Project 3: Reddit Classification

Naveen Parthasarathy - 4/25/2024

Table of Contents

Problem Statement

The objective of this project is to correctly predict which among the following subreddits a post is from -

1. r/LiverpoolFC - Liverpool Football Club (Best Club)
2. r/reddevils - Manchester United Football Club (Banter Club)

Data Sourcing

The data sourced was from the PRAW reddit API that allowed post titles, self texts, timestamps and subreddit information to be pulled.

Additionally, the data is from the ‘new’ category of posts.

Methodology - 1. Preprocessing

a. First a Lemmatizer was used on the title and self text of posts from both subreddits.
b. Then character and word counts of the titles and self texts of both subreddits were found.
c. Finally, both subreddit datasets were combined.

Longest and shortest titles by characters

```
longest_shortest(liverpool, 'title_count')
✓ 0.0s
```

	title	title_count
739	[Ornstein and Pearce]What happened with Xabi A...	300
933	[Premier League] Howard Webb explains the foll...	300
947	[Berger] Johan Bakayoko (20/🇧🇪) is a summer to...	298
168	Schweinsteiger on why Manchester United haven'...	297
915	[Martin Volkmar via Romano] Bayern's Uli Hoene...	297

	title	title_count
139	Fulham line up.	15
432	Well said boss🙌	15
198	We can do this.	15
443	PL Watch Thread	15
477	Sky predictions	15

Longest and shortest self_text by character

```
# Getting title character counts
longest_shortest_self(liverpool, 'title_count')
✓ 0.0s
```

	self_text	title_count
739		300
933		300
947	https://x.com/berger_pj/status/176977693092705...	298
168	Good to see how positively Schweinsteiger rega...	297
915		297

	self_text	title_count
139		15
432		15
198		15
443	Sigh….\n\nI don't blame y'all if you don't wan...	15
477	Sky is predicting we beat united 6-1. Surely n...	15

Methodology - 1. Preprocessing

After splitting the data into train and test, a **TFIDF vectorizer** was fit to the train data, then transformed both it and the test data.

Similarly, a **standard scaler** was fit on the train data and subsequently used to transform both train and test.

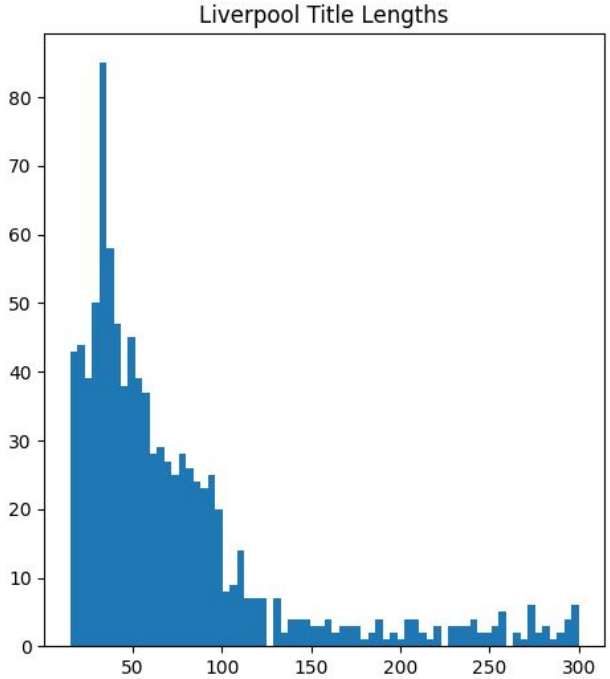
The resultant shape of the datasets are -

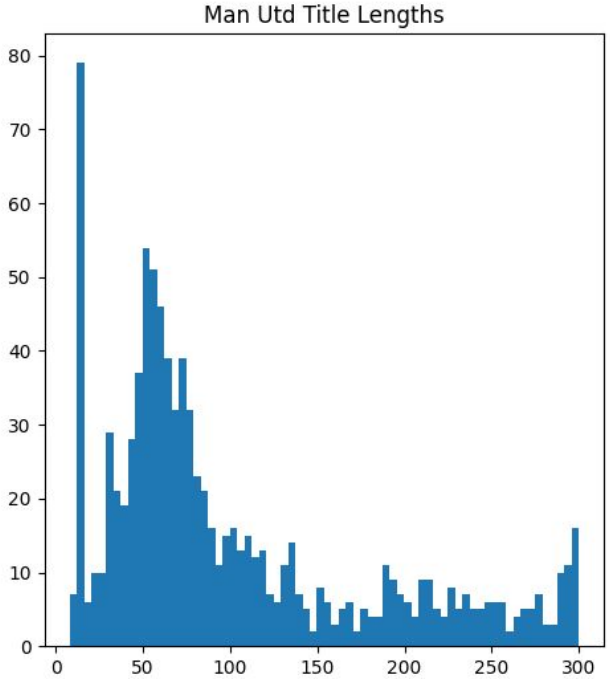
```
# Checking shape
X_train_new.shape, X_test_new.shape
```

✓ 0.0s

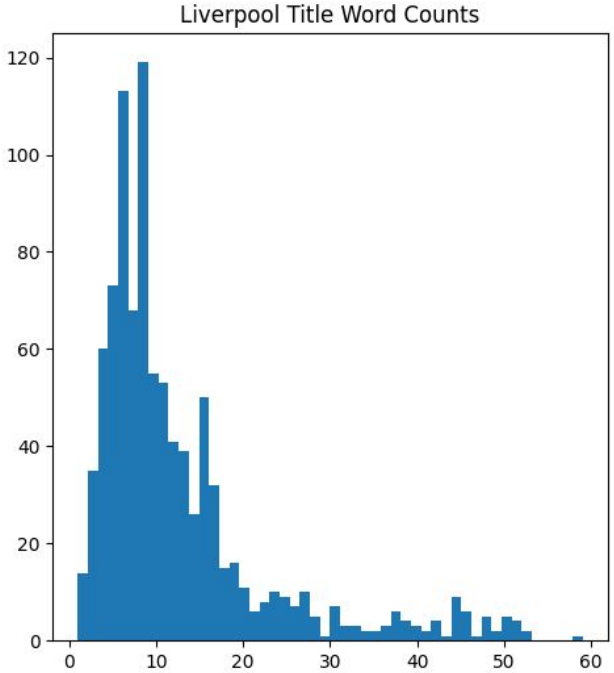
```
((1342, 8342), (576, 8342))
```

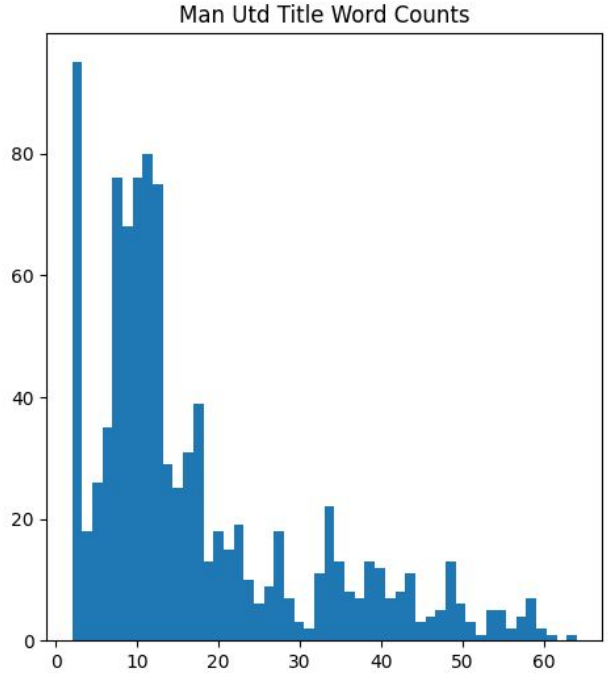
Methodology - 2. EDA





Methodology - 2. EDA





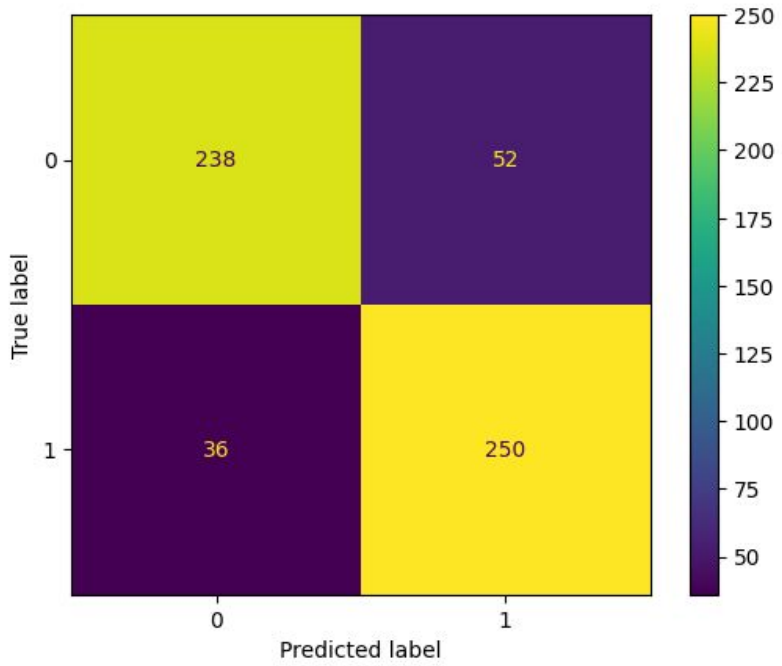
Methodology - 3. Models

Model 1 - Logistic Regression

Train Accuracy - 99.9%

Test Accuracy - 84.7%

Confusion Matrix



1- reddevils, 0 - LiverpoolFC

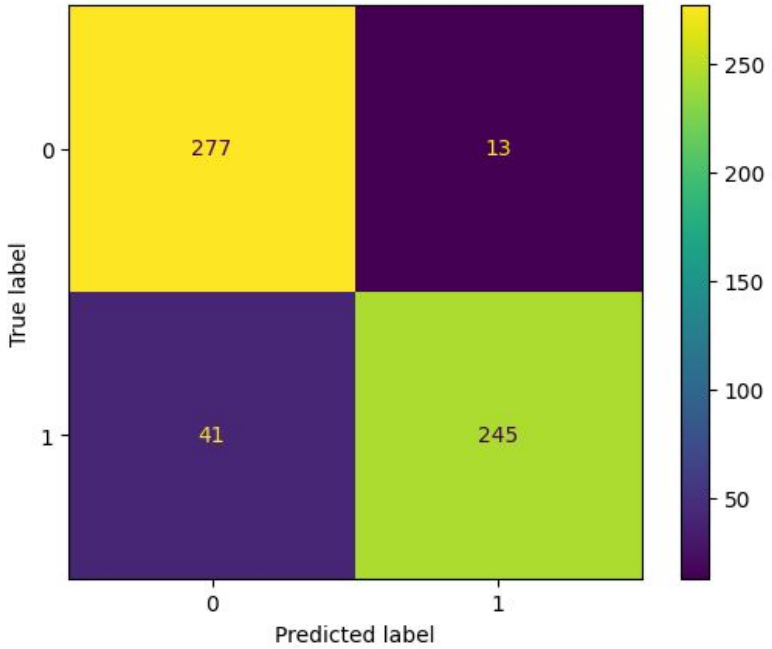
Methodology - 3. Models

Model 1 - Random Forest

Train Accuracy - 99.9%

Test Accuracy - 90.6%

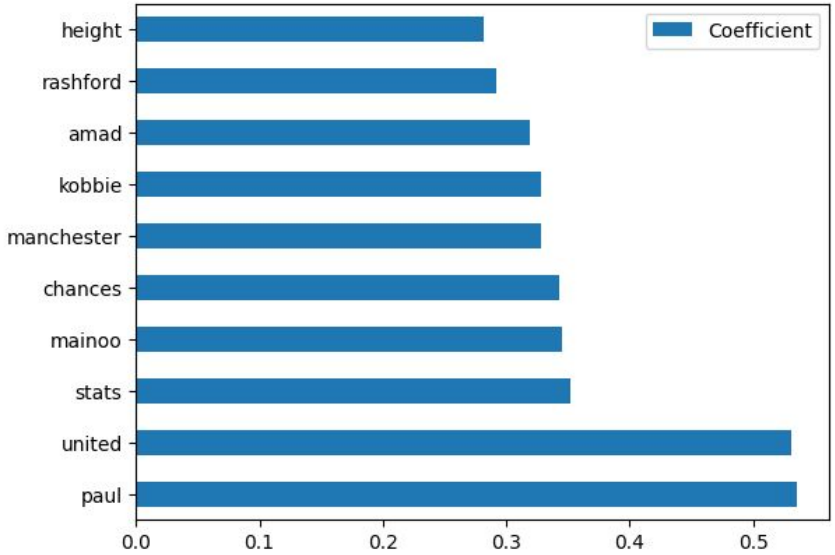
Confusion Matrix



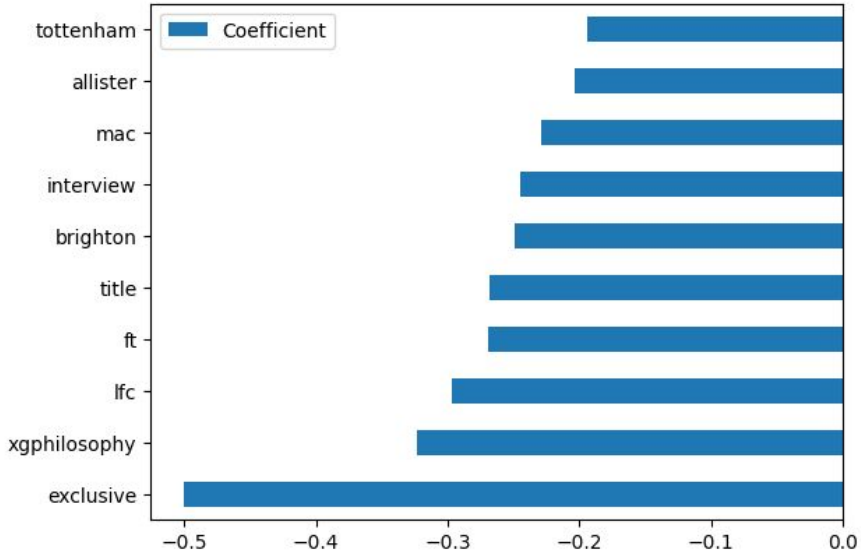
1- reddevils, 0 - LiverpoolFC

Inferences

Logistic Regression Model



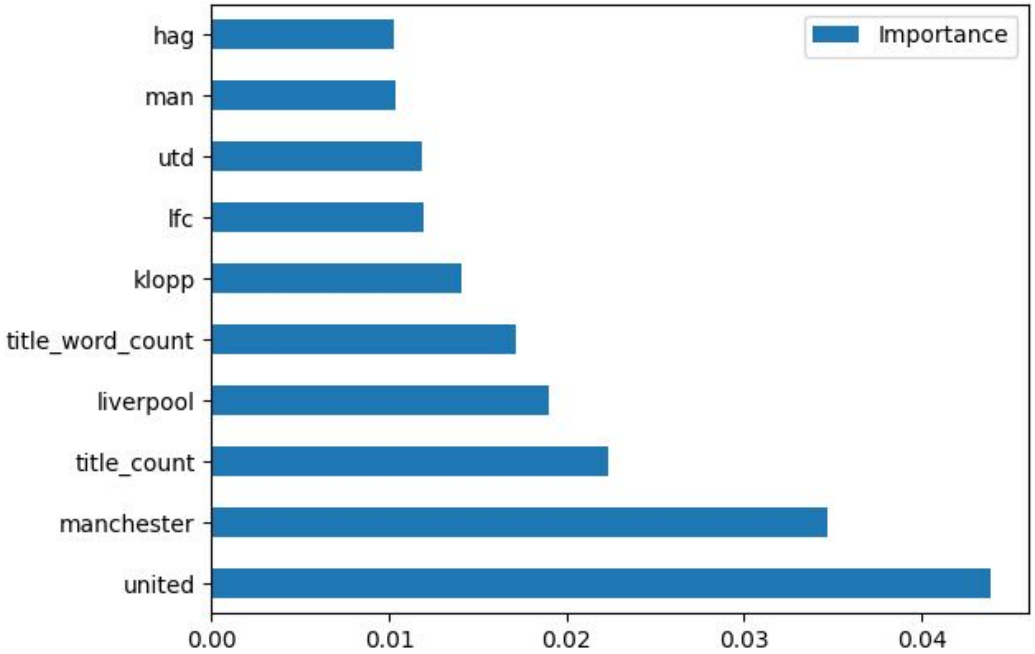
Positively correlated - Predictor of Man U



Negatively correlated - Predictor of Liverpool

Inferences

Random Forest Model



Conclusion

Both models overfit to the training data. This could be due to high variance, since the number of features in each model is very high.

Despite this, the testing accuracy of both models were high, with the Random Forest model performing best with a 90% accuracy on test data.

Future Work

Incorporating a variety of other models.

Exploring other vectorizers.

Implementing cross validation.

Using other features available from reddit API.


FIRMINO
9
VS
0 FULL TIME 5
Premier League


sky sports
DARWIN
27
7-0
Full-Time

Thank you