack of a longer history-data-based learning curve and the short text length of the messages. In this paper, we present a demonstration of our rule-based approach for classification of disease report messages. Our system is built based on the extraction of disease-related named entities.The type identification of the recognized named entities using the existing knowledge bases helps our system to classify a message as a disease report. We combine our approach with further text processing approaches like term frequency calculation to improve the effectiveness of the detection. Our experimental results show that the presented approach is capable of classifying the disease report messages with acceptable precision and recall. [1]

1 Motivation

People from all around the world use microblogging services on a daily basis and send messages, among others, about their current health condition. Those behaviors and tools on the Web provide us with the perfect technological-sociological background to develop a real-time disease and epidemic outbreak surveillance system. It would monitor the Twitter message data stream, decide if certain posts can be considered disease reports and cluster the appropriate ones based on the sender's geographic whereabouts.

To motivate the underlying problem which we address, we present a couple of example messages which come from a Twitter data stream:
,,This feeling sick is starting to bore me now. #headaches #fever piss right off"
,,Ew I think I have the worst fever in the book of fevers"
are messages that should be considered disease reports, whereas
,,I have bieber fever. justinbieber #BELIEVEtour"
,,my parents are arguing about saturday night fever is before or after grease um"

[1] This work has been partially supported by the "InnoProfile-Transfer Corporate Smart Content" project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder-Entrepreneurial Regions. Special thanks to Gary Ng for the implementation of the demonstration GUI.

are messages that definitely should not be considered disease reports.

In this paper, we present a demonstration[2] of our rule-based approach that can distinguish the disease reports from other microblogging messages containing the disease-related keywords.

Typical obstacles for classifiers of microblogging messages using standard machine learning approaches are:

Short Text Messages: Due to the length of the microblogging messages, the established text classification approaches, such as *Naïve Bayes*, turn out not to be very effective. Given the fact that the microblogging messages are rather short (140 characters is the Twitter limit), the training data set would have to be enormous for the algorithm to be able to distinguish the messages properly.

Term Frequency Equals Document Frequency: Again because of the length limit the terms are most often used only once in the messages which might make this numerical statistic not very meaningful.

Lack of the Learning Curve The microblogging messages are not per default labelled regarding their disease relevance and the classifier does not get any feedback during its lifetime unless the messages would be marked manually – in the contrary to, e.g., spam filters, where the users help the system to learn and improve it by saying that this email is or is not spam. As a result there is no relation between experience of the classifier and its performance.

Specific Language The colloquial Internet and microblogging diction differs from high language standards. This slang is often influenced by current offline events and thus constantly evolves requiring re-training the classifier.

The above mentioned problems clearly show that the existing text classification approaches cannot achieve high precision in the classification of the microblogging messages. Therefore we propose a new, content-based approach which performs the analysis and classification of short, specific text messages.

2 Rule-Based Content Analysis of Microblogging Messages

A classifier is required that can assign the microblogging messages to one of three classes: disease report (***DR***), no disease report (***NDR***) and possible disease report (***PDR***, which means that the poster's intentions were not clear).

We propose a rule-based classification that can classify each message by triggering multiple rules which are applied in a sequential order. A specific score is assigned to each rule and in case that it matches when analyzing a message m, we add the rule score to the overall score of m. At the end, we compare the calculated score to the fix *thresholds* and decide which class should a message belong to. The thresholds, a pair (t_n, t_p), should be interpreted as follows:

$$\begin{aligned} \boldsymbol{DR} &= \{m \mid score(m) \in (-\infty, t_n)\} \\ \boldsymbol{PDR} &= \{m \mid score(m) \in [t_n, t_p]\} \\ \boldsymbol{NDR} &= \{m \mid score(m) \in (t_p, \infty)\} \end{aligned}$$

[2] Our demonstrator is hosted at: `http://www.mi.fu-berlin.de/en/inf/groups/ag-csw/Research/Demos/`

Our rules are grouped in several categories. Provided that a message contains a disease-related keyword, it is processed by the following content analysis rules:

Category-1: Rules Based on Frequent Words: In the first step, we analyze the messages considering the most common words appearing in the disease report candidates.

We have set up a service which filters the Twitter stream based on the set of predefined keywords and stores the messages locally. We collected over 300 million candidates for disease reports between Jan 16^{th}, 2013 and Aug 17^{th}, 2013 and created taxonomies for each keyword by summing up the term frequencies of all words occurring in the messages containing this keyword. The stop words are removed and the remaining tokens are stemmed.

We analysed the lists of frequent words and manually assigned a classification sentiment for every *keyword* $\rightarrow$ *word* collocation. For this we used the Twitter search engine to obtain sample messages containing both words and, based on the context, assigned values from the range $[-5, \ldots, 5]$, with -5 meaning a definitely disease-related and 5 a definitely not disease-related collocation. Having that, we find all keywords in an analysed message and look for sentiments in the corresponding taxonomies. This way we calculate the *disease score* of a single message.

Category-2: Rules based on the Types of Detected Named Entities: We use DBPedia Spotlight [8] for the extraction of semantic resources from the text. Other existing public services like AlchemyAPI, OpenCalais and Zemanta cannot be used because of heavy processing load of the message stream. We use an internal mirror of the DBpedia and the DBpedia spotlight on a cluster of hosts. After the identification of the semantic sources, we query their types, like *rdf:type* (dbpedia-owl:Disease), *dbpedia-owl:type* to check the type of the recognized semantic concepts. This approach lets us derive more meaning from the tweets, e.g., take disease-related words from outside our taxonomies or the synonyms into account. Let us consider the following tweet: ,,*found out that I need surgery on the first week of holidays and I had glandular fever and Ross river fever in the past 6 months.Not happy*" In this example the message that would not be classified as disease report without the help of semantic concepts. The word *happy* would be the only match in the sentiments list (see **Category-3**) and no collocation from the taxonomy of *fever* would be found. DBPedia Spotlight however annotates *"glandular fever"* (dbpedia:Infectious_mononucleosis) and *"Ross river fever"* (dbpedia:Ross_River_fever) as diseases.

Category-3: Rules based on General Mood of Messages: To improve the classification precision, we apply further rules to extract the sentiment of the message.

We use the word list by Hansen et al. [3] to look up the message's tokens and thus calculate its *general mood*. Unfortunately, the disease report candidates generally tend to have a negative score calculated by this rule because of the diction that consists mainly of words with negative sentiment.

Category-4: Other Rules: The microblogging messages are often enriched with *emoticons* (also called *smileys*, conventional symbols for expressing emotions)

to put emphasis on the author's mood. For that reason we check the tweet for the presence of smileys.

Furthermore, when people tell about their illness when having a fever, they sometimes mention its height. We could assume that a number found in a message (restricted to ranges which correspond to fever temperatures, both in Fahrenheit and Celsius scale) means with high probability that it concerns a disease.

3 Related Work

The studies by Signorini [10] and Chew [2] exemplarily explain how massively the Twitter data stream is influenced by current real life events. Among many other triggers, like for example the American Idol contest, the authors of these publications took a look at the 2009 *Swine Flu* (H1N1) outbreak.

Stewart et al. [11] present different approaches of health-related Twitter surveillance. Discussing Early Warning, as well as Outbreak Control and Analysis Systems, they introduce several biosurveillance algorithms and techniques (Khan [6]; Hutwagner et al. [5]; Basseville et al. [1]) and use them to analyse the crowd's behavior during the 2011 *enterohemorrhagic Escherichia coli* (EHEC) outbreak in Germany. Their main focus was to detect aberration patterns when the observed variable (here: tweets containing the ,,EHEC" keyword but more generally: the number of tweets regarding diseases that do not reveal seasonal patterns) exceeds an expected threshold value. They used four different biosurveillance algorithms for early detection, each one of which proved to be at least one day faster than well-established early warning systems, like e.g. The Early Warning and Response System of the European Union[3], MedISys[4] or ProMED-mail[5].

Lampos et al. [7] investigate the 2009 *Swine Flu* outbreak. Hu et al. [4] aim to cluster Twitter messages by topic and extract meaningful human-readable labels for each cluster. They decompose the unstructured text using NLP and then transform the syntactic feature space (parse [sub]trees) into semantic feature space using WordNet and Wikipedia. Saif et al. [9] propose to add the semantic concepts of extracted entities as additional features for sentiment analysis.

Hansen et al. [3] analysed which tweets attract the biggest attention and are most likely to be retweeted. As a part of this publication one of the authors, Finn Årup Nielsen, prepared a list of 2477 English words rated for valence with an integer between minus five (negative) and plus five (positive).

The main difference of our approach with the existing approaches are that our approach is based on the collected heuristics, manually added sentiment scores and the semantic types of the extracted entities in the messages. Its advantage is that it does not require to have a learning loop (only an update of the keyword list / taxonomies) and has no shortcomings when applied to the very short messages.

[3] **EWRS**: `https://ewrs.ecdc.europa.eu/`
[4] **MedISys**: `http://medusa.jrc.it/`
[5] **ProMED-mail**: `http://promedmail.org/`

4 Conclusion

Conducted experiments show that our approach is able to detect the disease reports from the collection of disease report candidates with an acceptable precision and recall.

On this basis, we demonstrate a real-time system for classification of disease messages from mass of microblogging messages. Our system is a live service connected to the Twitter stream that receives messages and visualizes the disease reports on a map provided that they were enriched with the geographic whereabouts of the sender (sent from a mobile device).

Given that our system could receive a complete data stream, after some time it could extract anomalies based on the daily/weekly number of messages originating from a certain area. Such a tool could be a great complement to the well-established health surveillance systems.

References

1. Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
2. Cynthia Mei Chew. Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak. Master's thesis, University of Toronto, 2010.
3. Lars Kai Hansen, Adam Arvidsson, Finn Arup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. *CoRR*, abs/1101.0510, 2011.
4. Xia Hu, Lei Tang, and Huan Liu. Enhancing accessibility of microblogging messages using semantic knowledge. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2465–2468, New York, NY, USA, 2011. ACM.
5. L. Hutwagner, W. Thompson, G. M. Seeman, and T. Treadwell. The bioterrorism preparedness and response early aberration reporting system (ears). *J Urban Health*, 80:i89–96, Jun 2003. Hutwagner, LoriThompson, WilliamSeeman, G MatthewTreadwell, TraceeUnited StatesJournal of urban health : bulletin of the New York Academy of MedicineJ Urban Health. 2003 Jun;80(2 Suppl 1):i89-96.
6. Sharib A. Khan. Handbook of biosurveillance, m.m. wagner, a.w. moore, r.m. aryel (eds.). elsevier inc. isbn-13: 978-0-12-369378-5. pages 380–381, 2007.
7. Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, pages 411–416. IEEE Press, June 2010.
8. Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.
9. Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508–524, Berlin, Heidelberg, 2012. Springer-Verlag.
10. Alessio Signorini. Social web information monitoring for health, 2009.
11. Avaré Stewart and Ernesto Diaz. Epidemic intelligence: for the crowd, by the crowd (Tutorial). In *Proceedings of the 12th international conference on Web Engineering*, ICWE'12, pages 504–505, Berlin, Heidelberg, 2012. Springer-Verlag.