**ASSIGNMENT 1**

**SENTIMENTAL ANALYSIS TOOLS EVALUATION**

**Introduction**

Huge amount of data is generated on the social media platforms like Twitter & Facebook which is generally unstructured and has valuable hidden insights when correctly analyzed. The data can include tweets, comments, likes, opinion, feedback useful to drive business decisions when interpreted to its potential. Text analysis solves this problem of vast seemingly meaningless unstructured data to be transformed into valuable information.

**Problem Definition:**

The problem statement is to collect the vast amount of data available on the social media platforms like Twitter, Facebook, Instagram which is a representative of what the people think about Artificial Intelligence on these platforms. The tweets, comments from these platforms cannot be analyzed manually as it is an enormous pool of data which, needs to be collected and then analyzed using tools such as NLKT and SentiStrength which give the sentiment of the text in hand helping in deciding the public opinion about a topic. It also helps to analyze the current trending things about the topic, sentiment on the topic etc.

We need to evaluate how the two tools NLKT and SentiStrength behave on the data collected and to interpret their results to determine the best tool among these 2 for the sentiment analysis. Which of these 2 tools is the best & in what scenario for the sentimental analysis is a question whose answer needs to be known at the end of the analysis.

**Data Description:**

The Data about Artificial Intelligence, one of the trending topics today has to be collected from social media for its sentiment analysis. The data collected is “tweets” from the Twitter on this topic, cleaned before finally being analyzed by the tools. The data is fetched from Twitter using tags API scraping the tweets on AI to be downloaded in a CSV file. The data is a collection of tweets on AI meant to be analyzed to figure out the public emotion on AI.

**Methods:**

The Tweet collection is an unstructured text data which cannot be directly used to figure out the public sentiment about the topic. The data needs to be cleaned for it to be processed by the tools. Cleaning of data includes removing https links, handle names of people tagged, removing hash tags, removing duplicate tweets(re-tweets) etc. before moving ahead with the sentiment analysis.

All this pre-processing is done to make the data clean which helps to get accurate results. This clean data also consists of sorting meaningful tweets in a haystack of tweets most of which are random and do not make sense. After sorting the tweets to a bunch of meaningful tweets data is ready to be analyzed by the tools whose performance is to be evaluated. All these steps are crucial in making the computer understand the human language better along with deriving meaning of it like humans based on contexts, usage etc.

**Analytical Methods:**

There are 2 analytical methods used for sentiment analysis.

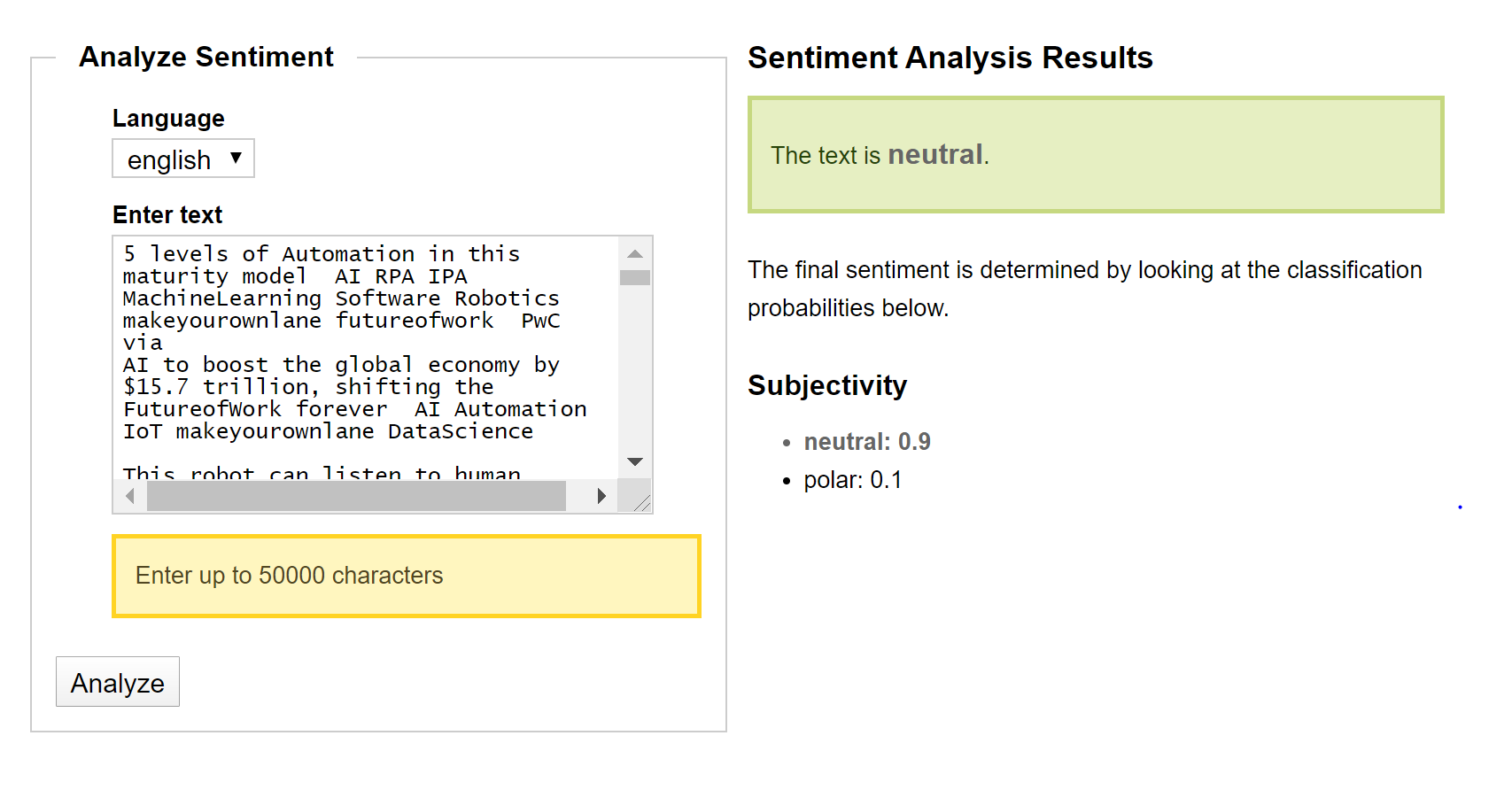
1. NLTK (Natural Language Tool Kit Text Classification using Python):
2. SentiStrength

There are 2 types of approaches for doing Natural Language Processing Sentiment Analysis.

1. Machine Learning Approach, where the model is pre-trained with labels positive, negative, neutral and the incoming text is broken into a pair of 2 or 3 or more words to get the overall sentiment of the sentence from the text.
2. Lexicon based Approach, where each word has a sentiment along with its polarity pre-defined and the overall sentiment is the aggregate of the sentiment of each individual word together with grammatical structure giving the maximum positive & negative sentiment of the text.

**NLTK:**

NLTK does classification of the text and 2 parameters quantify the results of sentiment analysis. Subjectivity and Polarity defining the sentiment. Neutrality is determined first and then Polarity if the text is not neutral. Here the tweets used show a Neutral Sentiment.



**Interpretation of the Analysis for NLTK:**

The NLTK gives output Subjectivity and Polarity where Subjectivity is how much of it is public opinion or sentiment. Objective refers to facts or figures. Here it suggests the tweets are more of opinion being neutral. The value of subjectivity is float which ranges from 0 to 1. The output represents a public opinion which is neutral.

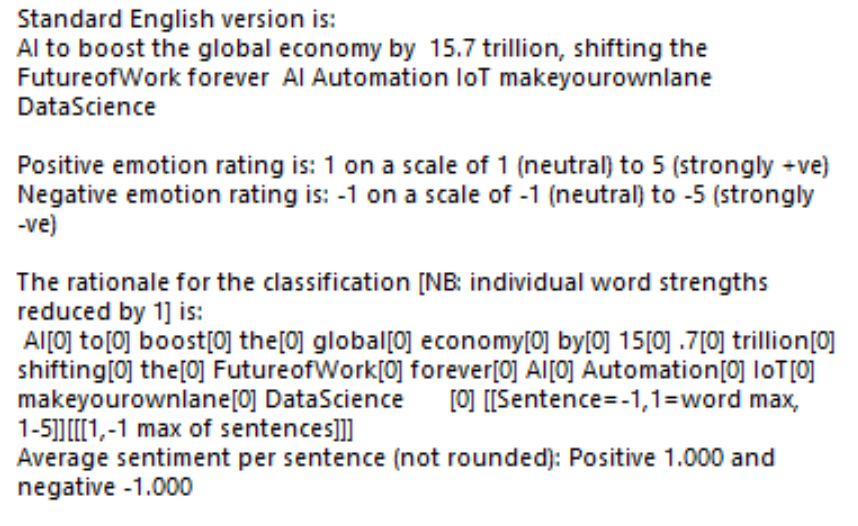
Polarity refers to whether the text has a positive, negative or neutral sentiment with the strength of its sentiment more positive or more negative etc. Polarity ranges from -1 to 1 with -1 being most negative and +1 being most positive.

We do not see the polarity here because the text is classified as neutral, if this was not the case it would give how much of the text is positive and how much negative.

It has a limitation of only 50000 words to be used for analysis so the data has to be fitted in such a way that it does not go beyond this limit.

**SentiStrength:**

SentiStrength uses Lexicon Approach wherein each word is already assigned a subjectivity and polarity and the overall sentiment is the total of all the scores of sentiments of each word. Using the tweet data to find the public emotion we see that SentiStrength gives individual words a sentiment score from -1 to -5 from negative to least negative and on the other side of the scale 1 to 5 least neutral to most positive. The tool has multiple tabs to do text analysis including special tab for text mining, convert text, sentimental analysis options etc. which can be very useful in some applications. There is no limit to the number of words it can handle. Passing a text file as a input is possible for analysis.



**Interpretation of the results of SentiStrength:** SentiStrength has given the sentence an overall sentiment of neutral -1,1. The overall classification for this text is the maximum positive and negative strength given in square brackets [-1,1]. Each word has also got its own sentiment score. It is very useful when analyzing each sentence as a separate entity for its sentiment.

**Conclusion:**

Having analyzed the tweet data collected from Twitter I see that NLTK does an overall text classification of it being neutral, positive or neutral whereas SentiStrength comes out with the sentiment for each sentence individually.

NLTK is better in this scenario as we get an overall sentiment for all the tweets combined. Having individual sentiment analysis would not be helpful as we are only interested in overall emotions of public on AI. SentiStrength makes it difficult to capture general opinion on a topic where huge amount of data is involved to be taken. Sentiment for each sentence needs to be again combined to get the composite public emotion which is a tedious task to do.

SentiStrength is useful in scenarios where individual tweets or comments sentiments are needed such as Tweet or a comment by a famous person on certain topic where individual tweet or comment can express different sentiment at different times.

The sample is a representative of the public opinion/emotion towards AI as it has been sorted from large number of tweets and covers tweets from broad spectrum of fields like medicine, technology, sports etc. in which AI is used and the people expressing their opinion which can be considered as representation of the population of tweets.