PREPROCESSING

A. Tokenisation -

After downloading the tweets using the tweet id’s provided in the dataset, we first tokenize the tweets. This is done using the Tweet-Tokenizer provided by the NLTK library. It is important to note that this is a twitter specific tokenizer in the sense that it tokenizes the twitter specific entries like Emoticons, Hashtag and Mentions too. After obtaining the tokenized tweet we move to the next step of preprocessing.

B. Replacing Emoticons -

Emoticons play an important role in determining the sentiment of the tweet. Hence we replace the emoticons by their sentiment polarity by looking up in the Emoticon Dictionary that we have created using Emoji Sentiment Dataset used in the ‘emoji’ python library.

C. Remove Url-

The url’s which are present in the tweet are shortened using TinyUrl due to the limitation on the tweet text. These shortened url’s did not carry much information regarding the sentiment of the tweet. Thus these are removed.

D. Remove Target-

The target mentions in a tweet done using ’@’ are usually the twitter handle of people or organisation. This information is also not needed to determine the sentiment of the tweet. Hence they are removed.

E. Replace Negative Mentions-

Tweets consists of various notions of negation. In general, words ending with ’nt’ are appended with a not. Before we remove the stopwords ’not’ is replaced by the word ’negation’. Negation play a very important role in determining the sentiment of the tweet. This is discussed later in detail.

F. Hashtags-

Hashtags are basically summariser of the tweet and hence are very critical. In order to capture the relevant information from hashtags, all special characters and punctuations are removed before using it as a feature.

G. Sequence of Repeated Characters-

Twitter provides a platform for users to express their opinion in an informal way. Tweets are written in random form, without any focus given to correct structure and spelling. Spell correction is an important part in sentiment analysis of user-generated content. People use words like ’coooool’ and ’hunnnnngry’ in order to emphasise the emotion. In order to capture such expressions, we replace the sequence of more than three similar characters by three characters. For example, wooooow is replaced by wooow. We replace by three characters so as to distinguish words like ’cool’ and ’cooooool’.

H. Numbers-

Numbers are of no use when measuring sentiment. Thus, numbers which are obtained as tokenized unit from the tokenizer are removed in order to refine the tweet content.

I. Nouns and Prepositions-

Given a tweet token, we identify the word as a Noun word by looking at its part of speech tag given by the tokenizer. If the majority sense (most commonly used sense) of that word is Noun, we discard the word. Noun words don't carry sentiment and thus are of no use in our experiments. The same reasoning go for prepositions too.

J. Stop-word Removal -

Stop words play a negative role in the task of sentiment classification. Stop words occur in both positive and negative training set, thus adding more ambiguity in the model formation. And also, stop words don’t carry any sentiment information and thus are of no use to us. We create a list of stop wo