OPINION MINING AND SENTIMENT ANALYSIS

BACK GROUND RESEARCH AND SIGNIFICANCE OF RESEARCH:

The main purpose of writing this paper is to present my approach about sentiment analysis and opinion mining. Human behavior get influenced a lot from the opinions that they develop from surrounding, or in today's time we should replace surrounding with the web. There used to be a time when we try to ask for suggestion from our friends and few experience people before buying anything. Now we heavily rely on the opinions or responses that we get about that article on web from N number of different users, which is no doubt a better method for analyzing some products. Opinion and its related concepts such as sentiments or emotions, or attitudes are the subject areas of doing sentiment analysis.

The rapid pace of social media, as per the latest data (318 million twitter users & 1.7 billion Facebook users) shapes up the opinion of population to a great extent. Opinions mining has been playing a key role from politics campaigning to driving marketing strategies of many products based companies. Sentiment analysis is a subset of many forms of natural language processing. Research in the areas of Natural Language processing has been going for around 50 years, as we always try to understand can machines process our Spoken languages as well as the humans do. During our study of sentiments, we also try to collect documents and try to segregate those documents as good or bad or positive or negative about certain products or events. It has gained a lot of significance for the last 10 -12 years, Companies are doing this tremendous job of understanding the problems of their customers about certain product, Election results are suggested by doing sentiment analysis state by state and there are many more ways in which this information is used.

RESEARCH OBJECTIVE AND AIM:

The main objective of mine is to improve the state of the art work in the field of opinion mining and try and come up with a different methodology to categories the sentiments of a document or a sentence. The state of the art work in this filed heavily relies on the supervised approach of machine

learning where humans have to analyze and label huge number of training sets and then using those training sets to provide labelling to another set of documents or reviews and now a day's tweets as well.

For Supervised Machine learning approach Few algorithms that are mostly used are Naïve Bayes classifiers, Support vector machines and lexicon based approaches for their abilities to classify the sentences on behalf of unigrams (one word). I have myself recently worked on building a sentiment analyzer using Naïve Bayes machine learning and SVM based approach on recent political campaign of Trump and Hilary. For Naïve Bayes or for that matter any supervised approach we have to feed the machine with huge number of labeled text documents to get some better results for some documents for which we would like to collect the opinions of a public.

The motive of this piece of work is not only to find the sentiments better, but to make it domain independent, and to make machines learn the semantic part of the language well using Unsupervised approach. The day we are capable to teach machines to understand the context of a sentence as humans do we could really do wonders in many fields.

RESEARCH PLAN AND OBJECTIVES:

Unsupervised Approach:

I myself have already built a sentiment analyzer using naïve Bayes model and SVM machine learning techniques after training our data sets to huge number of labelled documents. Main reason for working on Unsupervised approach is to remove the domain dependency which is a major problem in sentiment analysis. Using this approach, we need not annotate the training data which saves a lot of effort.

Inspired by Turning [4], I will try to use a lexicon which will break up the document word by word and assign a polarity to each and every word as per the defined in the lexicon and then assign the overall sentiment to a document by aggregating the sum of all polarities. A part of using unsupervised approach I have already completed and could be here which will be suggesting the percentage of tweets are positive and negative, along with the

emotions present in those (Lexicon Used NRC).

But using lexicon word by word, we don't care about the order in which word is being used, and also misses on the effect of negation used in the sentences, so we are neglecting the importance of order and thus we are missing the contextual meaning of a sentence at the same time we are also neglecting the effect of negation words which could completely change the meaning of the sentence. For instance, if we look at the sentence ***"I am having a lot of fun" versus "I am not having a lot of fun".*** First one is clearly a positive sentiment, however second one is not if we take only unigrams, but the impact of negation can be caught if use {bi –grams} or {N-grams}.

Deep Learning Approach:

NLP is hard for the reason as discussed earlier, the way people express themselves, Usage of a language, Use of emoticons and symbols a lot while expressing themselves, sarcasms and so many other issues. But its seems Deep learning could solve these problems to a great extent and I would also like to use some of the deep learning techniques to do the sentiment analysis. I would specially like to use Convolution Networks here with word2vec and then using word embedding and convolving the trigrams, multigrams over different layers and then finally doing the Max Pooling to find out the final vector or say the words which actually decide the sentiment of a sentence. This approach of doing the sentiment analyses is very fresh, however other approached like recurrent neural networks and recursive neural networks. I would also like to compare my results with the Cloud Built machine learning engines like the one by Google NLP API, IBM Watson and Amazon Machine learning capabilities.

DATA GATHERING AND AVAILIBILITY:

When it comes to gather data for Sentiment analysis, there is no dearth of the resources. Anything that goes on the web goes in the text format, so we web scrap any website like twitter, Facebook or amazon etc. Anything that is available on the web can be used for this project purposes. I have successfully build a data base from Twitter of around 20k using Twitter API and it can do a

wonderful job in fetching tweets on behalf of #some event, @followsombody, or any particular piece we are intrested in.

REFERENCES:

[1] James F. Allen – "Natural language Understanding ". Page [100-148]

[2] Bing Liu -"Sentiment Analysis and Opinion Mining". Page [30-56]

[3] Retrieved From https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html

[4] NLP Stanford Learning Group http://nlp.stanford.edu/sentiment

[5] Retrieved From http://nlpers.blogspot.com/search/label/machine%20learning

[6] P.D. Turney – "Semantic orientation applied to unsupervised classification of reviews. "Linguistics, pages [417–424, 2002.]