Aaron Cofield
Matthew Abraham

CIS 411 Final Report

1. Abstract

Our project will cover the analysis of emotion in tweets directed towards, in reply to, or mentioning political candidates. The goal here is to determine what the prevailing sentiment on social media is towards this candidate. To this end we are using an SVM classifier to judge the emotional sentiment of a body of tweets.

2. Motivations

For our project, we wanted to look at the problem of sentiment analysis, especially in determining sentiment towards public and prominent figures, like presidential candidates or celebrities. Specifically, we wanted to examine what methods or classifiers were best at determining sentiment in a given document. This could be useful in determining the public's attitude on social media, the attitude of news media (by way of examining articles or T.V. transcripts), and so on.

3. Related Work

The task of sentiment analysis has been a subtask at the past several SemEval conferences. Recent approaches include the use of SVM's, and neural network models such as convolutional neural nets or recurrent neural networks.

4. Dataset

For our data set, we chose the English E-c sets from the 2018 SemEval Task 1. This set is a small set of English tweets, roughly 13,000 in all. They are labelled with 11 possible emotions, including anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. This benefit of choosing this dataset was that we could have a more fine-grained classification of text, beyond simply a positive, neutral, or negative label. The dataset was also easily downloaded, being directly available from the SemEval site.

5. Methodology

To classify sentiment in tweets we collect several features including if the tweet has a question mark, the number of words, the first emoji in the tweet (or none if there are none), the number of positive words, and the number of negative words. The number of positive words is determined by the set intersection of the words in the tweet and a large dictionary of positive words here http://ptrckprry.com/course/ssd/data/positive-words.txt. The same is done for negative words. The classifier utilized is the scikit learn support vector machine.

6. Experimentation

Using just these features we are able to achieve about 55% accuracy.

7. Contribution

Matt - half report, data parser, collected some features

Aaron - half report, collected more features, fit and evaluate svm classifier

8. Conclusion

The few features we have only allow for 55% accuracy, adding a features such as including text from most popular positive/negative bigrams/unigrams would increase the accuracy.

9. References

Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

Semeval-2018 Task 1: Affect in Tweets. Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018.