NLP Task to Calculate Human Emotional Intensity using Statistical Techniques

> The dataset consisted of a lot of swear words, spelling mistakes, unnecessary noise (Twitter Usernames) and punctuation words. So, clearly, text-preprocessing was the first step.

Text pre-processing involved the following steps :-

1. Dealing with chat words - Since people often use informal language on Twitter, the first wise step was to convert chat words (such as LOL, ASAP, LMAO) into regular words.

2. Lowercasing - Lowercasing is done so that the CountVectorizer (used in Bag of Words to tokenize all the unique words from the corpus) does not tokenize two different versions of the same words (such as “phone” and “Phone” into two different tokens. This helps save computational space.

3. Removing Twitter Handles - Since Twitter handles contribute nothing to the emotional intensity of the user, and also contain special characters (@ symbols), it is best to remove them, to ensure a noise-free dataset.

4. Removing URLs - URLs also do not have any correlation with the emotional intensity of the user, so they also must be removed.

5. Removing Punctuations - Done to remove punctuation marks, as they do not contribute much to the overall emotional analysis.

6. Removing white spaces - White spaces are removed to prevent white spaces from being tokenized into the vocabulary.

7. Removing Stop Words - Since stop words provide zero to no contribution to the overall sense of the sentiment of the tweet, they are also removed.

8. Lemmatization - This helps prevent the tokenizer from creating two separate tokens for words which have the same root word (eg:- irritating and irritated have the same root word - irritate).

9. Tokenization - Unique words from the corpus are tokenized and are stored in the vocabulary.

10. POS Tagging - POS tagging helps the model identify adjectives and verbs which will help infer the emotional state of the user more accurately.

> Statistical Model
The statistical model we will be using for this process is going to be the simple technique of BOW (Bag of words). This will help calculate whether or not a particular set of words have occurred in the text of the tweet or not.

Moreover, we can employ the use of tf-idf to calculate the logarithm of the frequency of the words in the document divided by the total number of occurrences in the document.

Lastly, the array gathered from the Bag of Words technique is passed on to the model along with the answers (y-column) and passed into a model using Linear Regression to try and predict the emotional intensity of all the users in the dataset.

> Deep Learning Model - I wanted to use embeddings (word2vec technique) to try and capture the semantic sense of the data and provide more accurate results. Unfortunately, I could not get enough time to make a separate model which employed the use of DL models to make user emotional intensity predictions.