

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Advanced Machine Learning**

**Major Assignment – Visualising Tweeter Data**

Student ID: A00277436

*Brief Description: Objective is to use tweepy to get tweets of Donald Trump and then do sentiment analysis of those tweets: positive, negative, neutral.*

*Before Sentiment analysis, duplicates need to be removed along with stop words, special characters and numbers as well. First to annotate tweets using TextBlob and then with VADER API and to see how different the results are.*

Contents

[Introduction 3](#_Toc37679873)

[Design 6](#_Toc37679874)

[Implementation Details 6](#_Toc37679875)

[**Analysis** 7](#_Toc37679876)

[Visualisation 8](#_Toc37679877)

[Conclusion 13](#_Toc37679878)

[Appendix 1: Reflective Learning Log 14](#_Toc37679879)

[Appendix 2 : References 15](#_Toc37679880)

# Introduction

Sentiment analysis can show how positive or negative the text data is. There are many practical applications of this process. For example These reports could help companies in creating customer oriented strategies.

With the enhancement in Artificial Intelligence algorithms it is much easier now to handle and study textual data. Moreover these algorithms are getting high accuracy rates for their assumption of sentiments related to data.

Factors Related with Sentiment Analysis

There are three main factors:

1. Polarity

The polarity defines the phase of emotions expressed in the analyzed sentence. It ranges from -1 to 1 and goes like:

Positive

Neutral

Negative

On account of polarity the emotion and sentiment of the writer can be easily described. For

2. Subjectivity

Polarity alone is not enough to deal with complex text sentences. Sometimes the sentence needs more attribute analysis to check weather it is describing features or opinions on some object.

Subjectivity helps in determining personal states of the speaker including Emotions, Beliefs and opinions. It has values from 0 to 1 and a value closer to 0 shows the sentence is objective and vice versa.

3. Intensity

Along with the polarity and subjectivity another important factor is intensity. It defines how strong or weak the emotion is with respect to the context.

It contains the following range:

Strong

Medium

Weak

Neutral

4. Author

The person who has the ownership for that text or who has written the text content. The content may be a review, conversation, social media chat, feedback or response. This helps in managing a structured way for assigning expressions to a single entity.

TextBlob is an open source python library used for textual analysis. It is very much useful in Natural Language Processing and Understanding.

Features of TextBlob Library

1. Part of Speech Tagging: It is the process of tagging parts of a sentence based on their definitions. They could be verbs, adjectives

2. Sentiment Analysis: Analysing the emotion behind the text content as a whole or as a part.

3. Noun Phrase Extraction: Extracting phrases whose head are a noun or pronoun.

4. Text Classification: Classifying text based on multiple factors.

5. Sentence Tokenization: Segmenting a sentence into parts.

6. Lemmatization: Finding the root words so as to define context of each sentence correctly.

7. Correction of Spellings: Helping in correcting the spellings based on patterns and learning.

8. WordNet Integration: TextBlob makes it easier to integrate with WordNet which is a database of English language words.

9. n-grams: N-gram is a sequence of any N words which helps in deciding the next word in a sentence.

10. Adding New Models or Languages through extensions.

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

Advantages of using VADER

VADER has a lot of advantages over traditional methods of Sentiment Analysis, including:

It works exceedingly well on social media type text, yet readily generalizes to multiple domains

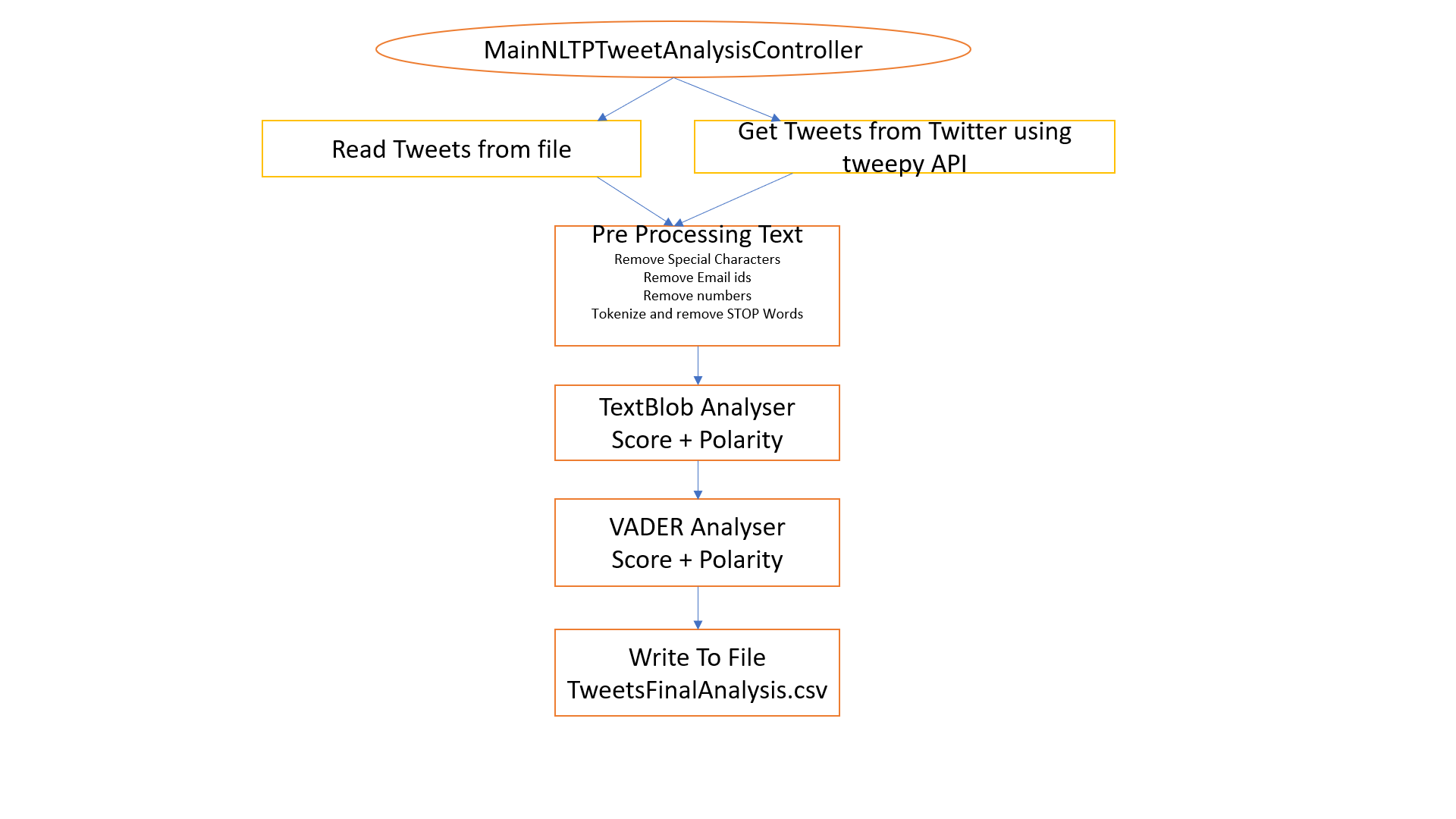
It doesn’t require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon

It is fast enough to be used online with streaming data, and

It does not severely suffer from a speed-performance tradeoff.

The results of VADER analysis are not only remarkable but also very encouraging. The outcomes highlight the tremendous benefits that can be attained by the use of VADER in cases of micro-blogging sites wherein the text data is a complex mix of a variety of text.

# Design



# Implementation Details

1. Source file of input of tweets is **trump\_data\_unique\_joined\_Removed\_PosNegNuet.csv**
2. Another source is directly from twitter website.
3. **MainNLTPTweetAnalysisController.py** is main python file which binds all other modules. Its main function is to read tweets from file **trump\_data\_unique\_joined\_Removed\_PosNegNuet.csv** and alose to get tweets from Twitter handle using tweepy API.
4. While reading tweets from both sources **Pre\_Processing\_Text.py** module is being called.
5. While pre processing text it mainly removes special characters and numbers along with email ids. It also tokenize and removes STOP words.
6. Next **TextBlob\_Analyser.py** module is called to get TextBlob scores and polarity of tweets.
7. Next **Vader\_Analyser.py** module is being called to get VADER score and polarity of tweets.
8. Finally all these raw tweets along with filtered tweets and TextBlob scores/polarity together with VADER sore/polarity is written to **TweetsFinalAnalysis.csv** file.

# **Analysis**

At the time of last run there were

Size of tweets: = 277

Tweets Read from the Web: = 94

Tweets Read from the File: = 183

TextBlob Score

Size of Positive tweets: = 101

Size of Negative tweets: = 43

Size of Neutral tweets: = 133

Positive tweets percentage: 36.46 %

Negative tweets percentage: 15.52 %

Neutral tweets percentage: 48.01 %

VADER Scores

Size of Positive tweets: = 147

Size of Negative tweets: = 82

Size of Neutral tweets: = 48

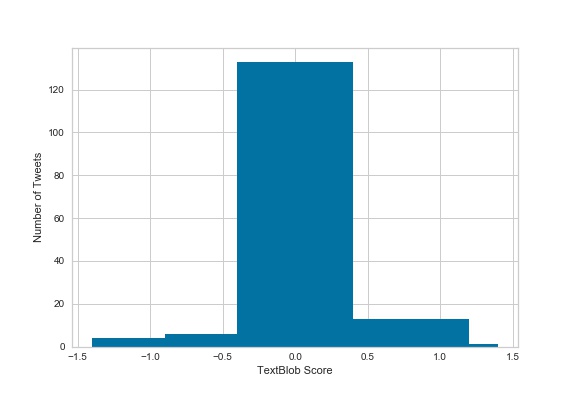
Positive tweets percentage: 53.06%

Negative tweets percentage: 29.60 %

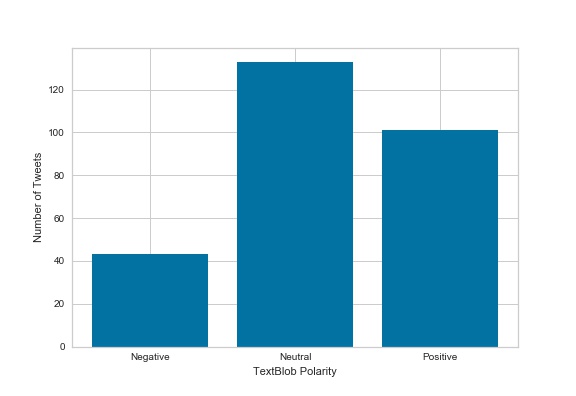
Neutral tweets percentage: 17.33 %

From the above figures it is very much clear that VADER scores are much more perfect as it is able to decide polarity of tweet with more efficient way because percentage of neutral tweets are much lower in comparison to TextBlob. Ambiguity in deciding which tweet s positive or negative is much clearer in case of VADER thus many of the neutral tweets of TextBlob are too labelled as negative or positive.

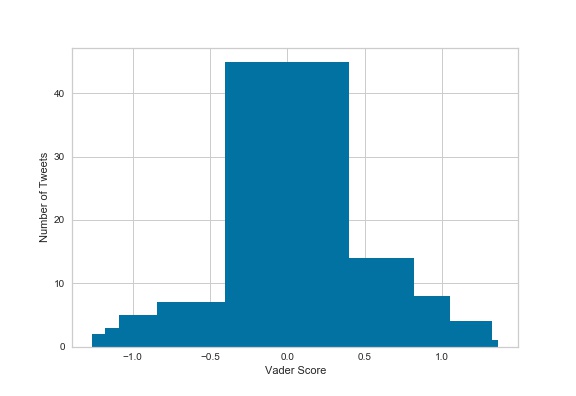
## Visualisation



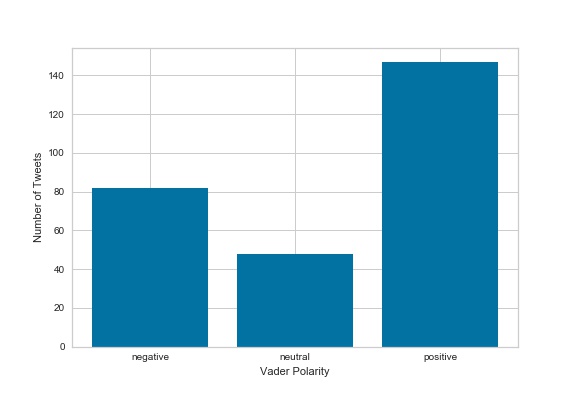
In TextBlob many of the tweets sentiment score is close to 0 and thus some where close to 50% of them are labelled as neutral. Same is evident from above graph.



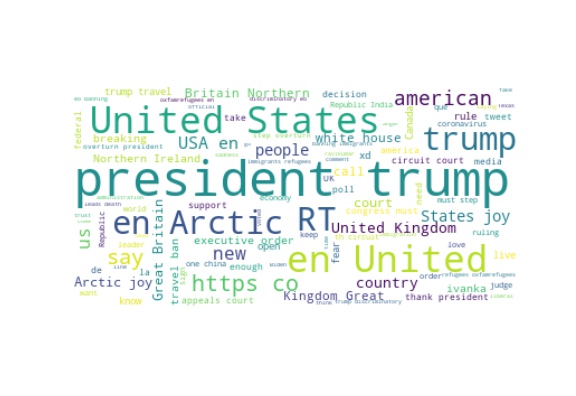
Majority of tweets are labelled as neutral by TextBlob.



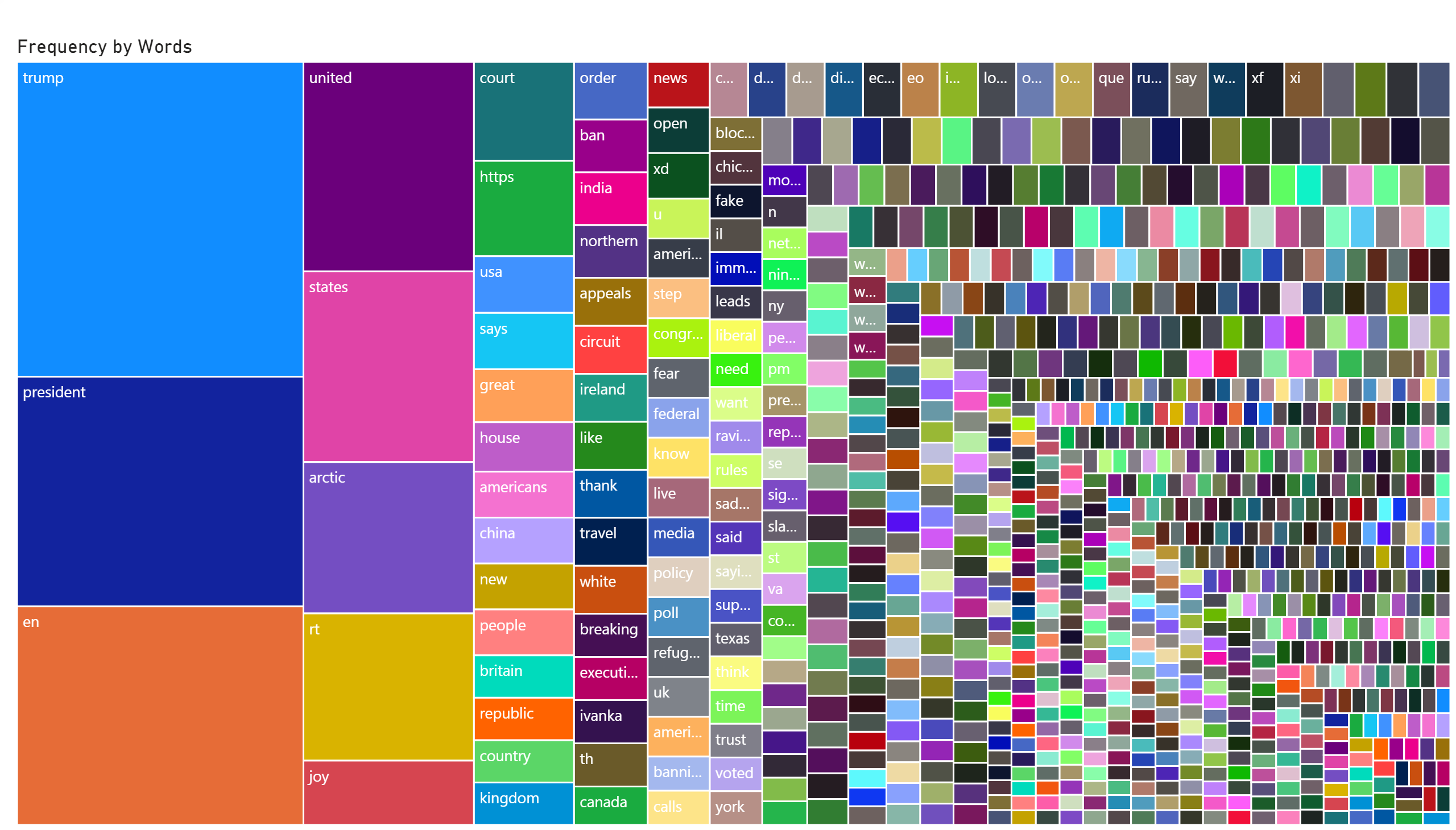
From above figure it appears that polarity score of most of tweets are close to zero but they are not zero and thus VADER is in better position to decide among them which one is positive and which one is negative.



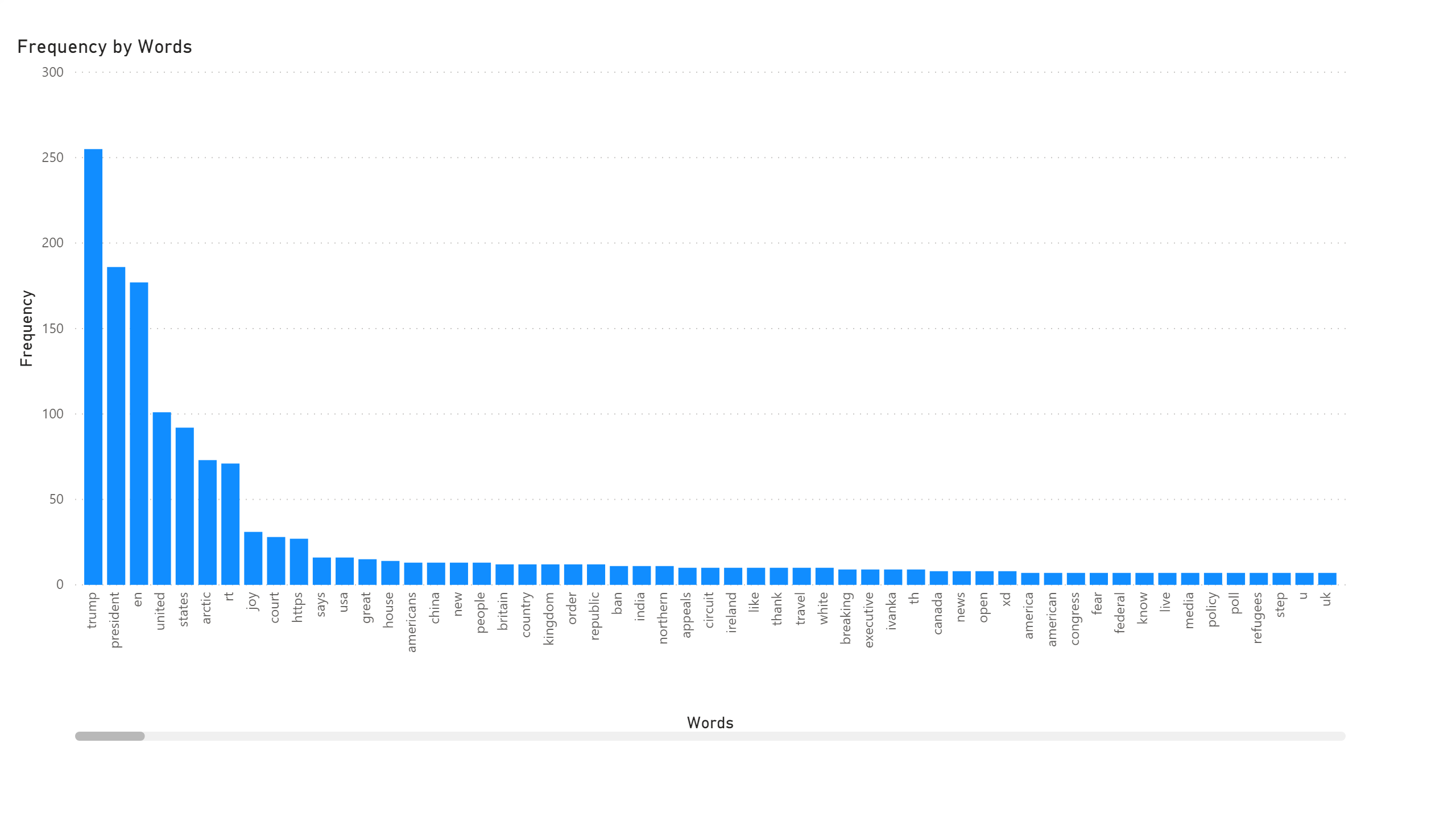
From above graph it is very much evident that neutral tweets are low where as good number of tweets are positive and then negative.



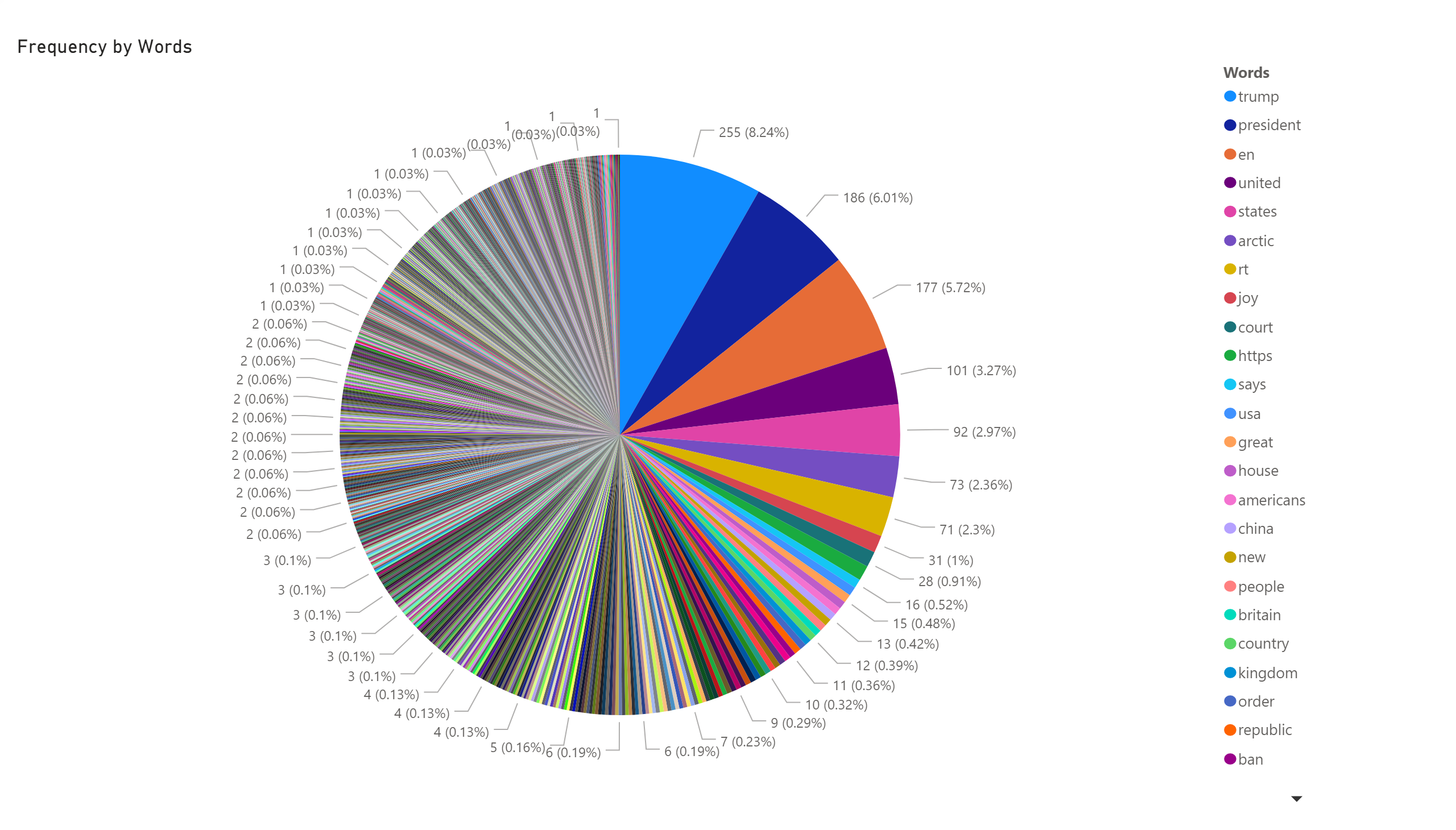
Above is prepared using WordCloud API. It is actually a cloud filled with lots of words in different sizes, which represent the frequency or the importance of each word. This is called Tag Cloud or WordCloud. To generate the same I wrote my own code which is there in WordCloud.py

[](https://app.powerbi.com/reports/3c4be8af-0bc6-4392-beff-00e0f480324e/ReportSection?pbi_source=PowerPoint)

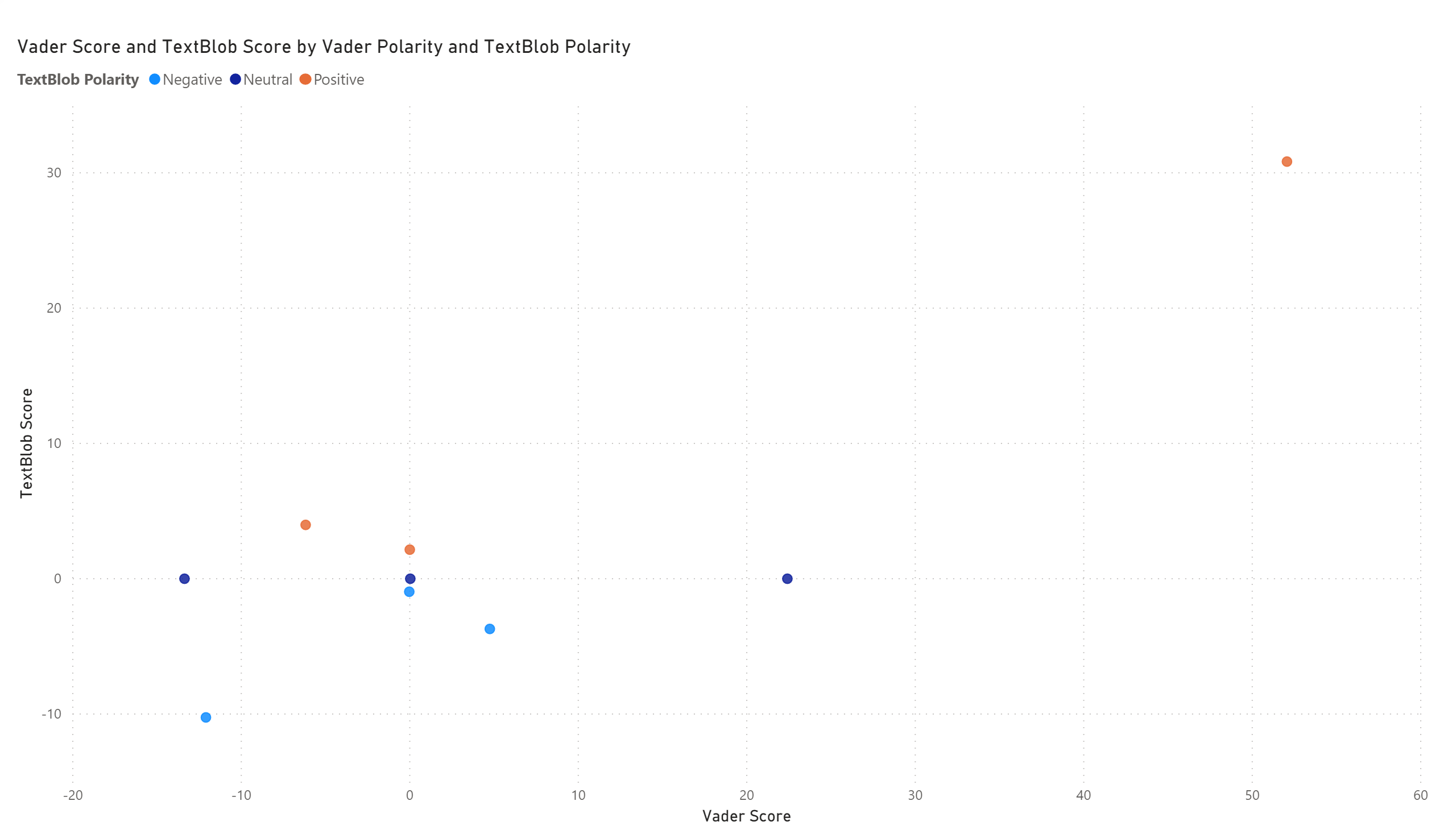
Here comes another figure that with special occupancy of word combined in all tweets together. From the above figure it is visible that the word “trump” occurred maximum and then next comes next word “president” and then “en” and so on. Depending on the frequency of word, space has been allocated to it. High occurrence leading to more space.

[](https://app.powerbi.com/reports/f7a0f572-73a5-4b05-a674-07a683fbf68e/ReportSection?pbi_source=PowerPoint)

Above graph shows frequency chart of words in tweets. Developed using Power BI.

[](https://app.powerbi.com/reports/0051952e-d729-4d4b-b92c-04f782c1966f/ReportSection?pbi_source=PowerPoint)

The above graph shows overall percentage of words in overall tweets. Developed using Power BI.

[](https://app.powerbi.com/reports/53c429c3-b671-4fab-bef8-050b7e200573/ReportSection?pbi_source=PowerPoint)

Above is a scatter plot between Vader and TextBlob scores and polarity. This one too is developed using PowerBI tool.

# Conclusion

From the above analysis it is very much certain that VADER is better API in semantic analysis of tweets in comparison to TexBlob. Its more efficient in deciding wether the tweet is positive or negative instead of terming it as neutral.

# Appendix 1: Reflective Learning Log

*Date:04 / 04 / 2020*

*Work Completed : Studied about VADER and TextBlob Semantic analytic API.*

*Understanding Achieved: 1st drafted design about the plan how I need to implement it.*

*Date:06 / 04 / 2020*

*Work Completed : Coded VADER and TextBlob API.*

*Understanding Achieved: Semantic analysis of tweets was implemented individually on tweets read from twitter using tweepy.*

*Date:08 / 04 / 2020*

*Work Completed : Coded VADER and API.*

*Understanding Achieved: Code to read tweets from file too was added to the project.*

*Date:11 / 04 / 2020*

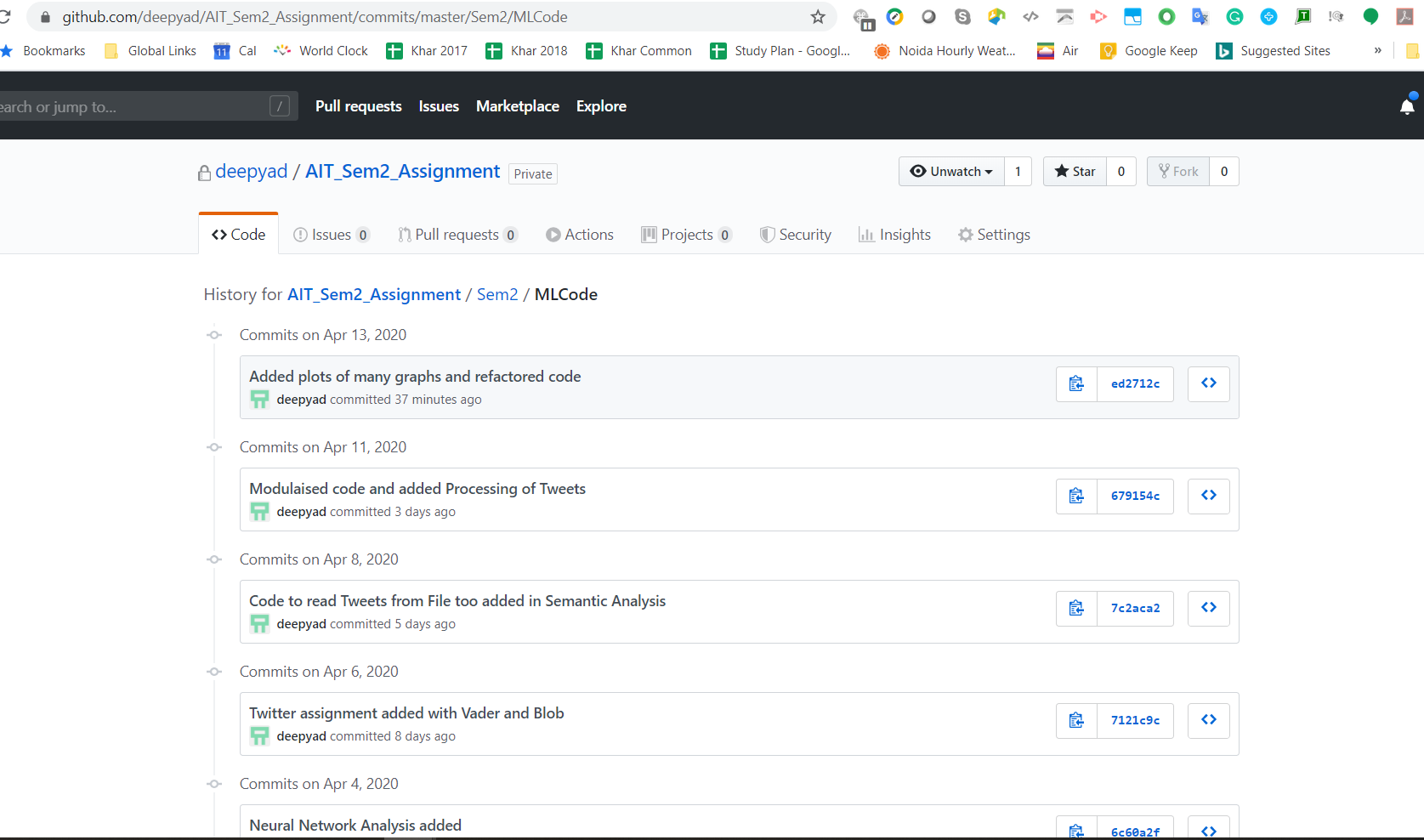
*Work Completed : Modularised code further.*

*Understanding Achieved: Modularisation of code with its functionality.*

*Date:13 / 04 / 2020*

*Work Completed : Added many figures and plot to the project*

*Understanding Achieved: Visualisation of tweets polarity and their score.Visual comparison betweek VADER and TextBlob APIs.*



# Appendix 2 : References

1. TextBlob

<https://www.presentslide.in/2019/08/sentiment-analysis-textblob-library.html?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com>

1. Sentiment Analysis using TextBlob <https://www.datacamp.com/community/news/sentiment-analysis-with-textblob-fdhk7okaa86>
2. WordCloud <https://www.datacamp.com/community/tutorials/wordcloud-python>
3. Text Analytics using NLTK <https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk>
4. VADER <https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f>