# DSC 680 -Project 1 Milestone 3

## Vijaykumar Mehtre

## Answers to Q&A

**Questions:**

1. What other algorithms can be applied to the model?

Answer: We have used Logical Regression and SVM however we could also tried Decision Tree and random forests. Decision Trees are easy to interpret but can overfit on text data. Random Forests can help reduce overfitting by averaging multiple decision trees.

1. What are the challenges in text mining from Twitter ?

Answer: Text mining from Twitter faces challenges such as short text length, informal language, and noise from spam or irrelevant content. Additional difficulties include handling multilingual data, special characters (hashtags, mentions, URLs), and detecting sentiment, sarcasm, and context. Effective preprocessing, advanced NLP models, and real-time processing frameworks are essential to address these issues and derive meaningful insights.

1. What are the challenges in data cleansing ?

Answer: Data cleansing challenges include handling inconsistent formats, missing values, and noisy or duplicate entries. Robust techniques are essential to ensure data quality and reliability for analysis.

1. What kind of data needs to be filtered from the text ?

Answer: Text data filtering involves removing noise and irrelevant content like spam and advertisements, handling special characters and formatting issues, and eliminating stop words that do not carry significant meaning. These steps help to focus on relevant and meaningful parts of the text, improving the quality of the analysis. This ensures cleaner data for more accurate text analysis.

1. Are there any available data sources to mine data for twitter analysis?

Answer : Twitter API: Provides access to real-time and historical tweet data, including the Standard, Premium, and Enterprise APIs for different levels of access.

Tweepy: A Python library that simplifies accessing the Twitter API for collecting tweets and user information.

1. What is the confidence level in predicting the sentiments ?

Answer: The model's confidence level in predicting sentiments is moderately high, given the accuracy and AUC of around 77%.

1. Which model provides the best prediction ?

Answer: Overall we found that Logistic Regression is the best model for anylyzing Sentiments on the dataset.

Logistic Regression is following the principle of Occam’s Razor which defines that for a particular problem statement if the data has no assumption, then the simplest model works the best. Since our dataset does not have any assumptions and Logistic Regression is a simple model, therefore the concept holds true for the above-mentioned dataset.

1. What ethical considerations were taken care while analyzing the data ?

Answer:When analyzing data, especially sensitive data like Twitter sentiments, several ethical considerations should be taken into account to ensure responsible and respectful handling of information.

1. What are the future improvements that you can specify for this analysis?

Answer: I would like to improve more on data cleansing and analysis .

1. Can this model be able to predict sentiment of any product which is tweeted?

Answer: While a well-trained sentiment analysis model can predict the sentiment of tweets about various products, its effectiveness depends on the quality and diversity of the training data, the sophistication of the model, and the handling of domain-specific language and context. Continuous monitoring and updating of the model are essential to maintain its accuracy and relevance.