ML Project Report

Group members:

- Shayan Mustafa (cs161041)
- Ahmed Kumail Pirzada (cs161048)
- Ehtasham Ali (cs161008)
- Fiza Ansari (cs161009)

Introduction:

We aim to predict the topic of tweets out of a labelled dataset with 12 categories (excluding RJ).

Dataset and Preprocessing:

Tweets are obtained from individuals at DHA Suffa University enrolled in Machine Learning course.

Overall tweets: 6471
Tweets left after cleaning: 2602

The following steps are applied to dataset to remove invalid and rejected data:

1. Dropping id, source and created_at columns as they are not necessary
2. Fill NaN tweets from wrong “   tags” column into proper “tags” column and drop “   tags” column
3. Drop NaN values
4. Drop tags which include categories “RJ”, “Rj”, “ET”, “EH”, “RH”

Following are the steps used for pre-processing:

1. Data is split into train_x, valid_x, train_y and valid_y using train_test_split from sklearn.
2. LabelEncoder is used to encode the categories
3. CountVectorizer is applied and fitted on text
4. Data is sent to preprocess function which removes STOP WORDS and SnowballStemmer and WordNetLemmatizer is applied to process text

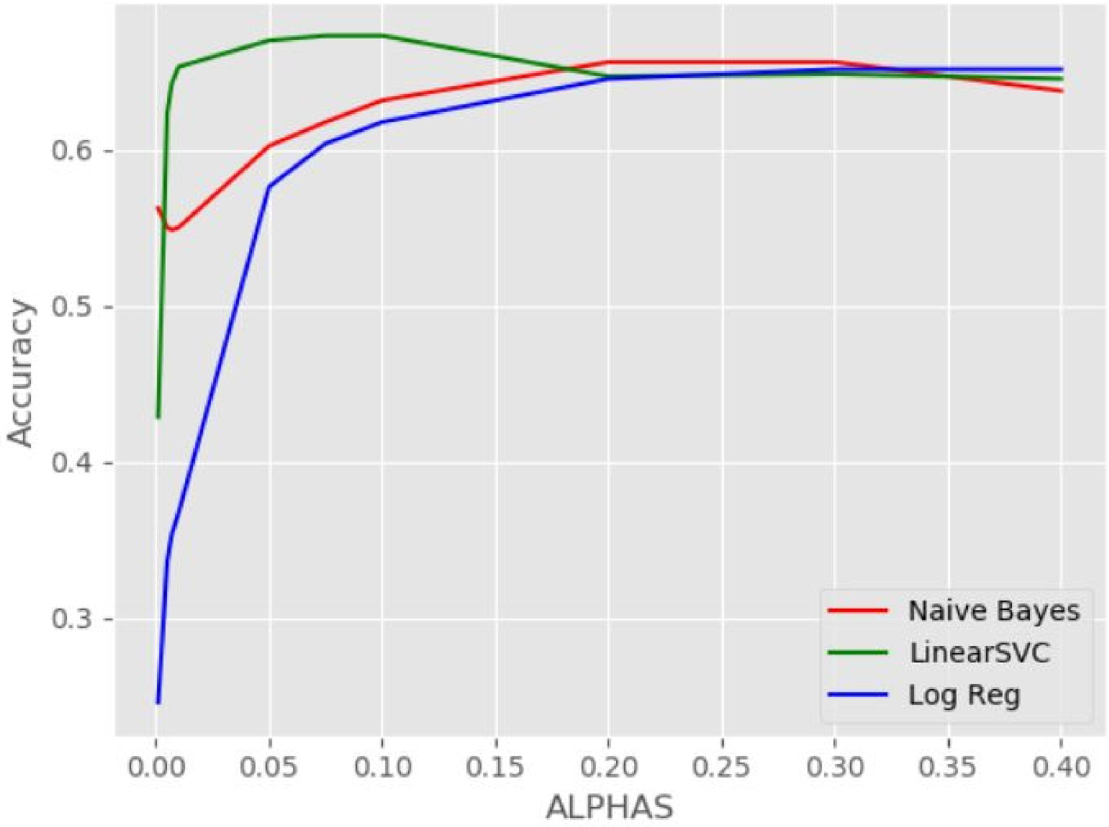
Model Selection:

As Naive Bayes, Linear SVM and Logistic Regression are all good choices to look for when it comes to multi-class text classification, we are going to try all three of them.

Naive Bayes gives us an accuracy of **63.1%** with alpha=0.1.
Linear SVM gives us an accuracy of **67.2%** with C=0.1.
Logistic Regression gives us an accuracy of **65.7%** with C=0.1 and lbfgs solver.
Therefore, we can say that Linear SVM is the best model for our case.

Next, we will run different values of alpha and C to check which gives us a better accuracy.

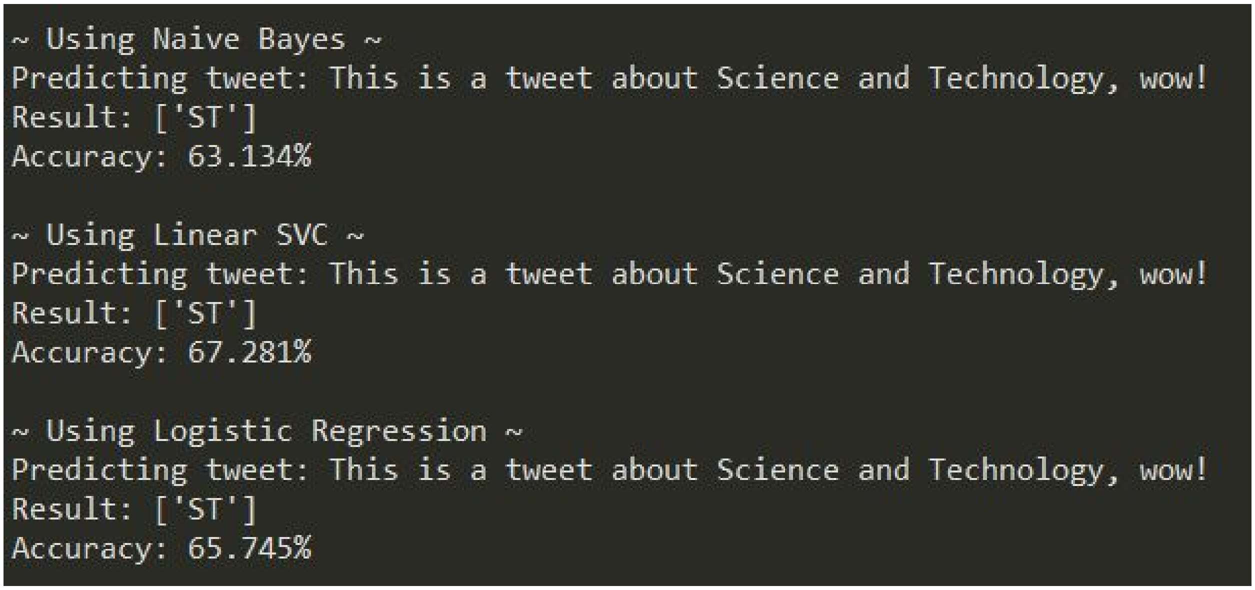
Using values from 0.001 to 0.4, we get the following results:



As we can see, Linear SVC provides us with the best result between C = 0.05 and C = 0.10.

Prediction:

If we predict a custom tweet, we can identify it using all three models accurately.

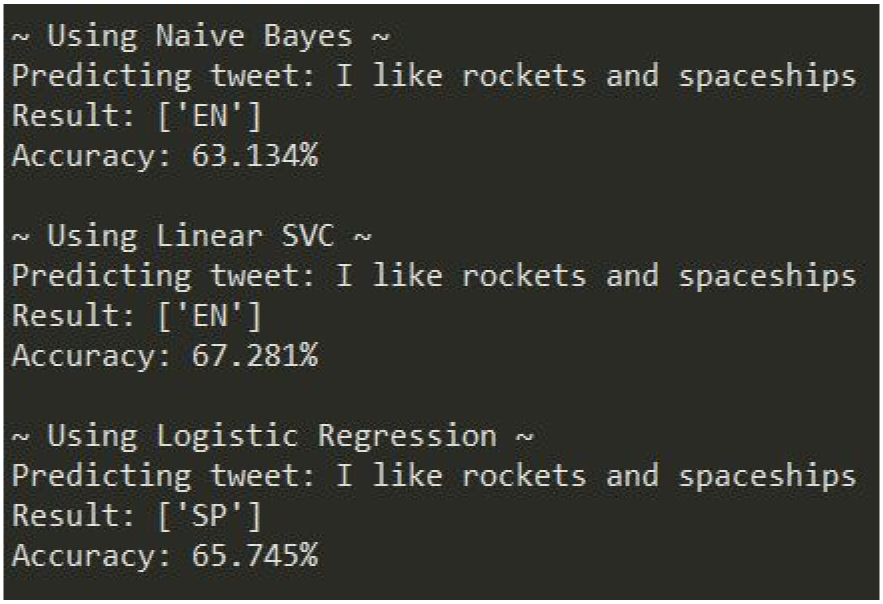


```
~ Using Naive Bayes ~
Predicting tweet: This is a tweet about Science and Technology, wow!
Result: ['ST']
Accuracy: 63.134%

~ Using Linear SVC ~
Predicting tweet: This is a tweet about Science and Technology, wow!
Result: ['ST']
Accuracy: 67.281%

~ Using Logistic Regression ~
Predicting tweet: This is a tweet about Science and Technology, wow!
Result: ['ST']
Accuracy: 65.745%
```

But if we run our models on a different tweet which is harder to predict, all three of them fail to provide the correct result, which should be **Science and Technology (ST)**.



```
~ Using Naive Bayes ~
Predicting tweet: I like rockets and spaceships
Result: ['EN']
Accuracy: 63.134%

~ Using Linear SVC ~
Predicting tweet: I like rockets and spaceships
Result: ['EN']
Accuracy: 67.281%

~ Using Logistic Regression ~
Predicting tweet: I like rockets and spaceships
Result: ['SP']
Accuracy: 65.745%
```

Conclusion:

In conclusion, we can say that we need a bigger dataset to get more accuracy. With around 2600 tweets this is the best accuracy we can obtain. Although, if the dataset is huge it will also give us a better fit. Therefore, the current models have a higher bias and they not fit for use in practical scenario.