Twitter Sentiment Analysis

Meenakshi Shankara
Bellevue University - Master of Science in Data Science
DSC680 Applied Data Science

07/23/2023

Table of Contents

INTRODUCTION

Twitter sentiment analysis analyzes the sentiment or emotion of tweets. It uses natural language processing and machine learning algorithms to classify tweets automatically as positive, negative, or neutral based on their content. It can be done for individual tweets, or a larger dataset related to a particular topic or event.

NLP - Natural Language Processing is one of the areas of artificial intelligence that works with the analysis, understanding and generation of living languages to interact with computers both orally and in writing, using natural languages instead of computer ones. We will be looking at some of the techniques in NLP.

Business Problem

Twitter Sentiment Analysis is important for understanding Customer Feedback, in reputation management, political analysis, crisis management, marketing research to name a few.

In this project, we try to implement NLP Twitter sentiment analysis models that help to overcome the challenges of sentiment classification of tweets. We will be building a model that will give us the most accurate prediction of the sentiment of the tweet.

DATA

Datasets

The dataset I will be working with is extracted from Kaggle website.

Twitter Sentiment Analysis | Kaggle

This is a sentiment analysis dataset of twitter. Given a message and an entity, the task is to judge the sentiment of the message about the entity. There are three classes in this dataset: Positive, Negative and Neutral. We regard messages that are not relevant to the entity (i.e., Irrelevant) as Neutral, although we will not be changing the sentiment value for this research. I will be using twitter_training.csv as the training set and twitter_validation.csv as the validation set.

Data Dictionary

There are four columns in each dataset.

Tweet ID – a unique tweet ID - Numeric

Entity – name of the product/service the user is tweeting about - String.

Sentiment – vibe of the message - String

Tweet Content – the text of the tweet – String

The training dataset has 74682 records while the validation dataset has 1000 records.

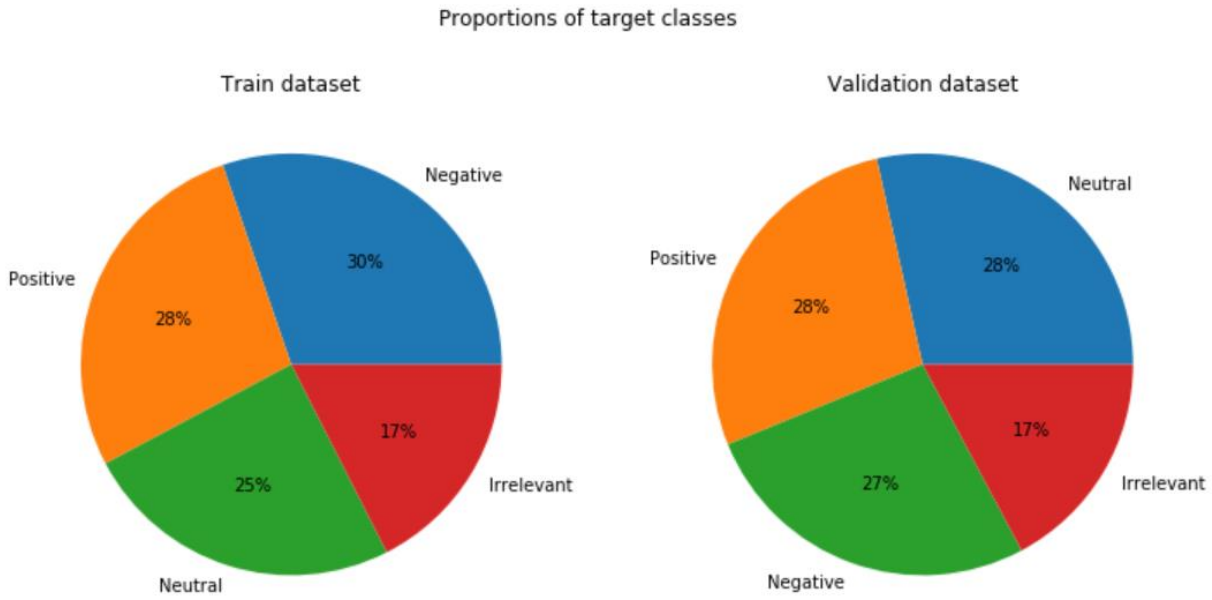
Data Preparation

The below transformations will be done on the dataset to prepare the data for modeling tasks.

- Remove all URLs, hash tags, usernames.
- Remove Emoticons.
- Remove all punctuation, symbols, numbers.
- Remove Stop Words
- make all text lower case.
- Handle NaN values
- Handle duplicate tweets.

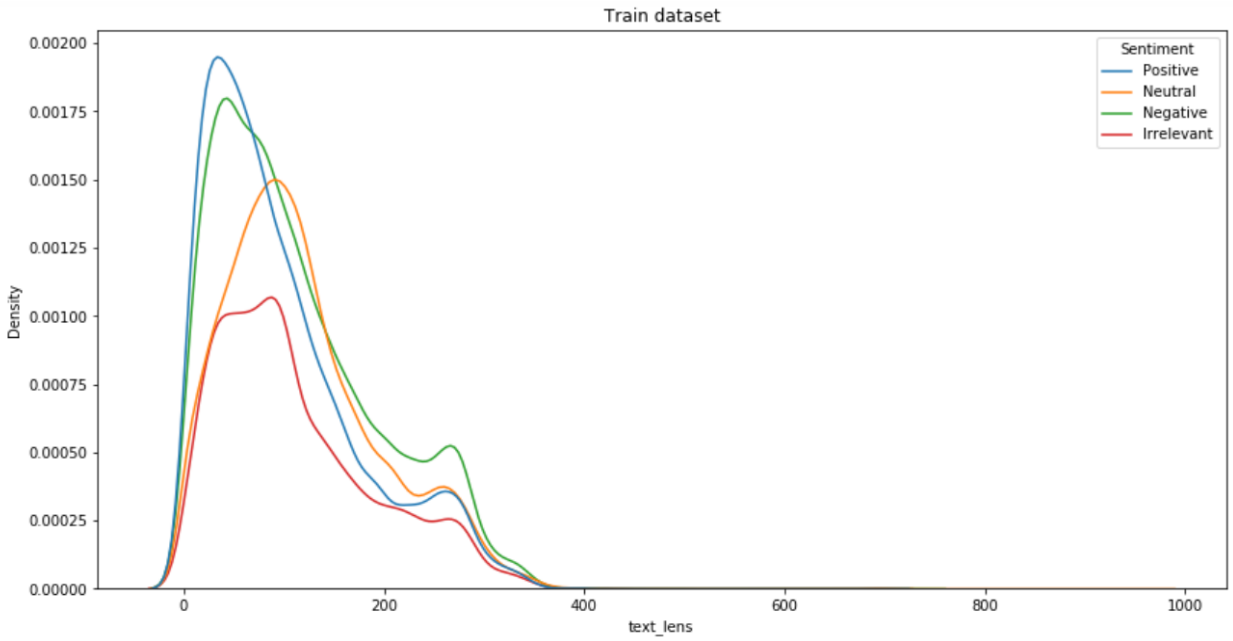
Data Visualization

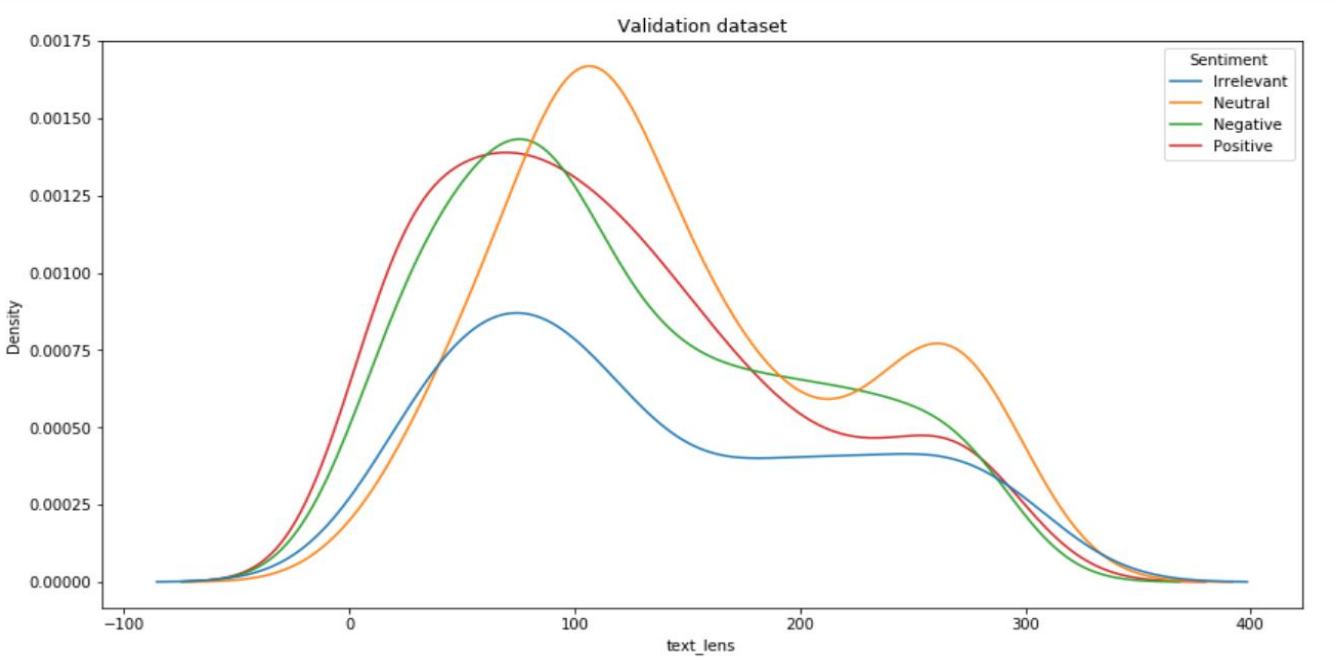
In this process, we will look at the distribution of the target values. In this analysis, we have 4 different sentiments, namely, positive, negative, irrelevant, and neutral. This is a way to gain more understanding of distribution of the target data in the training and the validation dataset.



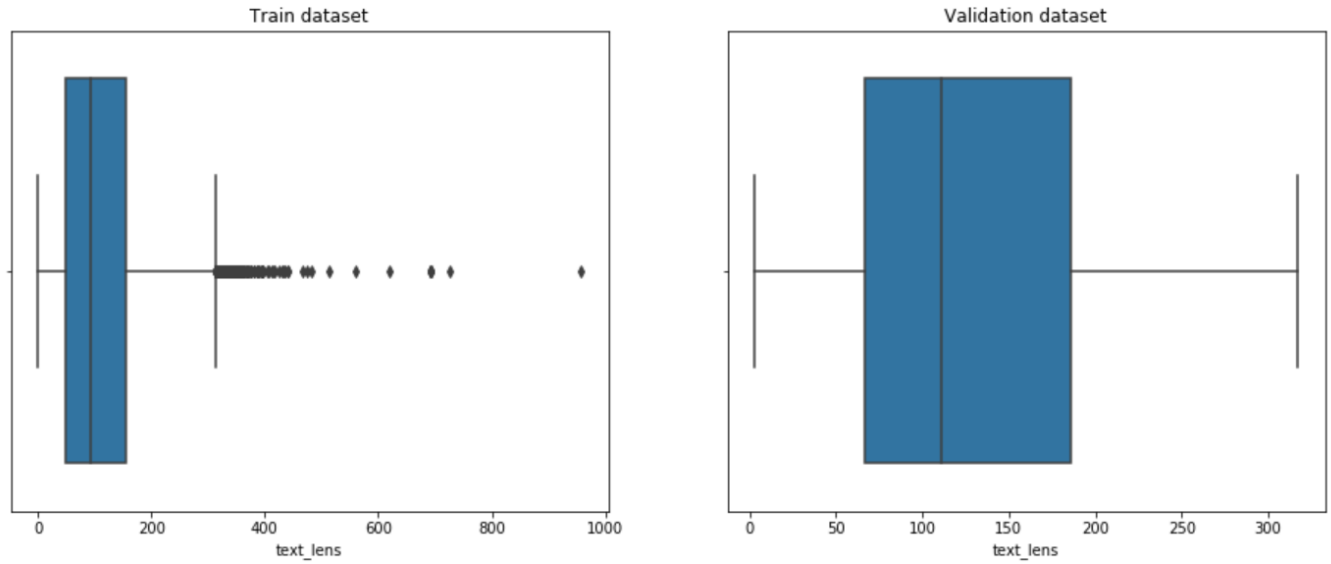
From the pie charts we observe that the target values distribution in both the datasets is like each other. We also observe that almost ~40-45% values are either neutral or irrelevant.

Next, we want to see if the lengths of the tweets in training and validation datasets are alike. The reason being if the lengths differ by a lot, then the accuracy of sentiment prediction may be affected by the words in the text.



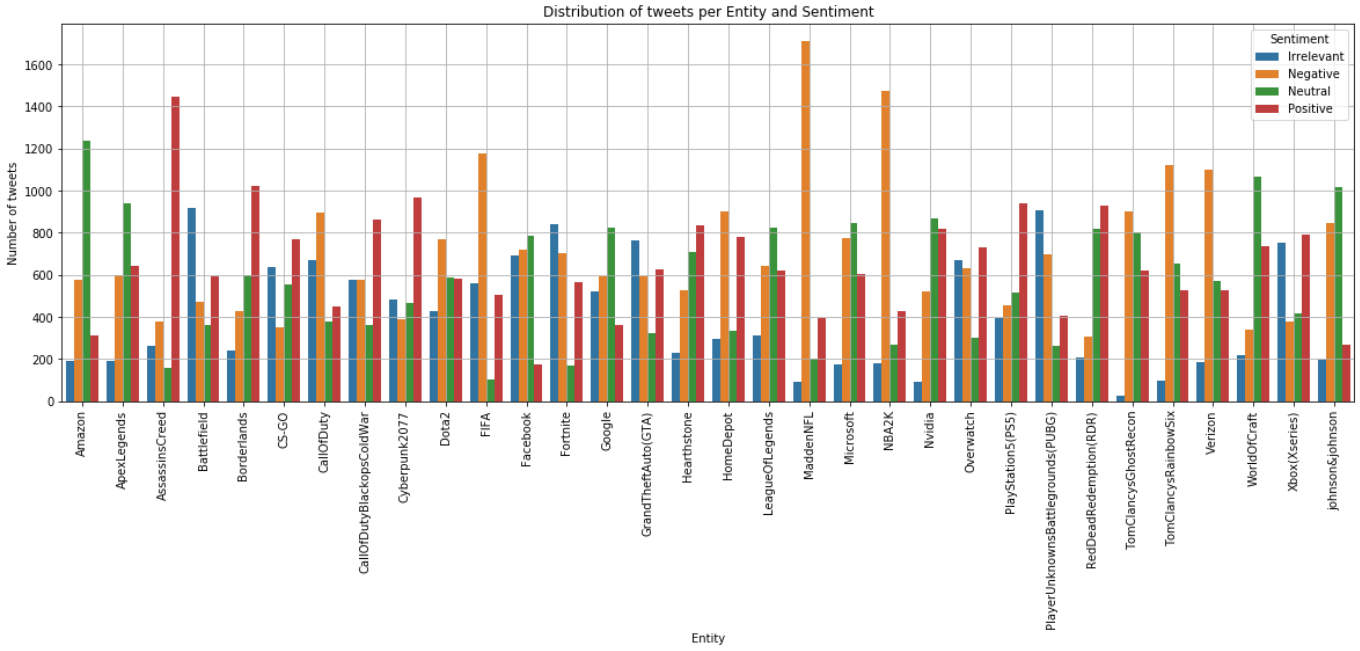


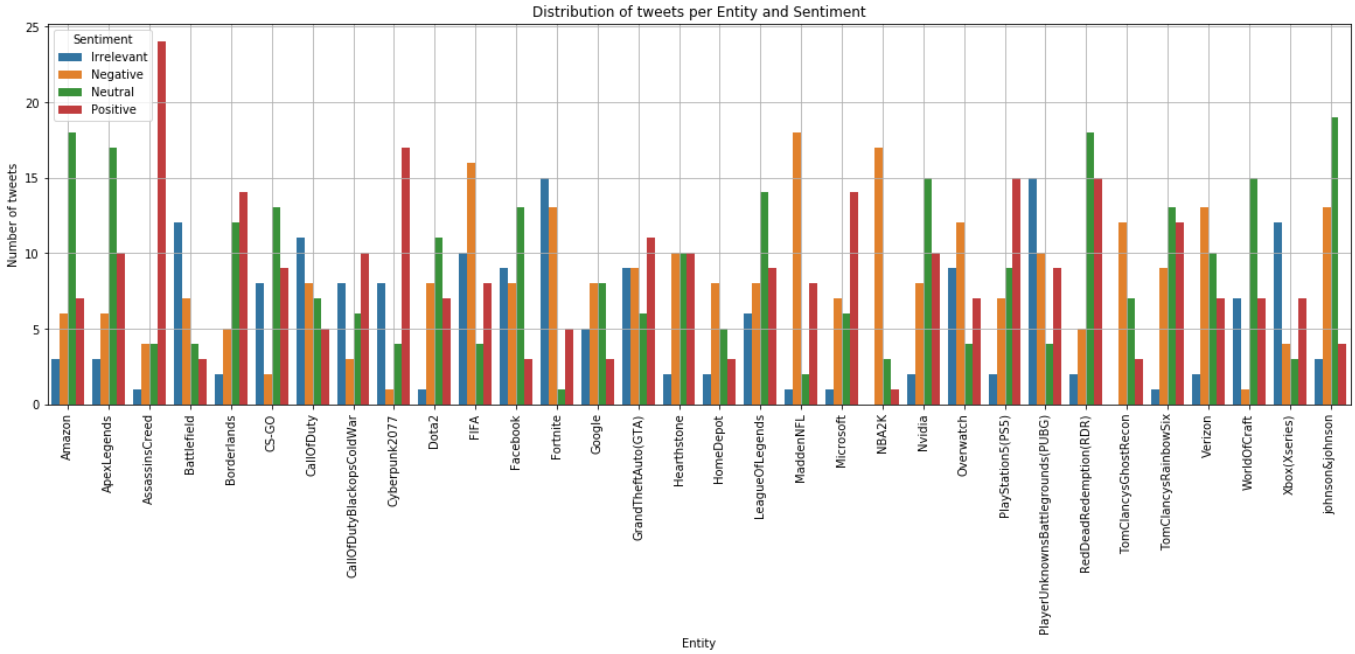
From the above graphs, we observe the text lengths of most tweets fall between 0-200 in both the Train and Validation datasets. The full range is similar as well, 0-400.



From the above graphs we observe that the tweet lengths in both the datasets have a similar range, although the training dataset seems to have some outliers. This is one area where we have to decide whether to remove the outliers or keep them.

Next, we are creating visualizations for analysing the distribution of sentiments for each Brand/Entity.





We observe that the distribution of sentiments for each entity is similar in both the datasets. For instance, we can see for brand ‘MaddenNFL’, it has more negative sentiments than positive sentiments in both the train and validation datasets.

The trend of sentiments in both the datasets is also similar. For instance, Madden NFL has the most negative sentiments and Assassins Creed has the most positive sentiments in both the datasets.

Modeling

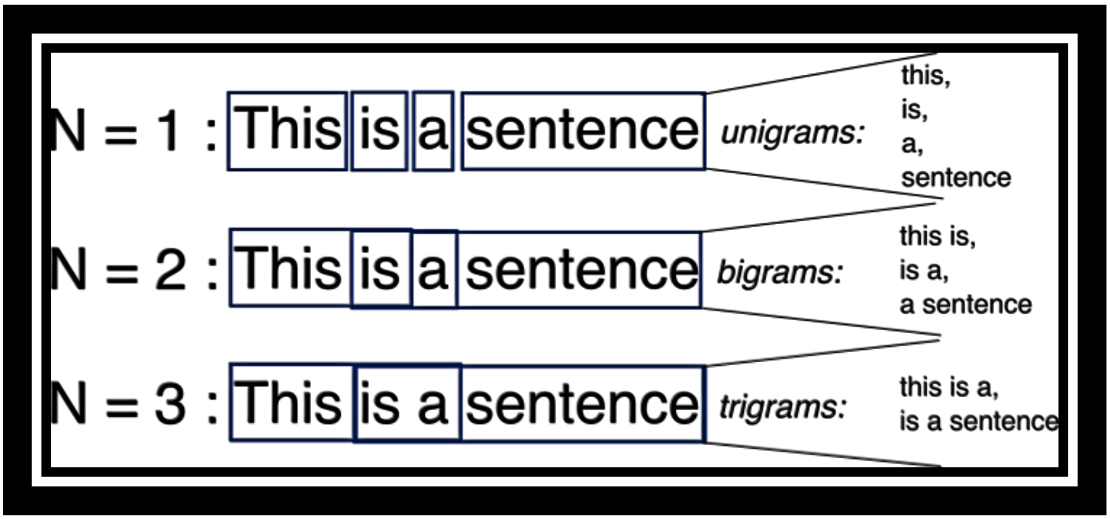
The following modeling techniques are used to determine which modeling technique works best on this dataset and the features that are mostly related or correlated to the happiness score.

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier
- Multinomial Naive Bayes

Bag of Words

We will be creating a Bag of Words with the n-gram value of 1 and then increase it to 4 to observe the models' behavior.

An N-Gram is a connected string of N items from a sample of text or speech. The N-Gram could be comprised of large blocks of words, or smaller sets of syllables. N-Grams are used as the basis for functioning N-Gram models, which are instrumental in natural language processing as a way of predicting upcoming text or speech.



Classification Report

For analyzing and comparing the results of each model, we will consider Precision, Recall, F1-Score and Accuracy.

1. Precision: Percentage of correct positive predictions relative to total positive predictions.

2. Recall: Percentage of correct positive predictions relative to total actual positives.

3. F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

Using these three metrics, we can understand how well a given classification model is able to predict the outcomes for some response variable.

4. Accuracy: The most common way to measure the accuracy of a classification model is by simply calculating the percentage of correct classifications the model makes:

Accuracy = correction classifications / total attempted classifications * 100%

Results

Bag of Words
Ngram = 1

Model	Data	Precision	Recall	F1-Score	Accuracy	
Logistic Regression	Train	0.84	0.84	0.84	83.65%	
	Validation	0.94	0.94	0.94	93.60%	√
Random Forest	Train	0.91	0.91	0.91	91.00%	
	Validation	0.96	0.96	0.96	95.70%	√
MultinomialNB	Train	0.76	0.75	0.75	75.01%	
	Validation	0.83	0.82	0.82	81.90%	
Decision Tree	Train	0.83	0.82	0.82	82.93%	
	Validation	0.9	0.9	0.9	90.20%	

With data fitting on Bag of words with ngram=1, Random Forest and Logistic Regression models performed the best with 95.7% and 93.6% accuracy on validation dataset respectively. Decision tree came third with 90.2% accuracy.

Multinomial performed the worst with 81.9% accuracy.

Bag of Words
Ngram = 4

Model	Data	Precision	Recall	F1-Score	Accuracy	
Logistic Regression	Train	0.92	0.91	0.91	91.43%	
	Validation	0.99	0.99	0.99	98.80%	√
Random Forest	Train	0.9	0.9	0.89	89.57%	
	Validation	0.96	0.96	0.96	96.40%	
MultinomialNB	Train	0.92	0.92	0.92	91.66%	
	Validation	0.98	0.98	0.98	97.60%	√
Decision Tree	Train	0.77	0.77	0.77	77.32%	
	Validation	0.9	0.9	0.9	90.20%	

Logistic Regression on Bag of words with ngram=4 is at 98.6% accuracy score on the validation dataset which is the highest among all the algorithms that we have researched.

Surprisingly, Multinomial came in second with 97.6% accuracy which is a huge jump from the previous modeling with ngram=1. Random forest was close with 96.4% accuracy.

Decision Tree remained at 90.2% accuracy.

Conclusion

A Twitter sentiment analysis determines negative, positive, or neutral emotions within the text of a tweet using NLP and ML models. Sentiment analysis or opinion mining refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often useful in generating a vast amount of sentiment data upon analysis. This data is useful in understanding the opinion of people on social media for a variety of topics.

From the various model results we find **Logistic Regression and Random Forest** to be stable candidates for building models for our analysis. **MultinomialNB** jumped from being the last to second in terms of accuracy for different combination of bag of words. This needs further analysis to refine the modeling solution.

Assumptions

The dataset contained null values and duplicate records. The decision to remove such records from the dataset may have affected the modeling results. But I was unable to find a logical way to fill in the null values without it affecting the final predictions.

Another assumption is that the target values in the datasets correctly represent the sentiment of the tweet. Usage of some words can mean differently in different contexts.

Limitations

The dataset considered for this prediction analysis contains only limited rows of ~71k rows in training dataset and about 1000 rows in the validation dataset. This data is only a small percentage of the actual real-time data. Therefore, the prediction accuracy could differ when used against the real-world data.

Challenges

The first challenge was to prepare the data appropriately for modeling. Extreme cleansing could result in fewer features and records available for the next steps.

There were a few data prepping steps like translating non-English tweets to English, that I was unable to perform due to package version incompatibility.

Recommendations

For this analysis, I have worked with Bag of Words for tokenizing the text. There are other tokenizing techniques available for usage. We observe that with different ngram values, the accuracy of the sentiment varies.

For the given data and the models that were used, we can predict the sentiment of the tweet with high accuracy. However, the recommendation would be to analyze the models with different tokenizing techniques and ngram values to find a better performing model.

Implementation Plan

The models Logistic Regression, Random Forest regression, and Multinomial NB can be implemented for predicting the sentiment of the tweets with very good results.

However, these models must be executed against real-world data for better evaluation.

Ethical Assessment

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions made from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project to a larger audience, the underlying assumptions and data considerations should be shared.

References

Twitter Sentiment Analysis | Kaggle

Logistic Regression Optimization & Parameters | HolyPython.com

Twitter Sentiment Analysis With Python | Introduction & Techniques (analyticsvidhya.com)

How to Interpret the Classification Report in sklearn (With Example) - Statology