Mining Twitter for Feelings and Opinions*

Marco Túlio C. Ribeiro[1], Adriano Veloso[1], Wagner Meira Jr.[1], Gisele L. Pappa[1], Letícia Cherchiglia[1], Leonardo Vilela Teixeira[1], Gustavo Brunoro[1]

[1] Dept. of Computer Science – Federal University of Minas Gerais – Brazil

{marcotcr,adrianov,meira,glpappa,letslc,vilela,brunoro}@dcc.ufmg.br

Abstract. *Social media channels are challenging how we communicate, with whom we communicate, and perhaps most fundamentally how people express their opinions and feelings about topics and entities in the news, products and services. As a result, online opinion has turned into a kind of virtual currency for business analists, marketers, and politicians. In this torrent of opinionated content provided by social media channels, what makes Twitter a unique channel for analysis is that users tend to be very succinct – the 140-character limit is very restrictive, and people tend to be very forthcoming with their opinions. However, Twitter also has a serious drawback – users are able to enter free text, making the vast number of linguistic approaches which are based on the use of controlled vocabulary not effective in practice. As an alternative, in this paper we present a highly effective machine learning based system for mining opinions on Twitter. Our system does not assume a controlled vocabulary, but training examples (i.e., annotated tweets) must be provided to the learning algorithm. In order to limit human intervention, we developed an active sampling strategy which is able to reduce exponentially the number of examples needed by the learning algorithm. In this paper we describe our system and illustrate its features with an analysis featuring the candidates for the 2010 Brazilian presidential elections. Our system is part of the project "Observatório da Web".*

1. Introduction

Sooner or later a business analyst will face questions such as "what do customers think about my product?", or a politician will face questions such as "what do people think or feel about my political proposals?". Ordinary people, which are usually the main target of such analysis, may also be interested in the common opinions and feelings of other people. Therefore, ordinary people may face questions such as "do other people like this product?" or "how is the popular opinion on the presidential candidates evolving?". The rise of Web social media channels has fueled scientists with large amounts of opinions, emotions, and other forms of online expressions, and this fast-growing mountain of Web data is the basic necessary resource for analysing opinionated content on a large scale. However, dealing with this information torrent is impossible with old, manual methods. As a result, an emerging field known as sentiment analysis is becoming more and more popular. Sentiment analysis involves classifying opinions present in text into polarity categories like "positive" or "negative" often with an implicit category of "neutral".

*This research is partially funded by InWeb - The National Institute of Science and Technology for the Web (MCT/CNPq/FAPEMIG), and by the authors's individual research grants from CAPES, CNPq, and FAPEMIG.

Amongst all social media channels available in the Web, Twitter has an unique property which facilitates sentiment analysis: 140 characters do not provide users enough space to explain, elaborate or get distracted from their main point. On the other hand, users may enter free text, reducing the effectiveness of methods based on linguistic analysis, which relies on controlled vocabularies. Furthermore, sentences (or tweets) with lots of sarcasm and irony are also commonly posted by Twitter users.

In this paper we present a sentiment analysis system which does not rely on controlled vocabularies - instead, it employs machine learning algorithms in order to differentiate positive and negative patterns in the text. At first, our system has to be trained, in order to identify sentiments correctly. Our training step employs association rules [Veloso et al. 2006] that are extracted from tweets annotated by our team, which basically means that every tweet in the training dataset is annotated by humans first. A typical approach for producing the training dataset is based on random sampling, that is, tweets are randomly selected and annotated. Although simple, this approach usually requires a large number of annotations, because it is common to annotate similar, or redundant tweets (i.e., tweets that share textual content). In order to limit human effort while producing the training dataset, our system adopts an active sampling approach, which selects the next tweet to be annotated based on the tweets already annotated. More specifically, the next tweet to be annotated is the one which will provide the highest information gain. This active sampling approach reduces the number of annotations needed by the system exponentially.

The system is trained on a daily basis, with few annotated tweets (about 20 new examples are annotated each day). Once trained, the system automatically predicts the polarity of a large number of tweets posted in the last 24 hours. This learning strategy not only avoids the use of controlled vocabularies, but also enables the system to detect tweets with irony and sarcasm (tweets with these idiomatic expressions are easily detected by humans during annotation, and the corresponding polarities are provided to the system during training). Furthermore, casual users of our system can give feedback about the predictions performed by the system. Specifically, users can agree or disagree with the predictions, and this collaborative feedback enables our system to improve with time.

We illustrate the features of our system using an important evaluation scenario: monitoring Twitter for trends in opinions about the candidates for the 2010 Brazilian presidential elections. Our system is able to capture opinions on a trend line so that users can identify sudden positive or negative spikes about the candidates. However, the formal definiton of opinion – a private state which is not open to objective observation or verification [Quirk et al. 1985] – makes an objective evaluation of this kind of system specially challenging. According to reports provided by the Amazon Mechanical Turk[1], humans only agree 79% of the time, meaning that even when the accuracy of such kind of system is well below perfect, statistically, it can be thought of as more accurate when compared to human analysis. In other words, usualy the automated analysis performed by systems like ours can be considered almost as good as human analysis.

[1] `http://aws.amazon.com/mturk/`

2. Related Systems

There are several sentiment analysis systems available. Some rely totaly on automated approaches [O'Connor et al. 2010], and some use a hybrid human-machine approach [Quinn et al. 2010, Hiroshi et al. 2004]. These systems consider word positioning and detect negators(like 'NOT', for example). The simplest systems work by scanning keywords to categorize a sentence as positive or negative, based on a simple binary analysis ("love" has a positive conotation, and "hate" is negative). These systems rely on a controlled vocabulary of selected keywords, and thus they fail to capture the subtleties present in human language: irony, sarcasm, slang and other idiomatic expressions.

The Financial Times recently introduced *Newssift*[2], an experimental system that tracks sentiments about business topics in the news, coupled with a specialized search engine that allows users to organize their queries by topic, organization, place, person and theme. Using *Newssift*, a search for "WalMart" reveals that the recent sentiment trend about the company is running positive by a ratio of slightly better than two to one.

Lexalytics[3] and *SAS Social Media Analytics*[4] are well established products which are able to analyze conversation data, identify the reputation and image of a brand, and to quantify interaction among traditional media/campaigns and social media activity. *Sentic*[5] and *SinoBuzz*[6] offer tools that help companies pinpoint the effect of specific issues on customer perceptions, helping them respond with appropriate marketing and public relations strategies. For Web surfers, sentiment analysis can be provided by lightweight tools like Tweetfeel[7], Twendz[8] and Twitrratr[9]. These sites allow users to analyze the opinion in tweets about particular topics. The features provided by these three systems is essentialy the same – the user searches for a specific topic, and a list of positive/neutral/negative tweets is returned to the user. Many of these tools are not open source, so we were not able to evaluate their performance and the way they work. However, some are clearly simpler, and seem to be based on the binary analysis we described earlier.

3. Opinion Mining System

Performing sentiment analysis on Twitter is definitelly challenging, because each post (or tweet) may point to many cultural factors and linguistic nuances, making it difficult to turn a string of written text into a simple pro or con sentiment. In this section we describe the basic steps of our sentiment analysis system.

3.1. Active Sampling

Our system requires examples in order to distinguish characteristics of positive and negative tweets surrounding a topic or entity (i.e., a candidate). These examples are provided by humans, as shown in Figure 1. For each entity (i.e., a candidate), the annotator has

[2] http://www.newsssift.com/
[3] http://www.lexalytics.com/
[4] http://www.sas.com/software/customer-intelligence/social-media-analytics/
[5] http://www.cs.stir.ac.uk/~eca/sentics/
[6] http://www.sinotechgroup.com.cn/index.php/product
[7] http://www.tweetfeel.com/
[8] http://twendz.waggeneredstrom.com/
[9] http://twitrratr.com/

Figure 1. Annotating tweets.

four options: (1) the polarity of the tweet is positive, (2) negative, (3) neutral, or (4) not applicable (this fourth option is necessary in order to protect the system against failures during the named entity recognition step, which, for brevity, follows an approach similar to [Wang et al. 2009]). The same tweet is analyzed by at least three annotators, and the positive and negative annotations will correspond to entries in the training dataset.

The simplest approach is to randomly select the tweets to annotate. Although simple, this approach leads to a large number of annotations, because there is a large number of tweets with similar textual content (i.e., re-tweets, copies etc.). The annotation of similar tweets provides redundant information, that is, there is no information gain. As a consequence, in order to acquire enough information, a very large number of tweets needs to be manually annotated, incurring in costly human effort.

A more effective approach is to actively select the tweets to be annotated. The intuition is simple: the next tweet to be annotated should be the one containing as many words not included in any tweet already processed as possible. In this way, the next tweet to be annotated will provide the largest number of unknown word. Popular words (like stopwords and candidates' names) will appear in almost every tweet, so they will not be relevant when selecting which tweet should be annotated next. As a result, a great portion of the vocabulary is quickly covered by the annotated tweets. In fact, we found that the number of annotations reduces exponentialy, making the system extremely practical [Mitchell 1997].

3.2. Demand-Driven Polarity Detection

Instead of classifying a tweet as positive or negative, our system computes the degree of positiveness for each tweet. The degree of positiveness ranges from 0 to 1, and it can be viewed as the likelihood of being positive. Higher/Lower values of positiveness indicate that the tweet is more positive/negative, while values around 0.5 indicate that the tweet is neutral. Our system employs association rules in order to calculate the degree of positiveness. Specifically, given the training dataset $\mathcal{T}$ and a tweet t, rules of the form $\mathcal{X} \rightarrow$ positive or $\mathcal{X} \rightarrow$ negative (with $\mathcal{X} \subseteq t$) are extracted from $\mathcal{T}$. Rules are extracted on a demand-driven basis, according to [Veloso and Meira Jr. 2007], and we recently proved in [Menezes et al. 2010] that this rule extraction process has polynomial complexity with the number of distinct words in $\mathcal{T}$.

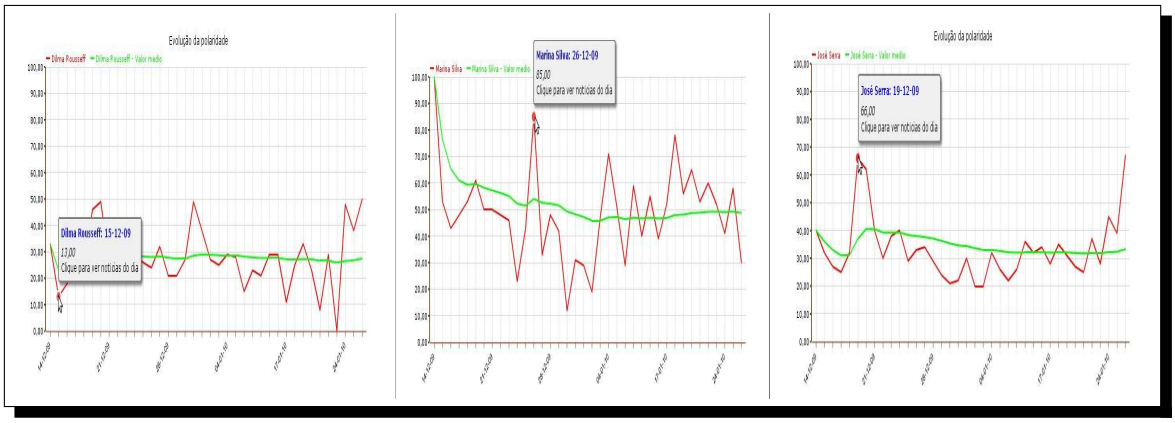

Figure 2. Polarity trend line.

Each rule r is viewed as a vote, which is weighted by its confidence value, denoted by $\theta(r)$ (the confidence is essentialy the conditional probability of the tweet t being positive or negative, given that $\mathcal{X} \subseteq t$). Votes for polarities positive and negative are summed according to Equation 1, and the positiveness is finally given by Equation 2.

$$
\begin{aligned}
s(t, \text{positive}) &= \sum \theta(\mathcal{X} \rightarrow \text{positive}), \text{ and} \\
s(t, \text{negative}) &= \sum \theta(\mathcal{X} \rightarrow \text{negative}), \text{ where } \mathcal{X} \subseteq t \qquad (1) \\
\hat{p}(\text{positive}|t) &= \frac{s(t, \text{positive})}{s(t, \text{positive}) + s(t, \text{negative})} \qquad (2)
\end{aligned}
$$

Figure 2 shows polarity trend lines for Dilma Rousseff, Marina Silva and José Serra. For all these candidates, our system was able to capture sudden positive and negative spikes. The user can click on a point in the trend line, and he will be redirected to the main news published about the candidate in the corresponding date. This allows us to interpret the reasons for the observed spike. For instance, in the leftmost graph, there is a negative spike on 15/december/2009, which happened as a result of an interview given by the candidate Dilma Rousseff in Copenhan, when she said: “The environment is a threat for the sustainable development”.

The graph in the middle shows spikes for Marina Silva. Specifically, on 26/December/2009, all major media channels reported: “Marina Silva leaves PT and joins PV”. Similarly, the rightmost graph shows spikes for José Serra. There is a big spike on 19/December/2009, one day after all major channels reported the resignation of Aécio Neves, who was Serra’s opponent within the party.

3.3. Quality Control

Our system is continuously evaluated in order to assess the quality of the predictions it performs. The evaluation procedure starts with the stratification of the predictions, according to their degree of positiveness. More specifically, the tweets are placed in 10 bins, so that tweets with positiveness between 0 and 0.1 are placed in the first bin, tweets with positiveness ranging from 0.1 to 0.2 are placed in the second bin, and so on. A sample of these bins are given to human evaluators, and they will judge if the tweet is

positive or negative. The ideal scenario is one where only 10% of the tweets placed in the first bin by our algorithm are indeed positive (as evaluated by the human evaluators), 20% of the tweets in the second bin are indeed positive, and so on. This ideal scenario is compared against the empirical scenario, and the divergence between the ideal and the empirical scenario gives the overall quality of our system (the lower the better).

4. Conclusions and Future Work

In this paper we present a new sentiment analysis system, which is part of the project "Observatorio da Web". The basic resource used by our system is a set of tweets, and our system relies on machine learning techniques in order to differentiate tweets with positive connotation from tweets with negative connotation. In order to limit human intervention during annotation of tweets, our system uses an active sampling technique, which reduces exponentialy the number of necessary annotations.

As future work we intend to perform a finer-grained analysis: feature/aspect-based sentiment analysis. This analysis refers to the study of determining the opinions expressed on different features or aspects of entities (i.e., candidates). A feature or aspect is an attribute or a component of an entity (e.g., the proposal for social healthy and public safety). Future work also includes the identification of the most influential opinion holders in Twitter, and to implement features to track Twitter activity during and after the political debates. Further, providing drill-down and roll-up services (i.e., polarity analysis in a country, in regions, in states etc.) are also targets of future work.

References

Hiroshi, K., Tetsuya, N., and Hideo, W. (2004). Deeper sentiment analysis using machine translation technology. In *Intl. Conf. on Computational Linguistics*, pages 494–501.

Menezes, G., Veloso, A., Ziviani, N., Moura, E., Almeida, J., Pappa, G., Lacerda, A., and Gonçalves, M. (2010). Demand-driven tag recommendation. In *European Conf. on Principles of Data Mining and Knowledge Discovery, to appear*.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Intl. Conf. on Weblogs and Social Media*, pages 11–21.

Quinn, A., Bederson, B., Yeh, T., and Lin, J. (2010). Crowdflow: Integrating machine learning with mechanical turk for speed-cost-quality flexibility. In *Human-Computer Interaction Lab Annual Symposium*.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, New York.

Veloso, A. and Meira Jr., W. (2007). Efficient on-demand opinion mining. In *Brazilian Symposium on Databases*, pages 332–346. SBC.

Veloso, A., Meira Jr., W., and Zaki, M. J. (2006). Lazy associative classification. In *Intl. Conf. on Data Mining*, pages 645–654.

Wang, W., Xiao, C., Lin, X., and Zhang, C. (2009). Efficient approximate entity extraction with edit distance constraints. In *SIGMOD Intl. Conf. on Management of Data*, pages 759–770.