**Technical Explanation of Lexicon Based Model**

**Proposed Model**

Tweets

Data Pre-Processing

Manual Labeling

Label Verify (Positive, Negative, Neutral)

Threshold

Lexicon

Threshold >0.05 (Positive)

Threshold < -0.02 (Negative)

Others (Neutral)

Scores (1-5)

-Precision

-Recall

-F1-Measure

-Accuracy

Evaluation Measures

**1- Dataset Acquisition**

Arabic tweets dataset was acquihired from the twitter. This dataset was in the sentence (tweet) and not labeled to any class. There were required to label the dataset to get the sentiments. The whole dataset was labeled manually by reading and analyzing each tweet which was very time taking task. It was also not the complete satisfactory results because sometimes machine understand the sentence on the basis of the trained model as compare to the human being analysis.

The major problem was faced to label the neutral class in the dataset because it was very confusing to identify the tweet is neutral or belong to the other class. For the experimental analysis this research study performed the label annotation.

In the dataset the tweet length was also counted to see the number of words in each tweet. That can define the strength of the tweet with acquiring the keywords.

In the lexicon model this research study used the same annotated dataset performed manually. But faced some issues due to the confusion in labeling and this issue was resolved to count the words which have more relevant to the defined classes like positive, negative and neutral.

**2- Data Pre-Processing**

Data pre-processing was the key point to handle the whole dataset used for the model sentiments. After dataset acquisition many steps performed to prepare the required dataset. The following point very necessary to apply:

* Lexicon score for word was taken from the tweet
* These score was added into the feature of the tweet
* Padding approach was used to handle the different length of the tweet
* Obtained features were available to pass the model
* Feature vectors obtained can be used in the machine learning model training

The sample data file with the screenshot can be seen in the figure 1 below which shows the words with the calculated score, but his process did on the run time and to show here the sample file is saved.

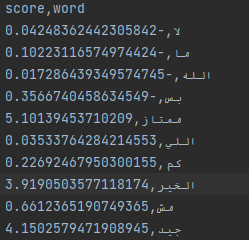


Figure : Tweet words with their score calculated

**3- Lexicon Approach**

After obtaining the pre-processed dataset the lexicon based approach was applied. Following points were covered in the lexicon approach such that:

* Import some necessary libraries like pandas that treat with the data manipulation
* Load the cleaned final dataset
* A function created which calculate the score values
* Each word was compared either it available in the lexicon or not
* The lexicon is the dictionary where many words available and here each word is compared to calculate the score values
* A threshold value defined such that threshold >0.05 results will be positive, threshold <-0.02 results will be negative and the others all considered as neutral.
* Final results were saved in the csv file

**4- Evaluation Results**

The evaluation of the model was performed to calculate the precision, recall, f1-measure and the accuracy for the model. These results were calculated to see the model performance. Example results can be seen below in figure 2:

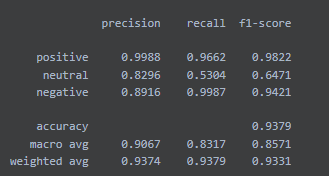


Figure : Lexicon based results