THE UNIVERSITY OF TEXAS AT AUSTIN

Semi-supervised learning for sentiment analysis on organized crime

Guillermo Cabrera [1] and Jason Baldridge [2]

[1] Department of Computer Science, 1 University Station C0500; [2] Department of Linguistics , 1 University Station B5100. Austin, TX 78712

Introduction

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). The study of the subjectivity of these opinions is referred to as sentiment analysis and has been applied to various domains (ex. Retail, politics, journalism). In this study we focus on sentiment analysis on blogs relating to organized crime, in particular, we study the scenario that Mexico is experiencing with the drug trafficking. We work with data from blogdelnarco.com, a controversial blog which attracts 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, yet there is also a high degree of noise with comments that fail to express sentiment towards an entity (ex. Army, police, government agencies).

We used a semi-supervised approach by means of a Transductive Support Vector Machine (TSVM) model to serve in sentiment classification. A TSVM can be thought of as simply a Support Vector Machine (SVM) trained with some labeled data, then unlabeled data is used to retrain the original model, and iterates this process until the unlabeled data has had an influence on the model.

Problem

- How does one define sentiment analysis in the context of organized crime?
- How to deal with complex use of language: Profanity, colloquial, new vocabulary, bad use of grammar and capitalization. Also need to take into account multiple aliases that refer to same entity.
- Why would this be helpful to government institutions?

Sample comments showing a few of the attributes mentioned above. 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).

(1) A la bio a la bao a la bim bom ba, los federales, ra ra ra / hooray for the federales.
(2) La ****** pfp solo viene a robar, matar y extorsionar, fuera / the ******* pfp only comes to steal, kill and extort, out.

State of the Art

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

- 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.
- Chen and Lin (2010) argue the importance of the class imbalance problem by stating that in blogs there will be far fewer instances that will be negative if there are a greater number of positive instances.
- There has been work such as in Bautin et al. (2008) that makes use of machine translation to convert text into English and then do sentiment analysis in this language. They mention that although the translation process has some negative effect, this was not a significant issue in their experiments.

Methodology

1. **Gathered** data for eight months of activity using Blogger API and iMacro plug-in for Firefox for text extraction.
2. **Cleaned** the text by removing diacritical marks and case folding tokens. Then removed stop words (list of 249) and split comments that had paragraphs into separate comments for cohesiveness.
3. **Filtered** out comments with entities not being tracked. Initial focus was on the President of Mexico (Felipe Calderon) for which we used an ontology of aliases that refer to him.
4. **Labeled** 420 instances with the help of humans as either positive or negative, discarding any neutral. with help of humans.
5. **Created** and trained model using SVM Light Java package.
6. **Classified** unlabeled instances using the model created in previous step. We then validate by using a 10-fold cross validation technique that compares accuracy of an SVM and TSVM model.

Hypothesis: The use of a TSVM for sentiment classification will produce higher accuracy when compared with a SVM.

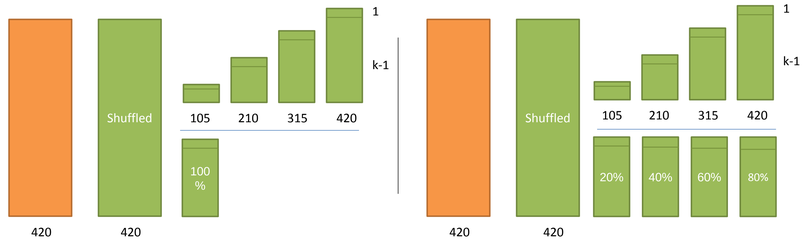

Figure 1: Experiments to test accuracy for SVMs (left) and TSVMs (right).

Results

For the case of an SVM, we can see how accuracy gets better as we include more labeled instances to train our model. On the other hand, our TSVM has a significant improvement in accuracy even though we never used 100% of all training instances to train the initial TSVM. Also, we can use a far less number of labeled instances (compare the 73.3% at 60% for TSVM to the 67.7% at 100% for SVM) and still get better results using a TSVM.

	Percentage Included			
Labeled Instance	20%	40%	60%	80%
105	70.7	70.7	**73.3**	72.0
210	69.5	74.3	75.5	**76.3**
315	67.5	71.3	73.9	**77.2**
420	73.4	71.3	76.6	**78.8**

Table 1: Accuracy comparison of TSVM (Highlighted values plotted in Figure 2).

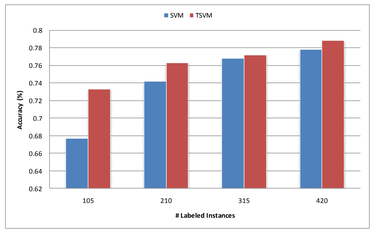

Figure 2: Accuracy comparison: SVM and TSVM

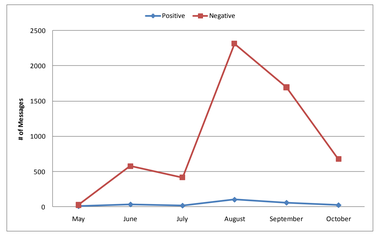

Figure 3: Sentiment over time for President

Future Work

Identified errors in our approach such as: including derogative aliases in our ontology for the President, taking a coarse granularity when breaking on paragraphs instead of sentences.

- Subjectivity classification before polarity classification.
- Plan to use openNLP Sentence Detector.
- Stemming even when bad use of orthography.
- Dimensionality reduction by means of minimal shadows (Pandey and Iyer (2009).

agarrar (grab/catch)

agarado	agarraditos	agarrara	agarraste
agaran	agarrado	agarraran	agarrate
agarando	agarrados	agarrarlo	agarre
agarar	agarralos	agarrarlos	agarremos
agararon	agarrame	agarrarnos	agarren
agaren	agarramos	agarraron	agarrenlos
garrence	agarran	agarrarse	agarrense
agarra	agarrando	agarrarte	agarres
agarraban	agarrandose	agarras	agarro
			agarron

Figure 4: Use of stemming to reduce dimensions

Conclusion

We introduced the problem with blogs in the organized crime domain, we also proposed that a TSVM would be the appropriate tool to do sentiment analysis on comments given the particular settings of the corpus. The results in our work clearly show the advantage of following a semi-supervised learning approach by means of a TSVM. Yet, one may argue that training TSVMs is time consuming, for instance, in order to classify the 5000+ instances a dual core machine took approximately three hours whereas a SVM took less than one minute to classify the same amount of instances.

Despite the time tradeoff, accuracy seems promising, especially after identifying sources of improvement. We have made the dataset we studied publicly available and invite you to play with it:

http://cs.utexas.edu/~gcabrera/data.zip