Sentiment Analysis in Twitter

Rohit Kumar Jha, Sakaar Khurana

Outline

Introduction

Problem Statement

Given a message, classify whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

Motivation

- In the past decade, new forms of communication, such as microblogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them.

- Tweets and texts are short: a sentence or a headline rather than a document. The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "re-tweet" and # hashtags, which are a type of tagging for Twitter messages.

- Another aspect of social media data such as Twitter messages is that it includes rich structured information about the individuals involved in the communication. For example, Twitter maintains information of who follows whom and re-tweets and tags inside of tweets provide discourse information.

Previous Works

Among the various machine learning algorithms that have been used for sentiment analysis Naive Bayes, SVM and MaxEnt have shown promising results in movie-review classification and subsequently in recent Twitter sentiment analysis research.

Bag of Words Model

- Use a word list where each word has been scored positivity/negativity or sentiment strength
- Overall polarity detemined by the aggregate of polarity of all the words in the text
- Achieves accuracy of 68.58% and becomes 72.81% when using discourse relations as well

Feature Extraction

In the world of microblogs, with prime focus set on Twitter, work done by Pak et al. confirm that a bigram model outperforms both unigram and trigram models while using a Multinomial Naive Bayes classifier. However, the reverse was true in the case of SVM and MaxEnt classifier studies conducted by Go et al. . Introduction of a combination of unigram and bigram in feature extraction promised better results in MaxEnt as well as NB classifiers.

Unigrams

- The easiest and most used approach
- Pang et al. reported an accuracy of 81.0%, 80.4%, and 82.9% for Naive Bayes, MaxEnt and SVM respectively in the movie-review domain
- Found to be closely similar to accuracies obtained in twitter classification which were 81.3%, 80.5%, and 82.2% respectively

Unigram+Bigram

- Both unigrams and bigrams are used as features
- In the movie-review domain, a decline observed for Naive Bayes and SVM, but an improvement for MaxEnt
- Recent research in the twitter research bed found that as compared to unigram features, accuracy improved for Naive Bayes (81.3% from to 82.7%), MaxEnt (from 80.5 to 82.7%) and there was a decline for SVM (from 82.2% to 81.6%)

POS Tagging

- Past experiments with POS tagging in feature extraction for sentiment analysis have yield little improvements
- The accuracy improves slightly for Naive Bayes but declines for SVMs, and the performance of MaxEnt is unchanged while classifying tweets with their individual accuracies being 81.5%,81.9% 80.4% respectively

Naive Bayesian Classifier

- Straightforward and frequently used method for supervised learning
- Provides a exible way for dealing with any number of attributes or classes, and is based on probability theory
- **Maximum entropy classifiers** are commonly used as alternatives to Naive Bayesian classifier because they do not require statistical independence of the features that serve as predictors
- Provides around 79% accuracy for tweets

Our Work

Features Considered

We plan to make use of following additional features apart from the ones mentioned till now.

Sentence Weightage

- If a tweet consists of more than one sentences, we give more weightage to sentences coming afterwards
- This is due to the tendency of most tweets to be conclusive in nature
- When testing it on small set of tweets, it improved accuracy by around 2.5%

Hashtags

- We plan to use the hash tags to get idea about the tweets
- The hashtags are like this: #IndiabeatAus #FinallySuccessful and so on
- These hashtags would be structured though not complete sentences
- So, we would need to parse these tweets before processing
- Hashtags like #happy, #good, #unhappy, etc give sufficient information about the polarity of the tweets

Abbreviations and Redundant/Repeated letters

- Due to the casual nature of Twitter language, several words (in many cases opinion words) are misspelt or often over emphasized due to which the classifer may not attribute polarity of this word (eg. looooooooove) to the actual word (eg.love) during training
- In words containing more than 3 occurences of the same letter together, these occurences are replaced with 2 instances of the letter. eg. haaaaaaaappy would be replaced by haappy , goooooooood would be replaced by good
- Created a list of common and most polular abbreviations of most commonly used words

Smileys

- Smileys are also a great source of information about the tweets
- Smileys have more wightage than the overall text of the tweets, and we give more weightage to smileys in sentences coming afterwards
- Created a list of all used smileys across different social networks

Other Ideas

- Try to incorporate the effect of modifiers like "very", "too", etc
- Consider this tweet: "Such a great knock. Team scored this at the loss of just one wicket." Now the problem is that it contains one word "great" and the other "loss", and so we would get the overall sentiment as neutral. but it is indeed positive. It is important to capture the idea, as to why it is so. The reason is that they say 'loss of "only" one', meaning at a minimal loss. So, if we capture this notion as well, we will get a pretty increase in accuracy. This is something that keeps appearing in texts, inclusing tweets. So, we plan to consider prepositions like "of", "in", "by", etc in viscinity of these sentiment/opinion words.

Datasets

- This free data set is for training and testing sentiment analysis algorithms. It consists of 5513 hand-classified tweets. Each tweet was classified with respect to one of four different topics. This has been obtained from the website of Sanders Analytics, a Seattle-based startup focused on data analytics. http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip

- Sentiment140 Lexicon: The sentiment140 corpus (Go et al., 2009) is a collection of 1.6 million tweets that contain positive and negative emoticons. The tweets are labelled positive or negative according to the emoticon. http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

- SEMEVAL 2013 has also provided with around 30000 labelled tweets for the "Contextual Polarity Disambiguation" problem and another 10000 for the "Message Polarity Classification" problem.
  http://www.cs.york.ac.uk/semeval-2013/task2/index.php?id=data

References

Alec, G.; Lei, H.; and Richa, B. Twitter sentiment classification using distant supervision. Technical report, Standford University. 2009

AR, Balamurali and Joshi, Aditya and Bhattacharyya, Pushpak. Harnessing WordNet Senses for Supervised Sentiment Classification. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP). 2011

Asher, Nicholas and Benamara, Farah and Mathieu, Yvette Yannick. Distilling opinion in discourse: A preliminary study. In Proceedings of Computational Linguistics (CoLing). 2008

Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In 23rd International Conference on Computational Linguistics: Posters, 36–44. 2010

Bermingham, A., and Smeaton, A. Classifying sentiment in microblogs: is brevity an advantage ACM 1833–1836. 2010

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. 2011

Dey, Lipika and Haque, Sk. Opinion Mining from Noisy Text Data. International Journal on Document Analysis and Recognition 12(3). pp 205-226. 2009

Elwell, Robert and Baldridge, Jason. Discourse Connective Argument Identification with Connective Specific Rankers. In Proceedings of IEEE International Conference on Semantic Computing. 2008

Questions?