Observatório da Dengue: Surveillance based on Twitter Sentiment Stream Analysis*

Ismael S. Silva, Janaína Gomide, Glívia A. R. Barbosa, Walter Santos, Adriano Veloso, Wagner Meira Jr., Renato Ferreira

Dept. of Computer Science – Federal University of Minas Gerais – Belo Horizonte – MG – Brazil

{ismael.silva, janaina, gliviaangelica, walter, adrianov, meira, renato}@dcc.ufmg.br

Abstract. *Social media channels became an option to measure the overall sentiment expressed by the online population about diverse subjects. However, performing sentiment analysis in these channels imposes several challenges to current classification-based techniques, which are typically used to perform sentiment analysis. Challenges include the need of a large amount of labeled data and constant training update in order to track concept drift and to provide up-to-date analysis. The core of the application to be described in this paper is an algorithm which takes as input a small training seed and produces classification models on-the-fly. This strategy keeps the model always up-to-date, better reflecting the sentiments being analyzed. By considering Twitter as a continuous source of epidemic information, our tool tracks the sentiment of the Brazilian population about dengue through the Twitter message stream, enabling real-time surveillance and arming public health agencies with the ability to perform more effective counter measures.*

1. Introduction

Knowledge discovery from web data offers strategic opportunities for various areas (e.g., epidemics surveillance, marketing, elections) [Chew and Eysenbach 2010]. In many cases, this discovery occurs through the analysis of the sentiments expressed by the online population on applications in which the data arrival process follows the data stream model, requiring the analysis to be performed in real time.

Currently, data stream analysis is identified as one of the biggest challenges faced by researchers in areas such as machine learning and data mining [Bifet 2010]. Among these challenges we may highlight: (1) classifiers have to operate with limited memory, processing time and labeled data and (2) there are constant changes in the data distribution and patterns related to labels (i.e., concept drift). Thus, methodologies to update classification models on real-time become necessary. Motivated by the above challenges, our aim is to provide a tool to monitor sentiments expressed by population about dengue, through the Twitter message stream. The algorithm used to classify messages with regard to the sentiment they express is presented in [Silva et al. 2011] and is described in Section 3.1, this algorithm provided gains that range from 7% to 58% on experiments of sentiment stream analysis.

* Tool available in http://goo.gl/y1BPB and http://goo.gl/mbck6. This research was sponsored by UOL (www.uol.com.br) through its UOL Bolsa Pesquisa program process number 20110215215201, CNPq, Capes, Fapemig and Inweb - the Brazilian National Institute of Science and Technology for the Web.

Twitter was used because it is one of the communication channels with the highest rise in recent years. In this communication channel millions of users are encouraged to express their opinions and to share information constantly. Further, the message generation model follows the data stream paradigm.

We perform sentiment analysis related to dengue, since this is a disease that affects much of the population and represents a major challenge for public health in Brazil [Gomide et al. 2011]. Monitoring what has been published about the disease (e.g., statements of patients and opinions) may bring great benefits to the population (e.g., dengue surveillance), as shown in [Gomide et al. 2011] the volume of messages reporting a sentiment of personal experience about dengue have high correlation with disease cases in real word.

Increasingly, tools for sentiment analysis become more relevant. Although some tools are already available, the application proposed here stands out from existing ones. It displays in real time the sentiment expressed by people in Twitter about a serious disease and provides online adaptation of training set to deal with concept drift. Our tool is part of the project "*Observatório da Web*" (http://www.observatorio.inweb.org.br), a project that comprises several features. However, in this paper, we focus on those that use sentiment stream analysis to perform dengue surveillance.

2. Related Work and Systems

There are many approaches that focus on data stream analysis using automatic classification techniques. In order to limit human intervention, automated alternatives that use distant supervision approaches (emoticons or tags) were proposed in [Go et al. 2009] but they are prone to error by definition and unable to capture other types of sentiments for which no emoticon (or tag) is associated with. Other alternatives to address the annotation burden include active and semi-supervised learning approaches [Wu et al. 2006, Ribeiro et al. 2010]. However, these approaches must employ procedures that would be too complex to be used on real-time data stream analysis.

Among the most representative tools for sentiment analysis, we may highlight Tweetfeel (http://goo.gl/ghMe6), Twendz (http://goo.gl/t9oLb) and Twitrratr (http://goo.gl/Jd2Og). Moreover, there are SAS Analytics Social Media (http://goo.gl/aIyZn) and Lexalytics (http://goo.gl/1aAig) that are tools able to analyze data to quantify interaction among traditional campaigns and social media activity. Sentic (http://goo.gl/xgLq0) and SinoBuzz (http://goo.gl/csoaT) offer tools that help companies to pinpoint the effect of specific issues on customer perceptions.

The proposed work differs from the aforementioned ones because it presents a tool based on the approach showed in [Silva et. al. 2011] to the classification of expressed sentiments by the Brazilian populations through Twitter, by adapting the classification model to the effects of *concept drift* without human intervention. Moreover, the topic accompanied (i.e., dengue) makes the tool a public utility service since it is capable of pointing out aspects such as: (1) disease outbreaks and (2) areas that need more investment in campaigns.

3. Sentiment Stream Analysis System

In this section we present the proposed tool with regard to the algorithms that support it, characteristics of the analyzed data and the visualization interfaces employed.

3.1 Effective Online Sentiment Stream Learning

In order to monitor dengue and the perception of people with regard to this topic, we built a tool to extract information about the disease from the Twitter message stream. Aiming at reducing the human effort in labeling messages for training and better capturing the variation in stream, we used the algorithm described in [Silva et. al. 2011]. This algorithm showed efficiency on very dynamic scenario adjusting the training set over sentiment stream. The main parts of this algorithm are summarized next.

Sentiment Scoring. We measure the likelihood of each sentiment using association rules [Agrawal et. al 1993]. An association rule is a structure like $\mathcal{X} \rightarrow s_i$ where the antecedent $\mathcal{X}$ is a set of terms (i.e., a termset), and the consequent s_i is the predicted sentiment label. The domain for $\mathcal{X}$ is the vocabulary of the training set, referred to as $\mathcal{D}$. The cardinality of rule $\mathcal{X} \rightarrow s_i$ is given by the number of terms in the antecedent, that is $|\mathcal{X}|$. The confidence of rule $\mathcal{X} \rightarrow s_i$, denoted as $\theta(\mathcal{X} \rightarrow s_i)$, is the conditional probability of sentiment label s_i given the terms in $\mathcal{X}$, that is $\theta(\mathcal{X} \rightarrow s_i) = \frac{\sigma(\mathcal{X} \cup s_i)}{\sigma(\mathcal{X})}$.

Once the classification model is extracted from $\mathcal{D}$, rules are used to score sentiments of messages that come later. Each extracted rule $\mathcal{X} \rightarrow s_i$ is a vote given to sentiment s_i. Given a new message (t), a rule $\mathcal{X} \rightarrow s_i$ is only considered as a valid vote if all terms in $\mathcal{X}$ are present in t. Votes for each sentiment (s_i) are averaged according to their θ values, giving the score for sentiment label s_i, with regard to message t, this score is denoted as $s(t, s_i)$. The scores are normalized, as showed in Equation 1. The scoring function estimates the likelihood of sentiment label s_i being the implicit sentiment in message t.

$$\hat{p}(s_i|t) = \frac{s(t,s_i)}{\sum_{j=0}^{k} s(t,s_j)} \quad (1)$$

Online Rule Extraction with Data Projection. Rules are extracted online, by projecting the training data on a demand-driven basis according to terms in t [Veloso and Meira, 2011]. This ensures that rules are extracted in polynomial time, what in practice makes the procedure very efficient.

Training Data Inclusion. In order to adapt the classification model accordingly, it is mandatory to gather the most current information emerging in the stream. Latest, most current training messages may be obtained by exploiting the predictions performed using the scoring function shown in Equation 1. These predictions may be used to assign a label to messages, generating labeled messages. We use a threshold δ_{min} $(0.5 \leq \delta_{min} \leq 1.0)$ and we include into $\mathcal{D}$ messages for which $\hat{p}(s_i|t) \geq \delta_{min}$.

The idea is to use δ_{min} to indicate the minimum reliability necessary to regard a labeled message as a correct one, and, therefore, to include it into the training data $\mathcal{D}$. Intuitively, if reliable predictions are indeed correct ones, then the training data will be continuously augmented with novel training information, keeping the training data up-to-date as the stream evolves.

Sub Judice Strategy. Some predictions are not reliable enough, given certain values of δ_{min} (i.e. $\hat{p}(s_i|t) < \delta_{min}$). An alternative is to abstain from using such doubtful predictions as the classifier does not have enough evidence for a reliable judgement, that is, we do not use the corresponding labeled messages for model

building and keep them *sub judice*. As new reliable labeled messages are included into $\mathcal{D}$, new evidence is exploited, hopefully increasing the reliability of the labeled messages that were previously hold and possibly releasing them, (i.e., when a labeled message is included into $\mathcal{D}$, the classifier re-evaluates all messages that are *sub-judice*). Messages are kept *sub judice* for a certain period of time (e.g. minutes, hours). After this period, all messages are necessarily processed.

3.2 Twitter Message Stream

Twitter message stream presents particular challenges, such as (1) cultural, linguistic and geographic factors that make the vocabulary nearly unpredictable and (2) the large data volume. For instance, in the period from 12/01/2010 to 04/18/2011 we collected 437,000 messages containing the word "dengue".

Since our application follows messages related to dengue, these were categorized according to the labels proposed in [Gomide et al. 2011]. The authors classified the messages as: (1) **Personal Experience** (e.g., "Did you know that I've had dengue? "), (2) **Irony** (e.g., "Practice sport, because stagnant water in your tiers may cause dengue"), (3) **Opinion** (e.g., "Very cool this campaign against dengue "), (4) **Information** (e.g., "Dengue virus type 4 in circulation"), (5) **Marketing** (e.g., "Everyone against dengue!"). Given the objective of the proposed application, we classify and also follow the evolution of each category as the messages arrive.

To create the initial training set for the algorithm, we use an active sampling approach [Ribeiro et al. 2010]. Then 100 messages were selected and these were labeled by two people. This training set is the same as used in [Gomide et al. 2011].

3.3 System Architecture

Architecture of our system is composed by (1) two redundant collectors, (2) a set of filters, (3) three *MongoDB* database servers and (4) one MySQL server.

The two collectors work simultaneously to ensure that most messages will be collected from Twitter streaming API. The filters are applied for language detecting, location identifying and URL expansion. After filtered, the messages are stored in a *MongoDB* database (a scalable and high-performance solution) to future queries. This database is replicated on three servers to prevent data loss. Then the messages are processed (e.g., by sentiment classifier) and aggregated (e.g., by time or localization). After that, the result is stored into a MySQL server.

This architecture has been shown robust and scalable to extract, transform and load of data streams. Some visualization, available on our system, are described next.

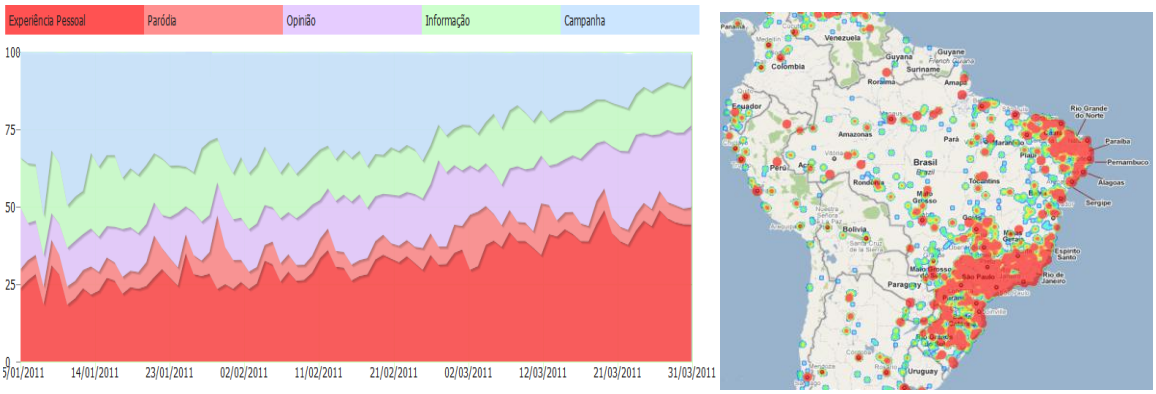
3.4 Visualization Tool for Tracking Dengue Evolution

To provide flexible and clear visualization, the proposed application offers different forms of data presentation, so that any web user can extract and analyze information.

The graph in Figure 1(a), which corresponds to one of the possible data views, shows the evolution of sentiments expressed by Twitter users, on the topic of interest distributed in the categories previously mentioned. In this chart the user can see which category (i.e., personal experience, opinion, irony, marketing or information) predominates in a given period. From this view, the users may perform several analysis,

for instance, one could check whether the increase of campaign messages occurred at the same time that the messages portraying personal experience decreased.

The second available feature is the geo-referenced data display. The application offers the user the possibility to check the intensity of sentiment categories related to dengue, addressed in each region of the country. An example of this view is illustrated in the heat map (Figure 1(b)), from where it is possible to check in which regions of the country there are more reports of personal experience in a month. On this map, the hot area (i.e. red) indicates more cases of disease. This visualization can indicate, for instance, regions with disease outbreaks,since, as showed by [Gomide et. al 2011] the volume of messages about personal experience with dengue have high correlation with statistics reported by the Brazilian Ministry of Health.



(a) Temporal evolution of sentiments categories

(b) Heat map of personal experiences per month

Figure 1. Aggregated Visualization by Time and Region Dimensions

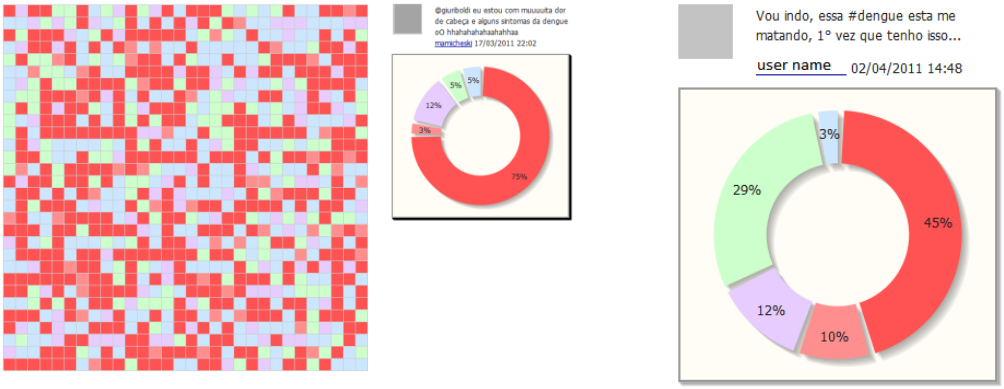
Another possibility of information can be viewed in Figure 2. Through this view the user can visualize a sample of the last collected messages and their respective categories. Clicking on a message the user can enjoy a radar chart (Figure 2(b)) that shows the proportion of each category contained in the message (i.e., assigned likelihood to the message belongs to the category), furthermore, it shows the message text. This data displaying option is interesting because apart from allowing verification of message content the user can also check which category predominates in the discussions at the last moments.

4. Conclusion and Future Work

Increasingly, Web 2.0 has proved to be a communication channel in which content producers and consumers mix and interact, thereby establishing new ways to create, organize, share and use knowledge. One of the generation sources of this content are the media social systems (e.g., Twitter, Facebook), where web users are invited and encouraged to share content and express their opinions and sentiments.

In this context, it is necessary to propose tools and techniques that provide the knowledge discovery and analysis from the data stream produced by the population in these applications. We present a tool for monitoring the sentiments expressed by the Brazilian population about dengue in the Twitter message stream.

This tool can be used by any web user with the purpose of acquiring knowledge about the disease state in Brazil. In addition, government agencies can use it to identify areas where disease outbreaks are occurring and through this information, conduct further investigation. In the future we intend to include warning panels of disease outbreaks and new visualization options.



(a) Panel with random sample of the last messages (b) Likelihood of each sentiment

Figure 2. Sampling of last collected messages

References

Agrawal, R., Imielinski, T., and Swami, A. (1993). "Mining association rules between sets of items in large databases". In SIGMOD, pages 207–216.

Bifet, A. (2010). Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams. In Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams, Albert Bifet (Ed.). IOS Press, 1-212.

Chew, C. and Eysenbach, G. (2010). "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak". PLoS ONE 5, 11: e14118.

Go, A., Huang, L. and Bhayani, R. (2009). "Twitter sentiment classification using distant supervision". In CS224N Project Report.

Gomide, J., Veloso, A., Meira Jr, W., Benevenuto, F., Almeida, V., Ferraz, F. and Teixeira, M. (2011). "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter". In WebScience.

Ribeiro, M. T. C., Veloso, A., Meira Jr, W., Pappa, G. L., Cherchiglia, L., Teixeira, L. V. and Brunoro, G. (2010). "Mining Twitter for Feelings and Opinions". In SBBD.

Silva, I. S., Gomide, J., Veloso, A., Meira, W., and Ferreira, R. (2011). "Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection". In SIGIR, pages 475-484.

Veloso A., Meira. W. Jr. (2011). "Demand-Driven Associative Classification". SpringerBriefs in Computer Science, March, 2011

Wu, S., Yang, C. and Zhou, J. (2006). "Clustering-training for Data Stream Mining". In ICDM - Workshop. IEEE Computer Society, pages 653-656.