Capstone project proposal on

Analysis of Tweets – Opinion Mining

By- Harshit Bajpai

Domain Background –

Sentiment analysis is the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral(Oxford dictionary). It is extremely useful for various businesses to monitor the attitude of consumers and then market their products accordingly. It helps businesses to accurately plan their marketing strategy. Not just businesses but political campaign runners also make use of this strategy to analyse the preferences of voters and their attitude towards the policies they are campaigning. There have been many previous.

Problem Statement –

For the purposes of this project we will be implementing a program that categorizes a user's tweets a positive or negative or neutral.

Datasets and Inputs –

In this project we use the dataset provided in SemEval. The dataset consists of tweet id's which are annotated with positive negative and neutral labels. We will be using AFINN dictionary for polarity. Apart from these we will be using Wordnet to look for synonyms if a word is not directly found in the AFINN dictionary. We will also be using emoticon dictionary, acronym dictionary, tweet downloader, tweet NLP, LibSVM, BeautifulSoup python, SciKit and SVMUTIL python libraries.(Links are available in the reference section.)

Solution Statement –

The solution to the problem is pretty straightforward. For each given username, we will classify the tweets either positive, negative or neutral.

The input will be passed using twitter APIs and parsed using tf.tokenizer(s) that divides a string into substrings by splitting on the specified string (defined in subclasses).

Baseline model –

In the baseline approach, we will –

- Clean the tweets
- Assign probabilities – Positive, Negative, Neutral
- Create a feature vector of tokens which can distinguish the sentiment of the tweet with high confidence. For example, presence of tokens like am happy!, love love , bullsh*t ! helps in determining that the tweet carries positive, negative or neutral sentiment with high confidence.(Emotion determiners)

Feature based model –

In our feature model also we will perform the same pre-processing steps as we used in Baseline model (listed in project design). After that we will obtain polarity score using AFINN dictionary. If a word is not directly, then we get all the synonyms from Wordnet and will try to match those synonyms in AFINN. If the word is still not found then we will do another search in senti-wordnet dictionary and if it is present then we assign it a score (between -1 to +1).

For all our experiment we will be using SVM (support vector machines) model.

Evaluation Metrics –

As listed above we will be using SemEval dataset. The dataset consists of tweet id's which are annotated with positive negative and neutral labels. The dataset is already divided into three sets: Training, Development and Testing. This will be used to evaluate and calculate accuracy of the model.

Project Design –

For this project, I am solving one of the problem sets they have in their course CS50. But what they have asked to do is that they have a list of positive and negative words and if the word matches in the positive word list they give it a score of +1 and if word matches in the negative word list, a score of -1 is given.

But this has many limitations. For example consider this statement – "Yeah, right!" .This statement is a negative statement but both 'yeah' and 'right' comes in the positive list and hence this would be rated as a positive tweet.

To overcome these many limitations we will do the following –

1. Pre-process the data – we do this by first downloading the tweets provided by the twitter and we tokenize them. However twitter allows limited number of words. So people use acronyms, hashtags etc. excessively.
   I have selected following things to be removed from data – Non English tweets, remove URLs, remove @, remove repeated characters (goooooood, hungggggrry etc.), numbers(numbers have no use in measuring sentiments)
   Also for emoticons which play important role in determining the emotion of a tweet, we will be replacing them by their sentiment polarity given in the emoticon dictionary.
2. We will then calculate probability of the phrases.
3. Train using SVM model

References –

- http://sentic.net/sentire2016ahlgren.pdf (Research On Sentiment Analysis: The First Decade)
- http://www.ijergs.org/files/documents/EMO-57.pdf (Paper on Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets)
- https://docs.cs50.net/problems/sentiments/sentiments.html
- https://www.brandwatch.com/blog/understanding-sentiment-analysis/
- https://en.wikipedia.org/wiki/List_of_emoticons
- http://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf
- https://blog.infegy.com/understanding-sentiment-analysis-and-sentiment-accuracy
- http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
- http://sentiwordnet.isti.cnr.it/
- https://wordnet.princeton.edu/
- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
- https://www.acronymfinder.com/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- http://scikit-learn.org/stable/index.html
- http://www-lium.univ-lemans.fr/sidekit/libsvm/libsvm_core.html