Sentiment for Cops and Robbers using Transductive SVM

Guillermo Cabrera
Department of Computer Science
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
gcabrera@cs.utexas.edu

Abstract

The proposed project will explore sentiment analysis on blogs relating to the organized crime domain. We focus on this domain given that blog posts exhibit a great number of named entities and make use of a new lexicon. A semi-supervised learning approach will be taken in order to correctly classify polarity in blog posts; we plan to use accuracy as the metric to evaluate the performance of our model.

1 Introduction and motivation

Blogs have become an integral part of the Internet user experience; users will normally visit and engage by expressing their opinion towards topics (blog posts). Many businesses, politicians and other entities have recognized the importance of gathering this kind of information; which is why sentiment analysis has gained popularity in the past years.

Sentiment analysis has been applied to various domains; however, we have not come across any work on organized crime. We believe this domain has unique properties that sets it apart from other domains and makes it an interesting problem to investigate. As a sample scenario, let us consider the current situation in Mexico, where a dramatic increase in violence due to organized crime activity has led to self-censorship. As a result, people are now turning to blogs as their source of information and sites such as blogdelnarco.com have attracted as many as three million hits per week. Individuals expressing their opinions on blog posts include: Civilians, police and criminals. All these entities interact in the blog by providing very strong opinions and many do so by making heavy use of profanity, colloquial language and other Internet jargon.

We propose a study of the sentiment associated with a small set of entities (army, police, and the president of Mexico) in the organized crime domain. As mentioned above, we claim that blog activity is significantly different than other domains. We plan to make use of a semi-supervised approach by means of a Transductive SVM (TSVM) [1] model to serve in sentiment classification. As a result of our study, we shall report on the accuracy of our model and compare that with existing supervised methods (Support Vector Machines).

In the past, Support Vector Machines (SVM) have demonstrated to outperform naive bayes or maximum entropy classification in terms of relative performance Pang et al. (2002). However, in our problem we have mostly noisy unlabeled data that we would like to use to train our TSVM, as well as exploit co-occurring patterns in text. Therefore, a supervised approach is not an affordable option for our problem, furthermore, given the the heavy use of colloquial language, made up verbs and nouns (ex. *troka* as a way to say truck, instead of the Spanish word *camioneta*), translating into English as is done in Denecke (2008) would not get good results.

Consider the case of the following two sample comments extracted from our dataset:

(1) a la bio a la bao a la bim bom ba, los federales, ra ra ra / hooray for the federales

(2) la p****e pfp solo viene a robar, matar y extorsionar, fuera / the f**** pfp only comes

[1] Also known as a Semi Supervised Support Vector Machine (S3VM)

to steal, kill and extort, out

An individual with knowledge of Spanish language and culture in Mexico might find it trivial to assign a positive label to (1) and a negative to (2), and know that both refer to the same entity (federal police). However, a machine might find it difficult given the heavy use of sarcasm, humor, intense use of profanity, multiple aliases for named entities and the inclusion of a new lexicon.

From a government perspective, the importance of having a list of entities with their associated sentiment would be important as it would give an insight into the feeling towards police or army extortion. One could also imagine that sentiment could be tracked over time for criminal groups or their members, allowing intelligence agencies to analyze possible expansion of certain organized crime groups.

We have collected close to 90 000 comments from 454 blog posts in blogdelnarco.com, the most popular blog amongst the other four most active blogs. Collecting the data posed a difficult task as the DISQUS[2] commenting system used by the blog does not allow extraction of comments, unless you are the blog owner. To overcome this problem we have used the Blogger API in combination with the iMacro plug-in for the Firefox web browser to scrape all of the comments. These posts range from mid October back to May 2010 and are now publicly available[3].

2 Related Work

A majority of the efforts in sentiment analysis in blogs takes mainly one of two forms, the first one is a knowledge based approach, where a dictionary determines the polarity for words and is then matched to the dataset. The other approach makes use of machine learning techniques where a classifier is used and is fed labeled instances as training data.

The work in Godbole et al. (2007) makes use of a knowledge based approach; lexicons are automatically generated for blogs in various sub domains including crime, then, this marked dictionary is correlated with the corpus in an effort to apply polarity labels to words. On the other hand, Melville et al. (2009) takes a supervised learning approach and constructs a generative model from a polarity annotated lexicon and then builds a model trained on labeled documents.

A similar approach makes use of semi supervised learning. The task of blog classification in Ikeda et al. (2009) tries to identify blogger's gender and age from a limited set of labeled data and a great number of unlabeled instances. It makes use of sub classifiers and trains them for different domains. The output of the classifiers give a weight which can be compared with the output of other classifiers to establish similarity between blogs in order to apply the correct label.

Furthermore, it is worth mentioning some of the problems inherent with sentiment analysis and blogs. Chen and Lin (2010) argues the importance of the class imbalance problem by stating that in blogs there are far fewer instances that will be negative. If there are a greater number of positive instances, then the classifier is highly likely to produce positive labels for the unlabeled instances. Also, when dealing with sentiment analysis in another language, there has been work where machine translation is used to convert text into English, such as the work from Bautin et al. (2008) where they explore the analysis of international news for blogs in nine languages. They mention that although the translation process has some negative effect, this was not a significant issue in their experiments.

3 Approach

Our problem involves determining polarity for entities, yet, we have mentioned that our data set is considerably noisy. As a first step, we will apply the technique to compress words outlined in Pandey and Iyer (2009), then, we will determine the frequency of all the words and remove the stop words that have a high presence on our set. This will reduce our vocabulary for our feature vectors and help eliminate words that do not contribute in determining polarity.

Next, we will filter out any comments that do not contain entities we are tracking. To achieve this, we will rely on named entity extraction from the Illinois Named Entity Tagger Ratinov and Roth (2009), this tools tags plain text with the help of user defined gazetteer. This means, we can provide our own ontology by which the system can recognize entities, thus, avoiding the problem of we might en-

[2] http://www.disqus.com

[3] http://www.cs.utexas.edu/ gcabrera/data.zip

counter with other NER systems where proper casing is required. We presume it might be needed to train the model in this package with our data, in order to achieve results as good as their reported 90.8 F score.

Having a smaller set of messages we will then use human annotators to label a subset of these messages. Starting with an arbitrary number of 500 messages, we plan to iteratively increase the number as we deem necessary for accuracy of our model. Human annotators will be presented with the blog post title, summary and comment to be evaluated, thus, providing them with the needed context in which the opinion was given.

Given our focus of the named entities as sentiments topics, a further division of the comments will be needed, so that we can represent the opinion of one entity per sentence, thus, it can be considered a separate feature vector. We conjecture that this approach will help in accuracy of predicting sentiment in unlabeled instances as it will eliminate any possible ambiguity in the case a second entity is present in the same comment. In the case of two or more entities appear in the same sentence, we will make one copy of the vector per entity and consider a single entity per vector.

Finally, having selected our contextual predicates and making up our feature vectors (as unigrams), we can then train a TSVM. This approach is an extension of the regular SVM but taking into consideration unlabeled data in training the model. In order to work with TSVM we will make use of SVM light [4], an existing implementation of SVM libraries that include support for TSVMs.

4 Evaluation

We plan to compare our approach of a TSVM (semi supervised) model with that of using a SVM (supervised). We make this comparison with the hope to demonstrate that our task will greatly benefit from unlabeled instances as training data. Accuracy on the number of correctly labeled instances will be used as the metric for evaluation. We will show this comparison by plotting one graph containing accuracy with respect to the number of labeled instances for both the TSVM and SVM models.

[4]http://www.cs.cornell.edu/People/tj

The proposed work outlined here is only part of a greater array of problems that could be explored in this domain, for instance, there could be yet another classifier that divides messages into those providing clues that could lead to the prosecution of a criminal; making use of graph theory, one could analyze the behavior of users and what kind of interaction they have with blog posts of a certain topic. We plan to make the dataset publicly available so that further investigation may take place.

References

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *International AAAI Conference on Weblogs and Social Media*, 2008.

Long-Sheng Chen and Li-Wei Lin. Two methods for classifying bloggers's sentiment. In *Proceedings of the International Multi Conference of Engineers and Computer Science*, Hong Kong, 2010.

Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *IEEE 24th International Conference on Data Engineering Workshop*, 2008.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Largescale sentiment analysis for news and blogs. In *ICWSM*, 2007.

Daisuke Ikeda, Hiroya Takamura, and Mnabu Okumura. Semisupervised learning for blog classification. In *Proceedings of Twenty Third AAAI Conference on Artificial Intelligence*, Hong Kong, 2009.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining knowledge with text classification. In *ACM Knowledge Discovery and Data Mining*, 2009.

Vipul Pandey and C.V. Krishnakumar Iyer. Sentiment analysis of microblogs. In *Stanford Machine Learning Course CS229*, 2009.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning. In *Proceedings of EMNLP*, 2002.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 2009.