MIT SCHOOL OF ENGINEERING

Department of Computer Science and Engineering

Mini Project Synopsis

Project Title: SENTIMENT ANALYSIS OF CUSTOMER'S FEEDBACK

Problem Statement:

The problem at hand consists of two subtasks:

* Phrase Level Sentiment Analysis:

Given a message containing a marked instance of a word or a phrase, determine whether that instance is positive, negative or neutral in that context.

* Sentence Level Sentiment Analysis:

Given a message, decide 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.

Abstract:

Sentiment analysis evaluates people's declarations from written language aiming to automatically determine opinions, sentiments, evaluations, appraisals and emotions, identifying whether the text expresses a positive, neutral or negative perspective.

This thesis performs the sentiment analysis in opinion articles in the specific domain of economics, exploring linguistics and computer science techniques and analysing the effects of adding domain-specific terms to a given lexicon.

Our objectives of this project are:

* To implement an algorithm for automatic classification of text into positive and negative.
* Sentiment analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest.

Our approach for sentiment analysis is based on the natural language processing and text detection and retrieval with the help of python.

The goal is to verify whether the combination of a general purpose lexicon and a domain-specific lexicon achieves better performance than just a general-purpose lexicon on its own.

Literature Review:

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Tumey, 2002; Pang and Lee, 2004) [4,5], it has been handled at the sentence level (Hu and Liu [2], 2004; Kim and Hwy, 2004 and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009). Microblog data like Twitter, on which users post real time reactions to and opinions about "everything", poses newer and different challenges. Some of the early and recent results on sentiment analysis of Twitter data are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and ParoubeK (2010) [3]. Go et al. (2009) use distant learning to acquire sentiment data. They use tweet sending in positive emotions like :-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, Max Ent and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature space, they try a Unigram, Bigram model in conjunction with partsof-speech (POS) features. They note that the unigram model outperforms all other models. Specifically, bigrams and POS features do not help. Pak and ParovbeK (2010) [3] collect data following a similar distant leaming paradigm. They perform a different classification task though: subjective versus objective.

For subjective data they coiiect the tweets ending with emoticons in the same manner as Go et al. (2009). For objective data they crawl twitter accounts of popular newspapers like "New York Times", "Washington Posts" etc. They report that POS and bigrams both help (contrary to results presented by Go et al. (2009)). Both these approaches, however, are primarily based on ngrarn models. Moreover, the data they use for training and testing is collected by search queries and is therefore biased. In contrast, we present features that achieve a significant gain over a unigram baseline. In addition we explore a different method of data representation and report significant improvement over the unigram models. Another contribution of this paper is that we report results on manually annotated data that does not suffer from any known biases. Our data will be a random sample of streaming tweets unlike data collected by using specific queries. The size of our hand-labelled data 6 will allow us to perform cross validation experiments and check forth variance in performance of the classifier across folds. Another significant effort for sentiment classification on twitter data is by Barbosa and Feng (2010).

They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labelled tweets for tuning and another 1000 manually labelled tweets for testing. They however do not mention

how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. We extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classifiers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally. Gamon (2004) perform sentiment analysis on feedback data from Global Support Services survey. One aim of their paper is to analyse the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contributes to the classifier accuracy. In this paper we perform extensive feature analysis and show that the use of only 100 abstract linguistic features performs as well as a hard unigram baseline.

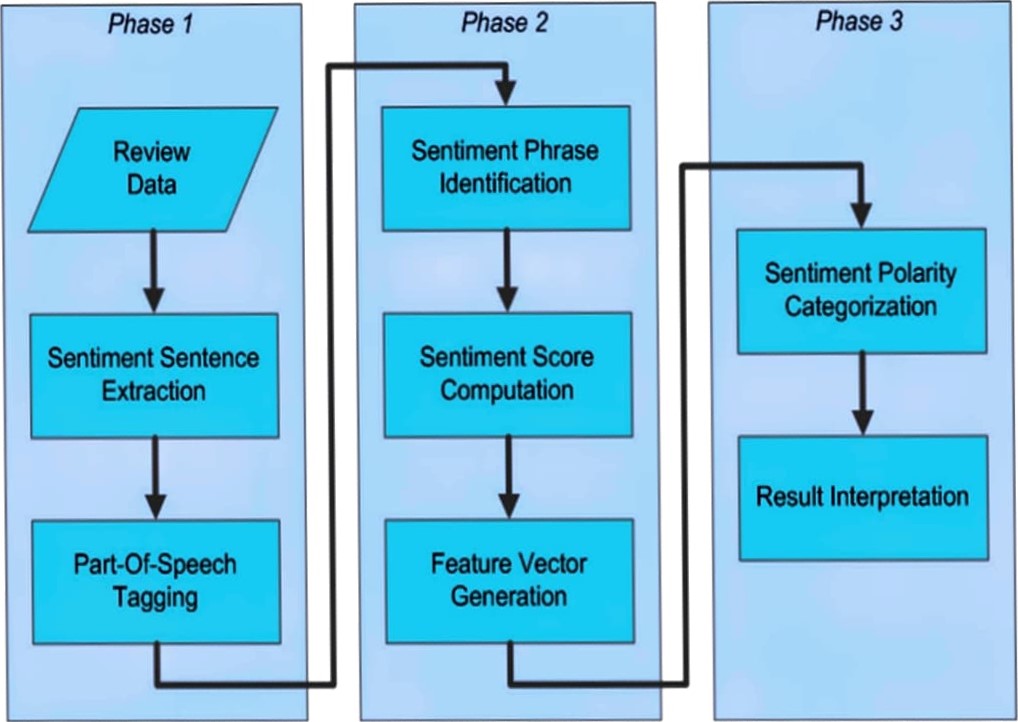
One fundamental problem in sentiment analysis is categorization of sentiment polarity [5,11]. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative(or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level [12]. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization. The entity and aspect level then targets on what exactly people like or dislike from their opinions.

For feature selection, Pang and Lee [4] suggested to remove objective sentences by extracting subjective ones. They proposed a textcategorization technique that is able to identify subjective content using minimum cut. Gann et al. [13] selected

76,799 tokens based on Twitter data, where each token is assigned a sentiment score, namely TSI (Total Sentiment Index), featuring itself as a positive token or a negative token.

Moreover, [8] showed that using the well-known "geo-tagged" feature in twitter to identify the polarity of a political candidates in the US could be done by employing the sentiment analysis algorithms to predict the future events such as the presidential elections results. Comparing to previous approaches in sentiment topics, additional findings by [10] showed that adding the semantic feature produces better Recall (retrieved documents) to compute the score) in negative sentiment classification.

Proposed System [Block Diagram]:



Conclusion:

The project: Sentiment Analysis of customer feedback will be completed using python as a language.

We were able to determine the positivity and negativity of each tweet. Based on the tweet we will we representing them in the form of graph. A small conclusion will also be shown during output representation based on product or brand entered. Our designed system is user friendly.

References:

1. Kim S-M, How E (2004) Determining the sentiment of opinions In: Proceedings of the 20th international conference on Computational Linguistics, page 1367. Associations for Computational Linguistics, Stroudsburg, PA, USA.
2. Liu B (2010) Sentiment analysis and subjectivity In: Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca. Liu B, Hu M, Cheng J (2005) Opinion observer: Analysing and comparing opinions on the web In: Proceedings of the 14th International Conference on World Wide Web, WWW '05, 342-351. ACM, New York, NY, USA.
3. Pak A, ParpubeK P (2010) Twitter as a corpus for sentiment analysis and opinion mining In: Proceedings of the Seventh conference on International Language Resources and Evaluation. European Languages Resources Association, Valletta, Malta.
4. Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts In: Proceedings of the 42 Nd Annual Meeting on Association for Computational Linguistics, ACL '04.Association for Computational Linguistics, Stroudsburg, PA, USA.
5. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr2(1-2): 1-135.
6. Twitter (2014) Twitter apis. https://dev.twitter.com/start.
7. Liu@ (2014) The science of detecting fake reviews. http://content26.com/blog/bing-liu-the-science-of-detecting-fakereviews/
8. Jahan>akhsh, K., & Moon, Y. (2014). The predictive power of social media: On the predictability of U.S presidential elections using Twitter.
9. Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews In: Proceedings of the 21st, Intemational Conference on World Wide Web, WWW'12, 191—200.. ACM, New York, NY, USA.
10. Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. The Semantic Web (pp. 508—524). ISWC

1 1 . Tan LK-W, Na J-C, Theng Y-L, Chang K (2011) Sentencelevel sentiment polarity classification using a linguistic approach In: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation, 77—87. Springer, Heidelberg, Germany.

1. Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
2. Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system In: Proceedings of the second ASE international conference on Big Data. ASE.
3. Joachims T. Probabilistic analysis of the rogghjo algorithm with TFIDF for text categorization. In: Presented at the ICML conference; 1997.
4. Yung-Ming Li, Tsung-Yung Li Deriving market intelligence from microblogs.