Project 3: Web APIs & NLP

Malcolm Lau • 22.01.2022

<Total Slides:20>

Background

- Common for inappropriate content such as illegal/undesirable activities to be posted online on Reddit before those posts can be removed.

- Reddit is quite hands-off with moderating subreddits, leaving it completely to the moderators to do so

Problem Statement

WHO ARE WE

Tech Consultants engaged by moderators of the r/Football subreddit

Our Task

To detect whether a post is related to soccer betting in the subreddit and flag it for removal for the benefit of users who are minors

Deliverable

To build a **binary classification model** that would classify whether a post belonged to r/Football or r/Soccerbetting.

The classification result will be used as a proxy to detect posts that contain betting/gambling content.

Scope

1. Methodology
    a. Data Cleaning
    b. EDA
    c. Modelling
    d. Selection
2. Results & Observations
3. Future Work

Methodology

Imputing missing values, removal of duplicate posts.

Creating benchmark model & candidate models for selection

Preprocess of text data, word frequency

Selection of best model

STEP 1: Data Cleaning | STEP 2: EDA | STEP 3: Modelling | STEP 4: Selection

	subreddit	selftext	title
0	football	How many other current footballers (age 19 to ...	Alex Ferguson said "Give me Zidane and 10 piec...
1	football	NaN	let me present the worst rating system in foot...
2	football	I want to inspire myself. I would prefer answe...	Who are some footballers with a great hard wor...
3	football	[removed]	Who are some players with the best mentalities...
4	football	[removed]	African Cup of Nation 2022 kicked off in Camer...

(4000, 3)

4000 posts obtained

1308 posts missing values

1230 posts removed/deleted

64 duplicate posts

STEP 1: Data Cleaning → STEP 2: EDA → STEP 3: Modelling → STEP 4: Selection

	subreddit	selftext	title
0	football	How many other current footballers (age 19 to ...	Alex Ferguson said "Give me Zidane and 10 piec...
1	football	NaN	let me present the worst rating system in foot...
2	football	I want to inspire myself. I would prefer answe...	Who are some footballers with a great hard wor...
3	football	[removed]	Who are some players with the best mentalities...
4	football	[removed]	African Cup of Nation 2022 kicked off in Camer...

```
(4000, 3)
```

4000 posts obtained

1308 posts missing values

1230 posts removed/deleted

64 duplicate posts

→

Missing values filled with content from 'title'

Removed/Deleted posts dropped

Duplicate posts dropped

STEP 1: Data Cleaning → STEP 2: EDA → STEP 3: Modelling → STEP 4: Selection

	subreddit	selftext	title
0	football	How many other current footballers (age 19 to ...	Alex Ferguson said "Give me Zidane and 10 piec...
1	football	NaN	let me present the worst rating system in foot...
2	football	I want to inspire myself. I would prefer answe...	Who are some footballers with a great hard wor...
3	football	[removed]	Who are some players with the best mentalities...
4	football	[removed]	African Cup of Nation 2022 kicked off in Camer...

(4000, 3)

4000 posts obtained

1308 posts missing values

1230 posts removed/deleted

64 duplicate posts

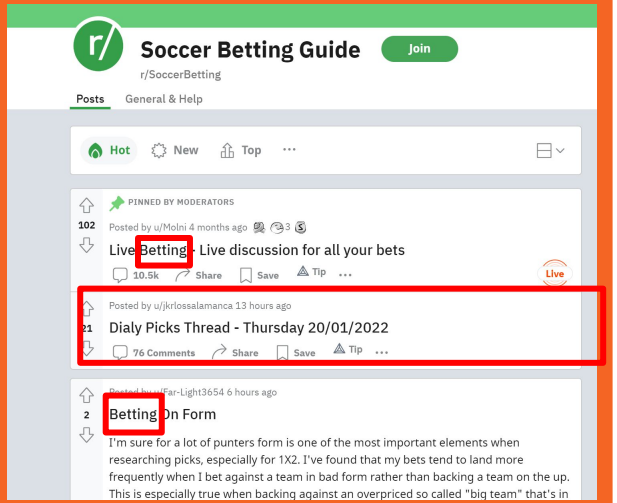
DataFrame Size After Cleaning

(2706, 3)

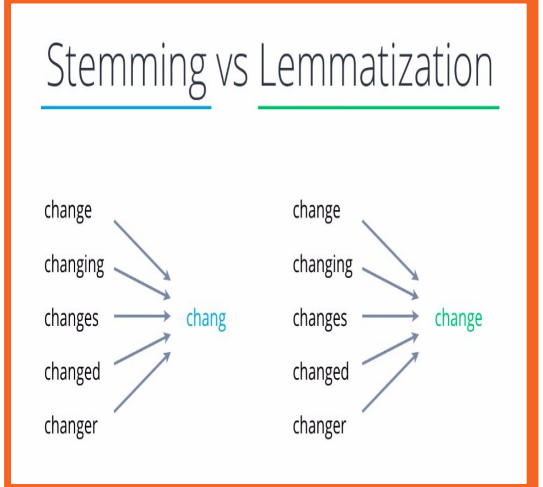
STEP 1: Data Cleaning | **STEP 2: EDA** | STEP 3: Modelling | STEP 4: Selection

Removing non-text characters

Removing Unhelpful Words

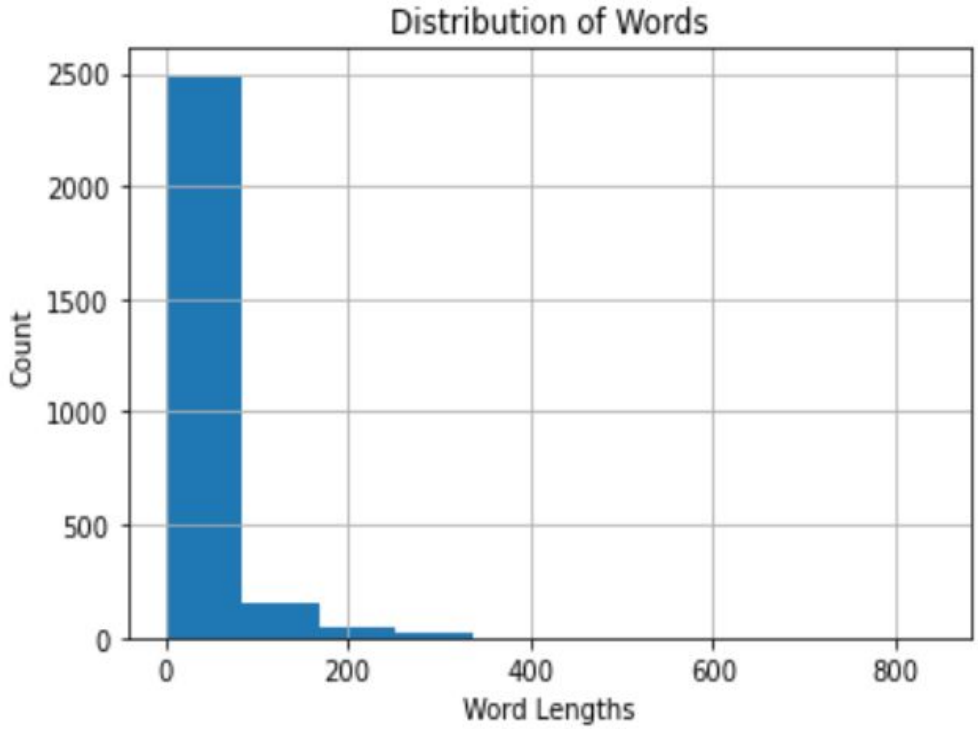


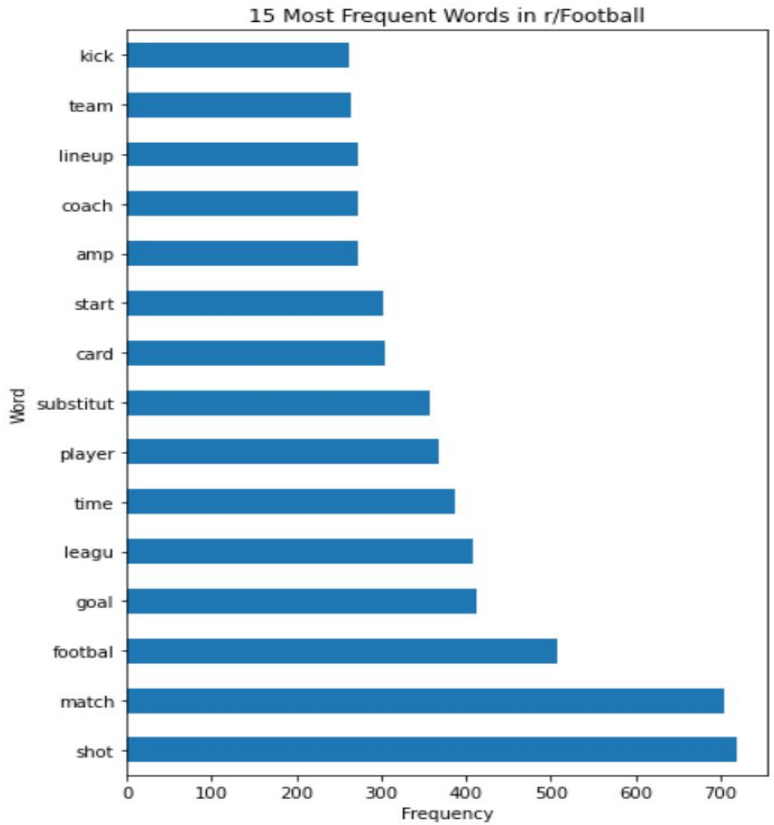
Lemmatization, Stemming

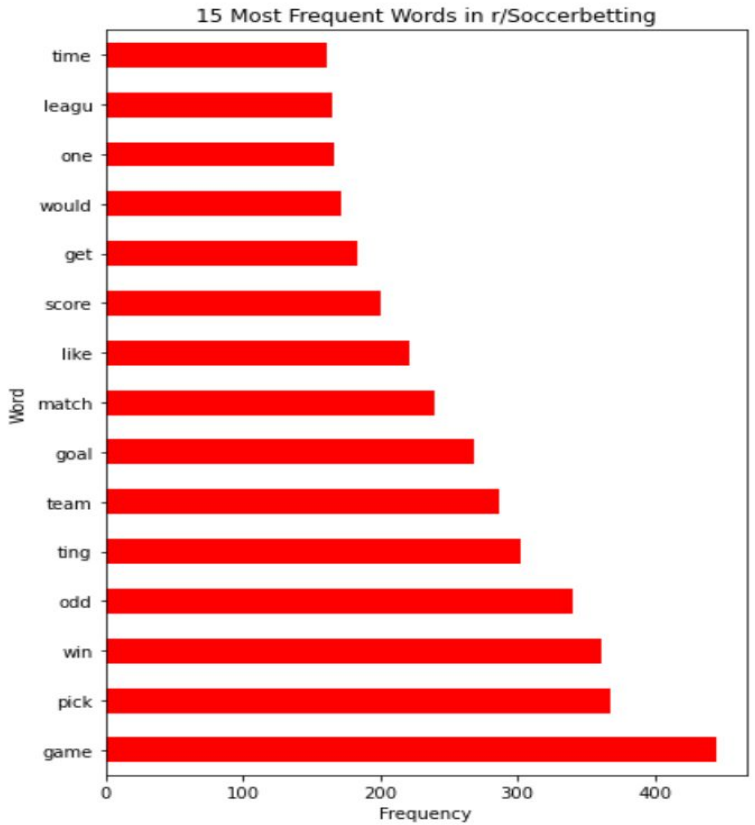


count	2706.000000
mean	27.090909
std	49.774499
min	1.000000
25%	5.000000
50%	11.000000
75%	27.000000
max	839.000000

1	0.56578
0	0.43422




15 Most Frequent Words in r/Football
kick
team
lineup
coach
amp
start
card
substitut
player
time
leagu
goal
footbal
match
shot
Word
0
100
200
300
400
500
600
700
Frequency


15 Most Frequent Words in r/Soccerbetting
time
leagu
one
would
get
score
like
match
goal
team
ting
odd
win
pick
game
Word
0
100
200
300
400
Frequency

{ 2706, 3 }

STEP 1: Data Cleaning | **STEP 2: EDA** | STEP 3: Modelling | STEP 4: Selection

2706, 3

Clean Data

4000 posts obtained

1308 posts missing values

1230 posts removed/deleted

64 duplicate posts

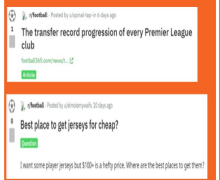
Missing values filled with content from 'title'

Removed/Deleted posts dropped

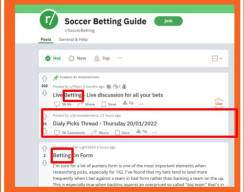
Duplicate posts dropped

Preprocess

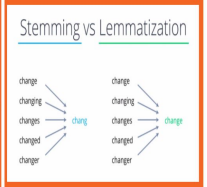
Removing non-text characters



Removing Unhelpful Words



Lemmatization, Stemming



STEP 1: Data Cleaning → **STEP 2: EDA** → STEP 3: Modelling → STEP 4: Selection

2706, 3

Clean Data

4000 posts obtained

1308 posts missing values

1230 posts removed/deleted

64 duplicate posts

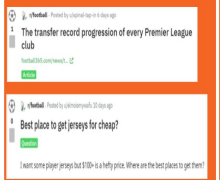
→ Missing values filled with content from 'title'

→ Removed/Deleted posts dropped

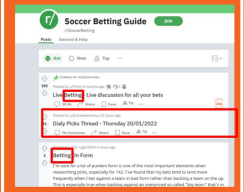
→ Duplicate posts dropped

Preprocess

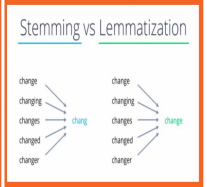
Removing non-text characters



Removing Unhelpful Words



Lemmatization, Stemming



→ 2706, 8336

Model	Text Vectorization
Multinomial Naive Bayes	**TF-IDF Vectorization**
K Nearest Neighbors	TF-IDF Vectorization
Random Forest	TF-IDF Vectorization
Logistic Regression	Count Vectorization

Performance Metrics to optimize for

r/Football

r/Soccerbetting

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Performance Metrics to optimize for

"0" r/Football

"1" r/Soccerbetting

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

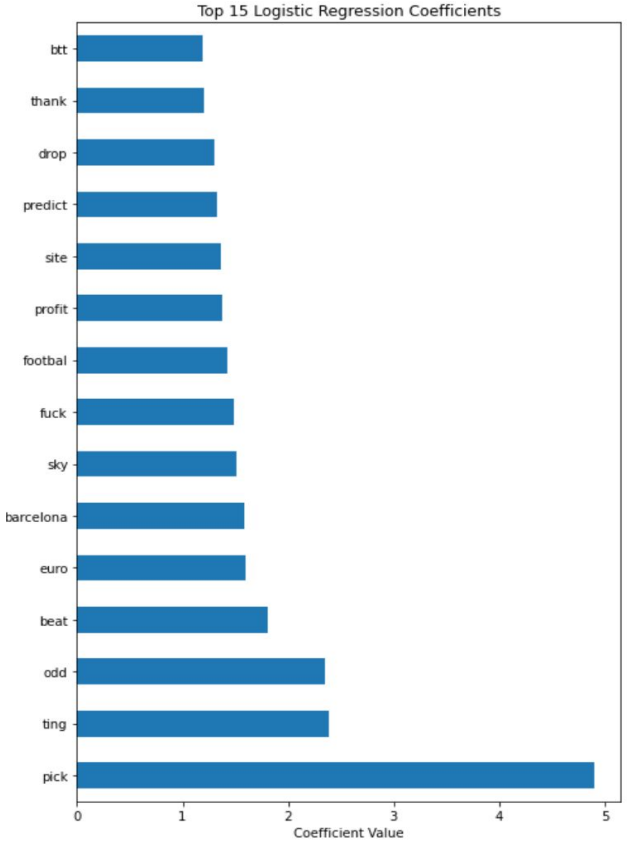
STEP 1: Data Cleaning → STEP 2: EDA → STEP 3: Modelling → STEP 4: Selection

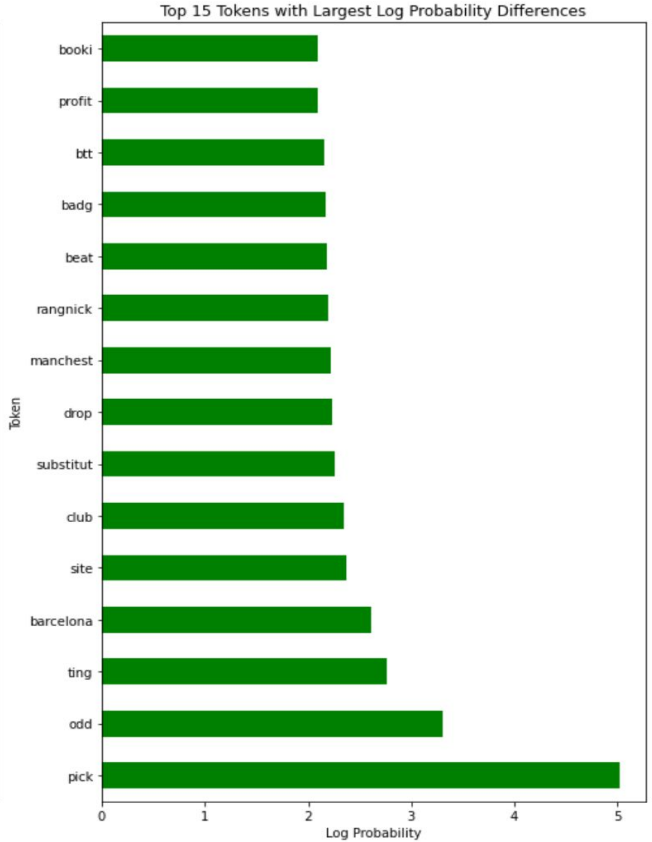
Model	Text Vectorization	Train Score	Test Score	Specificity	Recall	Precision	F1	AUC
Multinomial Naive Bayes	**TF-IDF Vectorization**	**0.972**	**0.913**	**0.867**	**0.948**	**0.875**	**0.925**	**0.966**
K Nearest Neighbors	TF-IDF Vectorization	0.916	0.897	0.823	0.953	0.875	0.913	0.952
Random Forest	TF-IDF Vectorization	1	0.897	0.806	0.966	0.867	0.914	0.957
Logistic Regression	Count Vectorization	0.989	0.894	0.816	0.953	0.871	0.910	0.957

STEP 1: Data Cleaning → STEP 2: EDA → **STEP 3: Modelling** → **STEP 4: Selection**

Model	Text Vectorization	Parameters
Multinomial Naive Bayes	**TF-IDF Vectorization**	{'tvec__max_features': 8000, 'tvec__min_df': 1, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': None} Multinomial NB Parameters: Default
K Nearest Neighbors	TF-IDF Vectorization	{'knn__metric': 'euclidean', 'knn__n_neighbors': 15, 'knn__p': 'uniform', 'tvec__max_features': 8000, 'tvec__min_df': 1, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': 'english'}
Random Forest	TF-IDF Vectorization	{'rf__max_depth': None, 'rf__max_samples': None, 'rf__n_estimators': 100, 'tvec__max_features': 8000, 'tvec__min_df': 1, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': 'english'}
Logistic Regression	Count Vectorization	Default

STEP 1: Data Cleaning | STEP 2: EDA | STEP 3: Modelling | **STEP 4: Selection**





Observations

- 60% of the features are common, which is expected. This is why the performance metrics are very close.
- In fact, the top 3 contributors for both models are identical. We can infer that the tokens: **pick, ting** & **odd** are the most important contributors to the model.

Next steps

Increase Model Vocabulary

Data can be obtained from other subreddits which centre on gambling in general such as **r/Gambling**, this will increase the model's ability to generalize when exposed to more generic words associated with gambling.

Improve Data Quality

One of the likely reasons for misclassification is typo errors from the user. Rectifying spelling mistakes will enable improve lemmatization and stemming performance, which will likely contribute to improvements in model performance as well.

Possible solution: Use Word module from **Textblob Library**

End of Presentation