Milestone report

Problem Statement:

Classifying Twitter sentiments based on positive/negative/neutral labels using NLP

Background:

Labels represent the sentiment of a text. These labels helps in analyzing the travellers' experience in using the US airlines - the actual feedback is from the text itself. The primary goal is to build a classifier that understands the text and assign an appropriate label to it. Background Each sentiment label is categorized as positive,negative and neutral. Therefore we will have a supervised, multi-class classifier with actual text as input. Background Each sentiment label is categorized as positive,negative and neutral. Therefore we will have a supervised, multi-class classifier with actual text as input. This piece of code is an exploration of Natural Language Processing (NLP). The goal of predicting the labels given a piece of text will deal with plenty of NLP topics such as NGrams, NamedEntityRecognition, Bag of Words. Finally, we arrive at a dataframe and we will be employing different machine learning algorithms on it to come up with the best model. The dataset contains the travellers' reviews on US Airlines in February 2015. There are a total of 14640 records of different airline reviews.

Dataset:

The dataset has been uploaded to the github repository.

tweet_id - user's unique ID on the twitter , int64
airline_sentiment - sentiment labels, string
airline_sentiment_confidence - degree of the sentiment, float64
negativereason - reasons out the negative sentiment, string
negativereason_confidence - degree of the negative sentiment, float64
airline - Specified name of the airline, string
airline_sentiment_gold - sentiment labels of airline's gold members, string
name - Specified name of the user on twitter, string
negativereason_gold - reasons out the negative sentiments of airline's gold members, string
retweet_count - twitter's retweet count,int64
text - user's reviews, string
tweet_coord - latitude and longitude coordinates, int64
tweet_created - time when the tweet was posted, string
tweet_location - place where the tweet was posted, string
user_timezone - time zone where the review was being posted, string.

Data Wrangling

The data is stored in a .csv file. It is downloaded and is stored in the form of pandas dataframe. It has a total of 14640 rows and 15 columns. The columns tweet_id and retweet_count are integers, airline_sentiment_confidence and negative_reason_confidence are floats while the rest of them are strings (objects). Also, user_timezone column has only 9820 records therefore it is missing some records.

The missing values in user_timezone is filled by 'forward fill' method of 'fillna' function.

After some statistical interference, it is noticed that the data is imbalanced. Hence the data is downsampled without replacement to match the minority class.

Sometimes we tend to use contractions in our language to convey the message which is easier than typing the whole word out. For example, instead of " we will" we might type it out as " we'll ", which is commonly referred to as the texting language. However, this makes it harder for the classifier to determine and analyze the sentiment. Hence I have come up with a map of some common contractions and have written a customized function 'expand' to expand the words if it come across any of them listed in the map. This helps in improving classifier's efficiency.

The function 'tweet_to_words' first expands the contractions, gets rid of words that have multiple repeating characters in them such as the word 'loooove' basically is 'love' and the user is trying to convey the degree of intensity by putting in multiple repeating characters. Then it gets rid of symbols, punctuations and numbers, stop words, words that have 'http' or 'www' in them, words whose lenghths are greater than 2. Finally the words are lemmatized which means that it converts all of them to their root words, for example, the word 'caring' has the lemmatized(root) word 'care'.

Inferential Statistics

The visual analysis shows that there are more negative tweets than the positive/neutral. So I did some statistical analysis to confirm the same. For simplicity, I have considered only the negative and the positive sentiments in the analysis.

```
Total number of negative sentiments: 9178
Total number of positive sentiments: 2363
The total entries in the field 'airline sentiment':14640
```

The central limit theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.

Central Limit Theorem assumes the following conditions:
i) Sample size should be "large enough".

Since the data is large (n > 30), this condition is met.

ii) np>= 10 and nq >= 10. The number of successes and failures of the samples should be greater than or equal to 10.

Success: There are 2363 positive sentiments out of 14640 > 10. Similarly, there are 9178 negative sentiments out of 14640 > 10.
Failure: There are 12277(14640 - 2363) sentiments which are not positive > 10. Similarly, there are 5462(14640 - 9178) which are negative > 10.
Therefore, this condition is met.

iii) Independence. Sample size should be less than or equal to 10% of the population size. We have seen this already in (i), hence this condition is also met.

iv) Randomization. Assuming that the samples are randomly chosen unless and until it is specified.
Hence CLT is applicable in this scenario and it is appropriate to use a significance test.

Null Hypothesis: Proportion of positive sentiments is equal to proportion of negative sentiments.

Alternate Hypothesis: Proportion of positive sentiments is not equal to proportion of negative sentiments.

I performed Bootstraping two sample test first and obtained the following result-

```
We can reject the null hypothesis with p-value:0.0
```

Then performed frequentist approach-

```
Proportion of positive sentiments (P1): 0.16
Proportion of negative sentiments (P2): 0.63

Combined sample proportion is 0.39415983606557375
Standard Error : 0.005711624850139768

p-value is : 0.00000
We can reject the null hypothesis.
```

After simulating the conditions for 10,000 trials, we got the same results following both the bootstrap and frequentist approaches, hereby concluding that the proportion of negative sentiments is NOT equal to positive sentiments. Hence the best recommendation that I found to balance this imbalanced data is to downsample the majority class without replacement (which is being mentioned in the above section).