Course: COMP 4750
Group Members: Eric Elli & Tyler Snow
Date: Novemeber 1st 2018

We have chosen Sentiment analysis as we believe this to be one of the most interesting fields of NLP since it is applicable to so much of human life and our written interactions. It is a great way of extracting human emotions, opinions, and attitudes from text. With there being so much available text to analyze. It is a frontier of computer science and everyday is teaching us more about ourselves individually and collectively. Sentiment analysis can also be applied to many real world applications in any profession, Since the expressed opinions of people have an impact on society. Given this the more opinions you can accurately process can help you identify trends significantly faster then without these systems. The development of our software will require machine learning which adds to our motivation as it is currently a prominent technique in the landscape of Computer Science and very applicable to a multitude of fields.

There has been a lot of previous work on sentiment analysis and its applications. Currently the two major classifications of sentiment analysis are broken down into the machine learning approach and the lexicon-based approach. The Machine learning approach uses several learning algorithms (both supervised and unsupervised) to classify data. The Lexicon based approach uses a dictionary containing positive and negative words to determine the sentiment polarity.[1] Each of these classifications have been further broken down into more specific techniques. There are many machine learning and lexicon-based sentiment analysis algorithms that exist. [2] They has been a lot of research on the individual approaches as well as hybrid approaches. As is to be expected there is not one clear choice to use and the optimal approach depends on the constraints and parameters of the specific problem (The domain). The lexicon-based approach faces the disadvantage that the strength of the sentiment classification depends on the size of the lexicon. As this lexicon grows it becomes more error prone and is very time consuming. Machine learning classification requires that the data set be labelled and be very large in order to train and test. Hybrid approaches tends to improve accuracy overall as it can take from the best of each respective method. [1]

Our first step will be to implement some of the machine learning algorithms to compare and contrast. Support vector machines, Naive Bayes, and decision trees will be some of the algorithms we plan to experiment with. Once we have our algorithms chosen we will then being to train and test it on reviews. Then we will compare the results of the algorithms. We will test our machines using test data and compare the run-times and success-rates of all the algorithms. Then we will generate a lexicon to integrate into our systems and retrain and test them again. We will perform the same run-time and success-rate comparisons and check those versus the pure machine learning approach.

References:

[1] Anuja P Jain, Padma Dandannavar, “Application of Machine Learning Techniques to Sentiment Analysis”, 2nd International Conference on Applied and Theoretical Computing and Communication Technology(iCATccT), 2016

[2] Walaa Medhat, Ahmed Hassan, Hoda Korashy, “Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal ,2014

[3] Bing Liu, “Sentiment Analysis and Opinion Mining”. Morgan & Claypool Publishers, 2012, ISBN: 9781608458844

[4] Duyu Tang, Meishan Zhang, “Deep Learning in Sentiment Analysis”. Springer Nature Singapore Pte Ltd, 2018, ISBN: 9789811052095

[5] John Blitzer, Mark Dredze, Fernando Pereira, “Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification”, 45th Annual Meeting of the Association of Computational Linguistics, 2007