Summarizing Public Opinions in Tweets

Sentiment Analysis using Twitter corpus

Nibir N. Bora

School of Computer Engineering
KIIT University
Bhubaneswar, Odisha, India

1. Introduction

- *Sentiment Analysis* or *Opinion Mining*, the study of computationally determining whether a piece of text is indicative of positive or negative sentiment, is one way of aggregating the overwhelming amount of user generated content available on the internet today.
- **Applications:**
  - Knowing how users feel about a product or service can help in business decisions for corporates.
  - Political parties and social organizations can collect feedback about their programs and legislation.
  - Artists, musicians and other entertainment icons can reach out to their fans and assess the quality of their work.

  Broadly, it can serve as an automatic polling system, relieving any manual intervention.
- Machine Learning techniques are often applied to sentiment classification, which require a labeled training set of considerable size. Manually labeling large datasets is a tedious and expensive task.
- **We introduce a novel automatic labeling (*noisy label*) technique, *sentiment suggestive words*, using which we created an annotated dataset of 1.5 million tweets. A Naive Bayes Classifier is trained using this training set.**
- **We also study a combination of *minimum word frequency threshold* and *categorical proportional difference* as the feature selection method.**

2. Background

Noisy Label:

A noisy label is an element within the piece of text which provides information about the class to which the text belongs. e.g. the emoticon "`:-)`" in the sentence "`Saw it for the first time, it was a great show, can't wait for next week :-)`" gives a hint that the sentence might contain a positive sentiment.

Table 1: **Previous studies and concerned noisy label methods.**

NAME	*NOISY LABEL*	*DOMAIN*
Pang et al. (2002)	Author ratings	Movie review
Read (2005)	Emoticons	News posts
Mishne and de Rijke (2006)	Author tagged moods	Blog posts
Go et al. (2009)	Emoticons	Twitter
Yessenov and Misailovi (2009)	Manually labeled	Movie review
Duurkoop (2010)	Hashtags	Twitter
Pak & Paroubek (2010)	Emoticons	Twitter

Feature Selection:

Feature selection, used in order to improve the performance of the classifier, is a method to select a portion of the feature set, generated by the trainer, which is most likely to serve in classification.

Minimum Word Frequency Threshold: All features with frequency below a minimum threshold frequency (calculated as a percentage on the maximum frequency of any feature) are removed from the feature set.

Categorical Proportional Difference: It is a measure of how equal two numbers are. If a feature is prevalent in either positive tweets or in negative tweets then it's CPD will be close to one whereas if it occurs almost evenly in both positive and negative tweets then its CPD will be close to zero.

Experimental Setup 3.

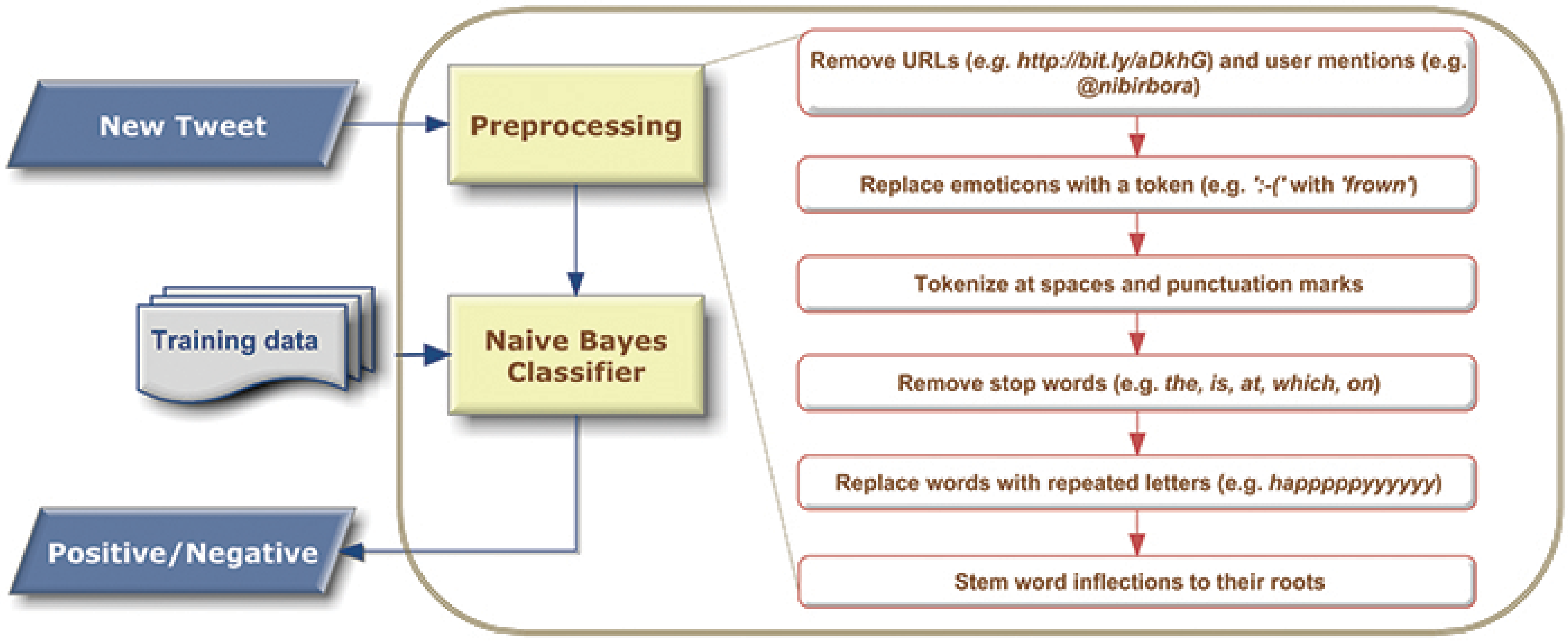
New Tweet
Preprocessing
Training data
Naive Bayes Classifier
Positive/Negative
Remove URLs (e.g. http://bit.ly/aDkhG) and user mentions (e.g. @nibirbora)
Replace emoticons with a token (e.g. ':-(' with 'frown')
Tokenize at spaces and punctuation marks
Remove stop words (e.g. the, is, at, which, on)
Replace words with repeated letters (e.g. happppppyyyyyy)
Stem word inflections to their roots

Figure 1: **Overview of the experimental setup**

The training set:

- A set of 40 words was prepared, each indicating certain sentiment value, and were manually categorized as being positive or negative.
- A tweet is labeled as positive if it contained any of the positive sentiment words, or negative if it contained any of the negative sentiment word.

Table 2: **List of positive and negative sentiment words**

Positive sentiment words	*Negative sentiment words*
amazed, amused, attracted, cheerful, delighted, elated, excited, festive, funny, hilarious, joyful, lively, loving, overjoyed, passion, pleasant, pleased, pleasure, thrilled, wonderful	annoyed, ashamed, awful, defeated, depressed, disappointed, discouraged, displeased, embarrassed, furious, gloomy, greedy, guilty, hurt, lonely, mad, miserable, shocked, unhappy, upset

Table 3: **Example of a few labeled tweets**

Tweet	*Label*	*Matched sentiment word*
Arrived in Basel. Brilliant sunny weather. I'll go to the botanical gardens first. It's a wonderful place to think and write.	positive	wonderful
I should really eat something, but i'm just not a fan when it's this hot and miserable. :/	negative	miserable

The test set:

A hand labeled test set of 198 positive sentiment tweets and 204 negative sentiment tweets was used to check the accuracy of the classifier.

Results 4.

- The maximum accuracy achieved was **83.33%**.
- Maximum accuracy with CPD alone as the feature selection method was **83.08%**.
- Using Categorical Proportional Difference (CPD) as a feature selection method increased the accuracy of the classifier, showing a peak at CPD value 0.25.
- Minimum word frequency (MWF) threshold did not contribute as a feature selection method. The accuracy decreased gradually with an increase in the MWF threshold.

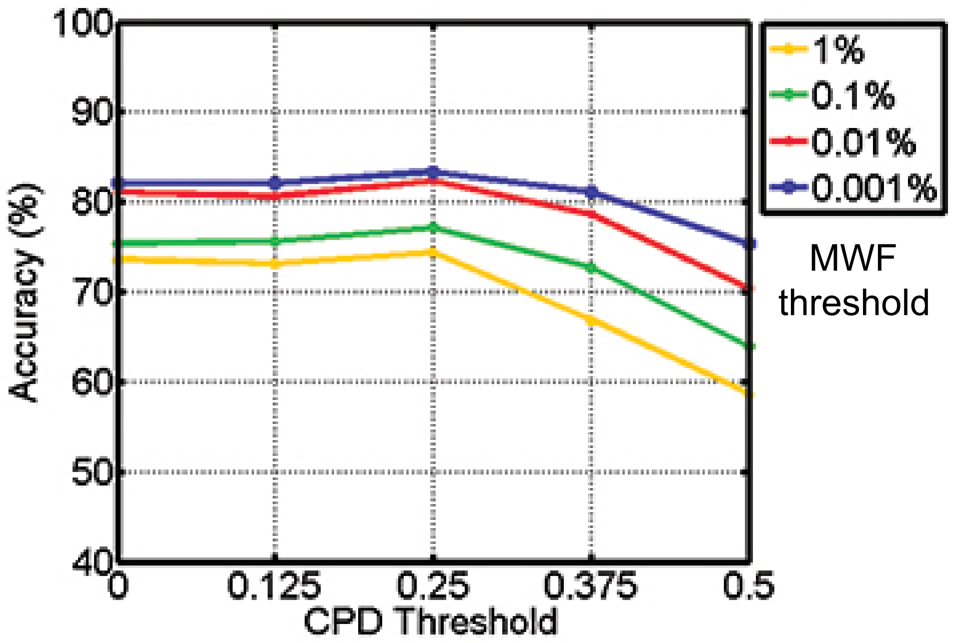

Figure 2: **Accuracy of the classifer at different MWF thresholds and CPD thresholds.**

Conclusion 5.

- Sentiment suggestive words can be effectively used as noisy label for a Twitter corpus. A training set labeled using sentiment suggestive words is as good as an emoticons annotated dataset.
- Categorical Proportional Difference (CPD) performs well as a feature selection method with Twitter corpus (CPD value 0.25 being a descent choice).
- Setting a minimum word frequency threshold does not perform well as a feature selection method with Twitter corpus.

References

1. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: Proceedings of the Asia Pacific Finance Association Annual Conference (APFA). (2001)
2. Duurkoop, J.: Real-Time Happiness. Bachelor thesis, University of Amsterdam, Faculty of Science, Science Park 904, 1098 XH Amsterdam (2010)
3. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford University (2009)
4. O'Keefe, T., Koprinska, I.: Feature Selection and Weighting Methods in Sentiment Analysis. In: Proceedings of the 14th Australasian Document Computing Symposium. (2009)
5. Lake, T.: Twitter Sentiment Analysis. Technical Report, Western Michigan University, Kalamazoo (2011)
6. Mishne, G., de Rijke, M.: Capturing Global Mood Levels using Blog Posts. In: AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pp. 145-152. The AAAI Press, Menlo Park (2006)
7. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA). pp. 19-21. Valletta, Malta (2010)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86 (2002)
9. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 271-278. (2004)
10. Porter, M.F.: An algorithm for suffix stripping. In: Program, 14(3) pp. 130-137 (1980)
11. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. (2005)
12. Simeon, M., Hilderman, R.: Categorical Proportional Difference: A Feature Selection Method for Text Categorization. In: Proceedings of the 17th Australasian Data Mining Conference, pp. 201-208. (2008)
13. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics, pp. 417-424. (2002)
14. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML97), pp. 412-420. Nashville (1997)
15. Yessenov, K., Misailovic, S.: Sentiment Analysis of Movie Review Comments. 6.863 Spring 2009 fnal project, CSAIL, MIT (2009)

http://mejaj.nibir.me nibirbora@gmail.com