Fam Jiang Yuan 1121116369

Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Alexander Pak, Patrick Paroubak

Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Batiment 508,
F-91405 Orsay Cedex, France
alexpak@limsi.fr, pap@limsi.fr

Abstract

Microblogging nowadays became a most common communication tools for all Internet users. There are millions of users share their thoughts or opinions on different aspects of their daily life. Meanwhile, millions of corpus or even more, appear every day in popular social websites. Thus, microblogging websites are the idea platforms for data mining and sentiment analysis. In this research, we will concentrate on one of the most popular microblogging website, Twitter, in order to perform sentiment analysis. After that, we show how to collect a corpus from Twitter automatically and use it to train a sentiment classifier, which able to determine positive, negative, or neutral sentiment of a corpus. Throughout our experimental evaluations, we found that our methods are efficient and performs better than the previous existing methods. Besides that, we only worked on English, however, our proposed method can adapt on any language.

Problem Solved

In order to collect those corpus, we used a Twitter API to collect those valued information and then perform sentiment analysis by building a sentiment classifier, which able to determine positive, negative, or neutral sentiment for the collected corpus automatically.

Claimed Contribution

First and foremost, we collect a corpus with positive and negative sentiment, and a corpus of objective texts by using our method that doesn't required human effort for classifying the collected corpus. Next, we performed a statistical linguistic analysis of the collected corpus and used the corpus to train a sentiment classifier for microblogging. Finally, we carried out an empirical evaluations by implementing the sentiment classifier on the real microblogging posts to prove that our methods are efficient and performs better than previous existing methods.

Related Work

This research was referred to an existing work that presented by (Pang & Lee, 2008), who describe the existing methods and approaches for an opinion-oriented information retrieval. In (Yang, Lin, & Chen, 2007), who used an emotion classifier to determine the sentiment level of the collected corpora by using Support Vector Machine (SVM) and Conditional Random Field (CRF) learning machine. J.Read (Read, 2005), who used emoticons to train the classifiers, which are SVM and Nave Bayes, both classifier were able to obtain 70% of accuracy during the test set. In (Go, Huang, & Bhayani, 2009), authors used the similar approach, which using emoticons to get corpus sentiment on Twitter to collect the data and performed a sentiment analysis. So, the best obtained result goes to Nave Bayes classifier as the classifier obtained 81% of accuracy on the test set.

Methodology

For corpus collection, we used a Twitter API and same approach as in ((Read, 2005); (Go et al., 2009)) to collect a corpus of text posts and classify into positive sentiment, negative sentiment, and objective texts (no sentiment, neutral, or facts) with two types of emoticons, which are Happy emoticons and Sad emoticons. After that, those collected corpora were used to train a classifier to determine positive and negative sentiment of documents. We used English language in our research. For corpus analysis, we used a tool called TreeTagger (Schmid, 1994) to tag all the posts in the collected corpora and then distribute the Part of Speech (POS)-tags. Moreover, we have compared the distributions of POS-tags and the results show that the positive texts are characterized by the positive ending, and the negative texts are characterized by the words of loss and disappointment. For training the classifier, we used the existing of an n-gram as binary feature to extract the information. We have tested with unigrams, bigrams, and trigrams. Three of the n-gram perform differently at different aspects. Furthermore, we build a sentiment classifier by using the multinomial Nave Bayes, since Nave Bayes obtained the best results among SVM and CRF. Next, we have hand-annotated the outputs by testing our classifier on real Twitter posts. The results are presented in Table 1.

Sentiment	Number of samples
Positive	108
Negative	75
Neutral	33
Total	216

Table 1: The Patterns of the evaluation dataset.

We evaluate the accuracy (Manning & Schütze, 1999) of our classifier using:

$$accuracy\,of\,classifier = \frac{N(correct\,classifications)}{N(all\,classifications)} \tag{1}$$

We determine the accuracy across the classifiers decision (Adda, Mariani, Lecomte, Paroubek, & Rajman, 1998) using:

$$classifier\,decision = \frac{N(retrieved\,documents)}{N(all\,documents)} \tag{2}$$

The testing results are recorded and presented in graph.

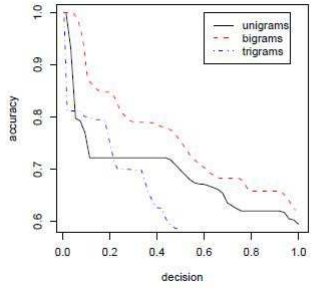

Figure 1: The comparision of the classification accuracy when using three of the n-grams.

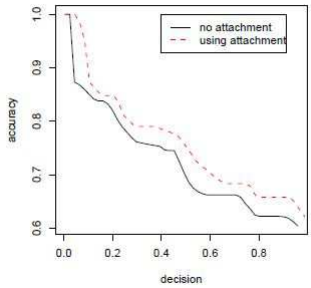

Figure 2: The impact when negation words were attached.

Throughout our testing results, we found that the best performance is obtained when using bigrams, as bigrams provide a balanced coverage (unigrams) and good capability to capture the expression patterns of sentiment corpus (trigrams).

Conclusion

Microblogging nowadays became a most common communication tools for all Internet users. Microblogging websites is a platform where millions of corpus or even more, appear every day in popular social websites. Thus, microblogging websites are the idea platforms for data mining and sentiment analysis. In this carried out research, we showed a method that can collect a corpus automatically and to be used to train a sentiment classifier. Next, we used Tree Tagger for POS-tag and observed the distribution among positive, negative, and neutral set. We trained a sentiment classifier known as Nave Bayes that uses n-grams and POS-tags as features. Our trained classifier is capable to classify positive, negative, and neutral of document.

Future Extension

Throughout this research, I have learned some basic knowledge about data mining, and training a learning machine, in order to make the machine to make decision.

However, the presented methodology in this research were limited in one language. So, as the future extension, the authors may plan to collect a multi-language corpus from microblogging websites and compare the collected corpora across with difference language to build a multi-language sentiment classifier.

References

Adda, G., Mariani, J., Lecomte, J., Paroubek, P., & Rajman, M. (1998). The grace french part-ofspeech tagging evaluation task. In *in proceedings of the first international conference on language resources and evaluation (lrec.*

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Entropy*, *17*.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the acl student research workshop* (pp. 43–48).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (Vol. 12, pp. 44–49).

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Web intelligence, ieee/wic/acm international conference on* (pp. 275–278).

Paper Title	Twitter as a Corpus for Sentiment Analysis and Opinion Mining
Author(s)	Alexander Pak, Patrick Paroubek
Abstract/Summary	
Problem Solved	
Claimed Contributions	
Related work	
Methodology	
Conclusions	
Future Extension	
References	