SportTopic Categorizer

The filter that will help categorize posts about football and basketball that have come in abundance in webpage

Problem Statement

In an online community platform, users often engage in discussions across a wide range of topics. However, with an increasing volume of content, it becomes challenging for users to navigate and find posts that align with their specific interests. To address this issue, we aim to develop a text classification model that can accurately categorize posts into one of two categories: "Basketball" and "Football".

Questions

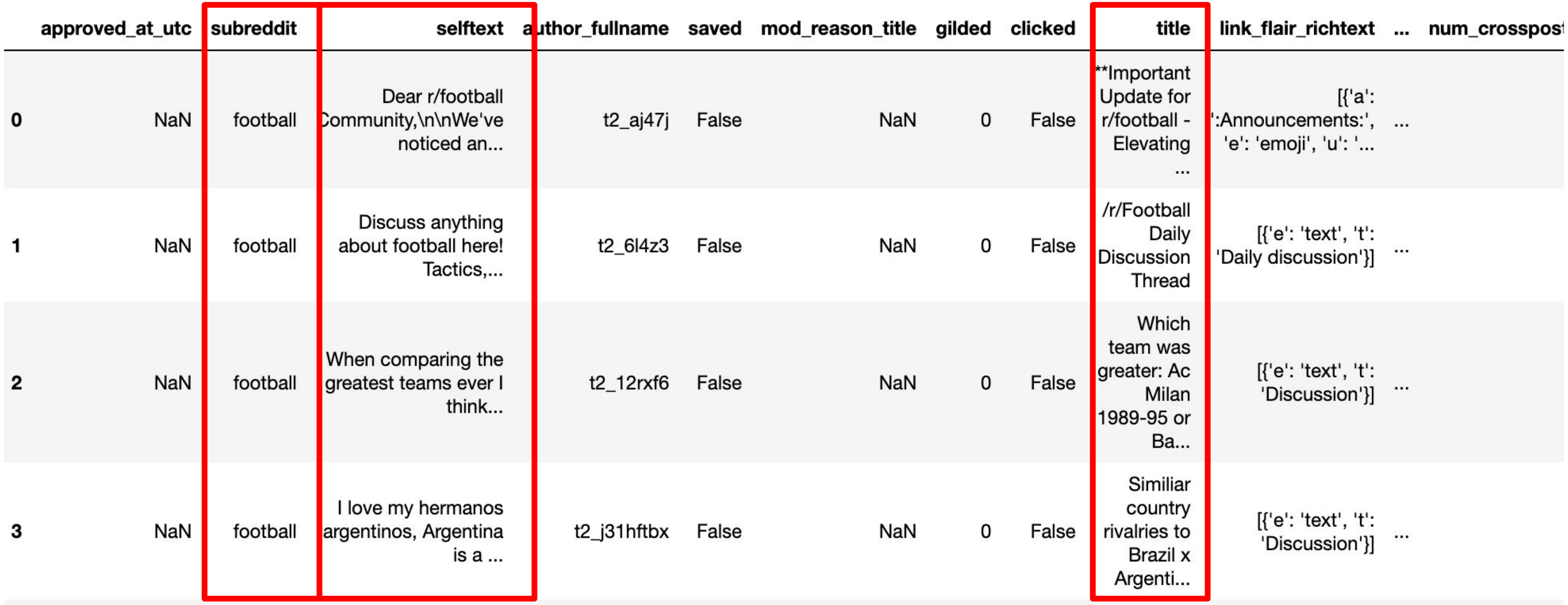
1. How can we effectively differentiate between posts related to basketball and football?
2. What features of the text data are most influential in making accurate classifications?
3. How can we ensure the model performs reliably on new unseen posts?

Exploratory Data Analysis and Cleaning Data

Dataset

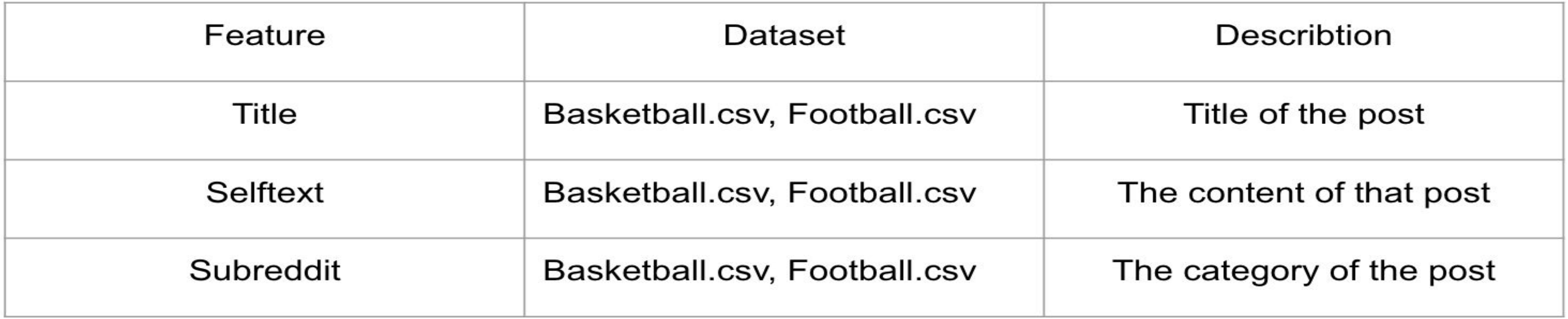
1.Basketball.csv (7441 rows , 114 columns)

2. Football.csv = (7431 rows , 114 columns)



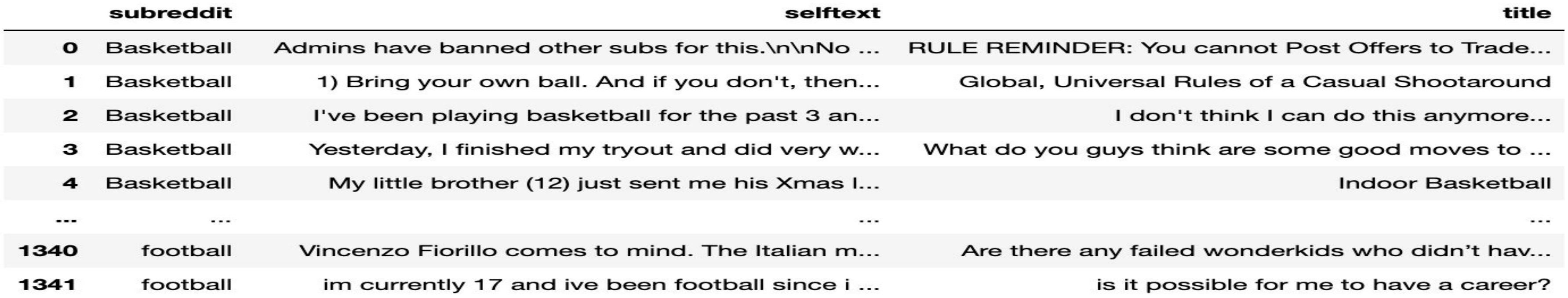
	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	link_flair_richtext	...	num_crosspost
0	NaN	football	Dear r/football Community,\n\nWe've noticed an...	t2_aj47j	False	NaN	0	False	**Important Update for r/football - Elevating ...	[{'a': ':Announcements:', 'e': 'emoji', 'u': '...	...	
1	NaN	football	Discuss anything about football here! Tactics,...	t2_6l4z3	False	NaN	0	False	/r/Football Daily Discussion Thread	[{'e': 'text', 't': 'Daily discussion'}]	...	
2	NaN	football	When comparing the greatest teams ever I think...	t2_12rxf6	False	NaN	0	False	Which team was greater: Ac Milan 1989-95 or Ba...	[{'e': 'text', 't': 'Discussion'}]	...	
3	NaN	football	I love my hermanos argentinos, Argentina is a ...	t2_j31hftbx	False	NaN	0	False	Similiar country rivalries to Brazil x Argenti...	[{'e': 'text', 't': 'Discussion'}]	...	

Exploratory Data Analysis and Cleaning Data



Feature	Dataset	Describtion
Title	Basketball.csv, Football.csv	Title of the post
Selftext	Basketball.csv, Football.csv	The content of that post
Subreddit	Basketball.csv, Football.csv	The category of the post

Merging the data from "basketball.csv" and "football.csv" into one. and Drop duplicate , Drop Null values

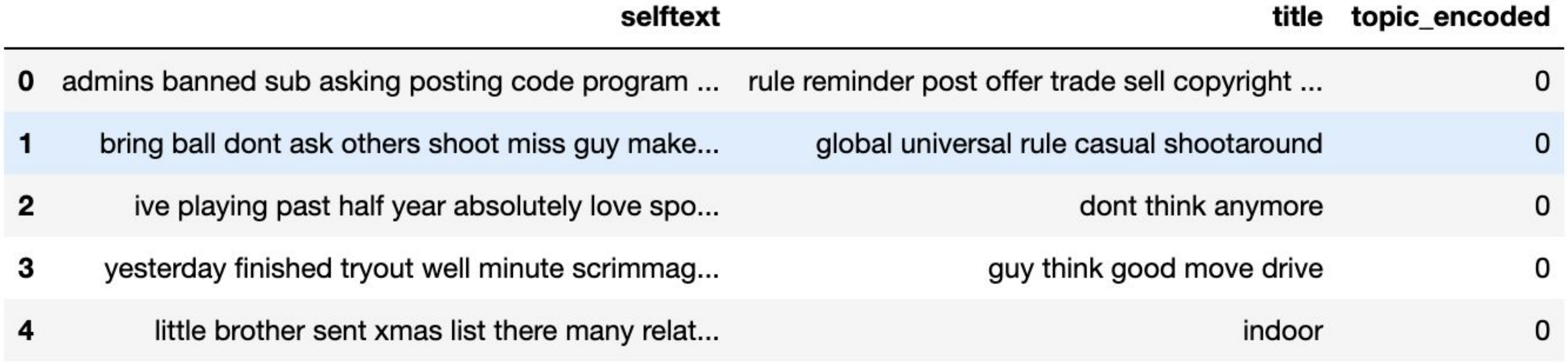


```
      subreddit                                            selftext                                              title
0    Basketball  Admins have banned other subs for this.\n\nNo ...  RULE REMINDER: You cannot Post Offers to Trade...
1    Basketball  1) Bring your own ball. And if you don't, then...     Global, Universal Rules of a Casual Shootaround
2    Basketball  I've been playing basketball for the past 3 an...              I don't think I can do this anymore...
3    Basketball  Yesterday, I finished my tryout and did very w...  What do you guys think are some good moves to ...
4    Basketball  My little brother (12) just sent me his Xmas l...                                  Indoor Basketball
...         ...                                                ...                                                ...
1340   football  Vincenzo Fiorillo comes to mind. The Italian m...  Are there any failed wonderkids who didn't hav...
1341   football  im currently 17 and ive been football since i ...               is it possible for me to have a career?
```

Exploratory Data Analysis and Cleaning Data

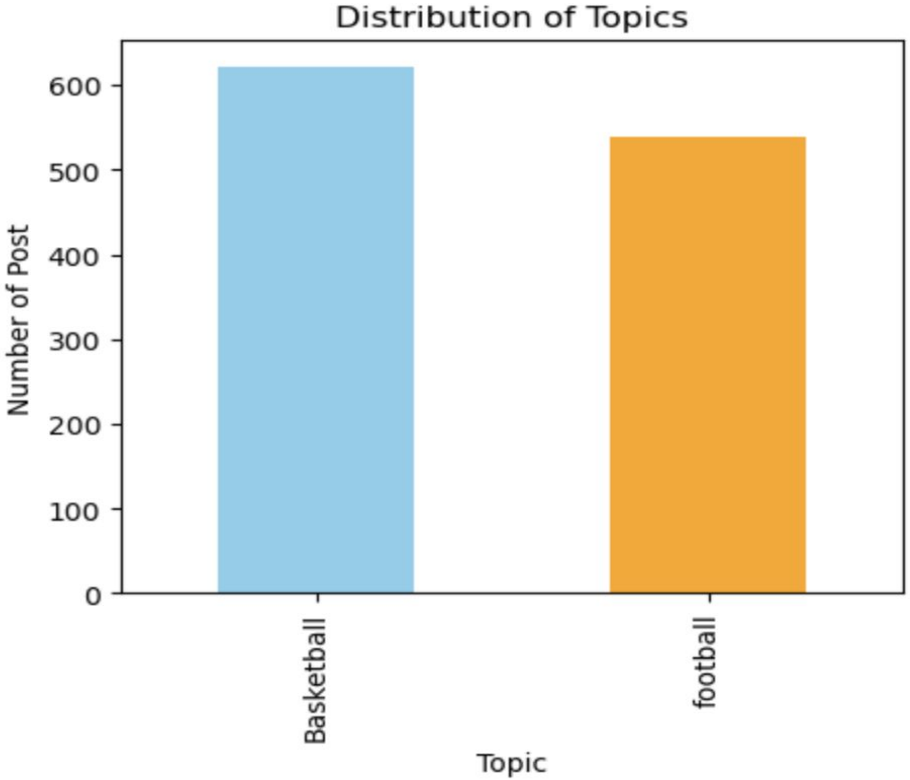
Define Function to clean data

- Use regular expression character that is not an alphabetical letter (lowercase or uppercase) or a whitespace character replace with withespace
- Remove words “football” , “basketball” from all of text
- Lemmatizer
- Change Subreddit to Topic_encoded 0 : Basketball , 1 : Football



	selftext	title	topic_encoded
0	admins banned sub asking posting code program ...	rule reminder post offer trade sell copyright ...	0
1	bring ball dont ask others shoot miss guy make...	global universal rule casual shootaround	0
2	ive playing past half year absolutely love spo...	dont think anymore	0
3	yesterday finished tryout well minute scrimmag...	guy think good move drive	0
4	little brother sent xmas list there many relat...	indoor	0

Exploratory Data Analysis and Cleaning Data



Basketball   621 posts

Football   538 posts

Preprocessing

In this experiment, I explore different feature combinations to identify the most effective approach for our text classification task

X = Selftext

X = Title

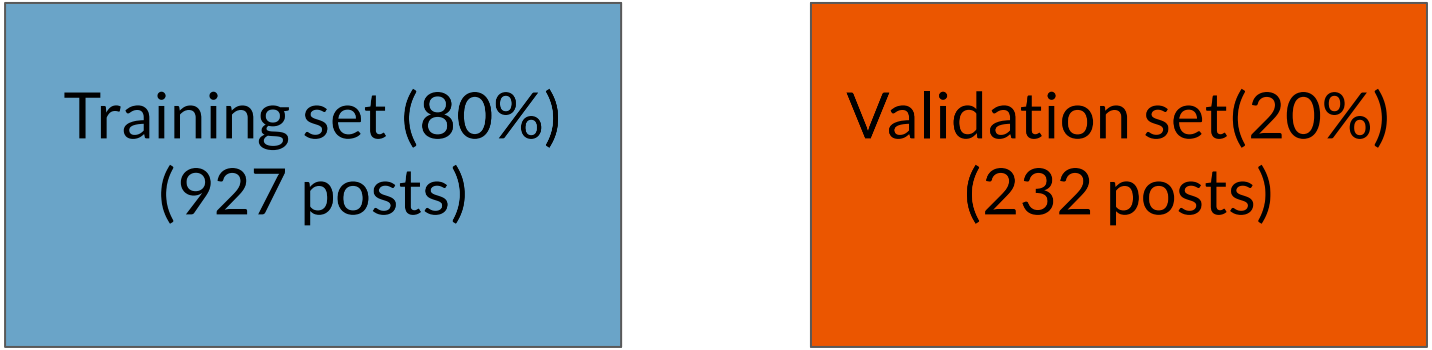
X = Title + Selftext

Y = Target

To understand which features contribute most to model performance.

Preprocessing

Data Split for Training and Validation



Training set (80%)
(927 posts)

Validation set(20%)
(232 posts)

Modeling & Evaluation

Evaluation Metric : F1 - score

Since this problem is important in terms of both Precision and Recall, especially to prevent misclassification in both "Basketball" and "Football" categories, the F1-score is a suitable metric

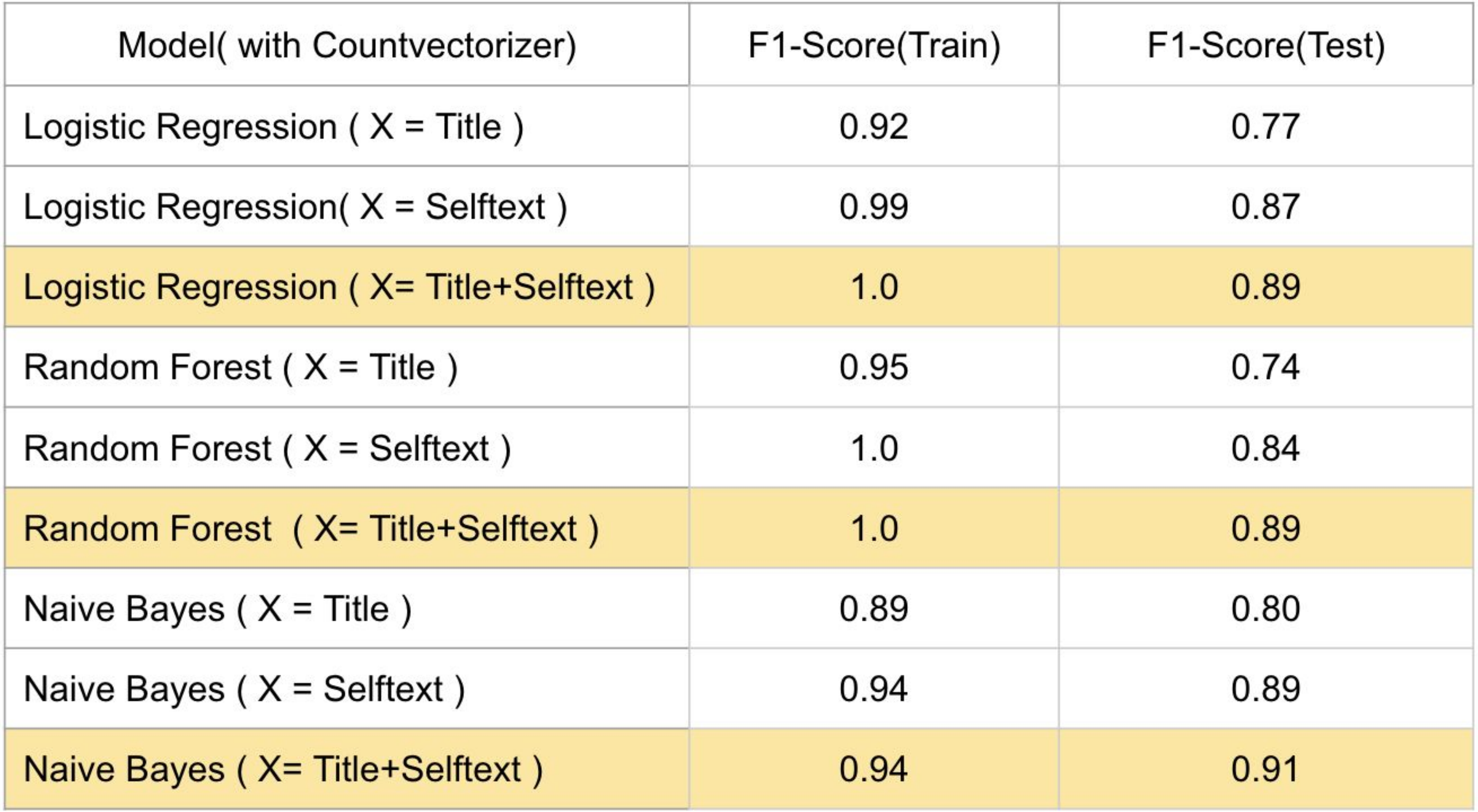
Explore and evaluate with 3 Models and 2 Vectorization techniques

1. Logistic Regression
2. Naive Bayes
3. Random Forest

Vectorization

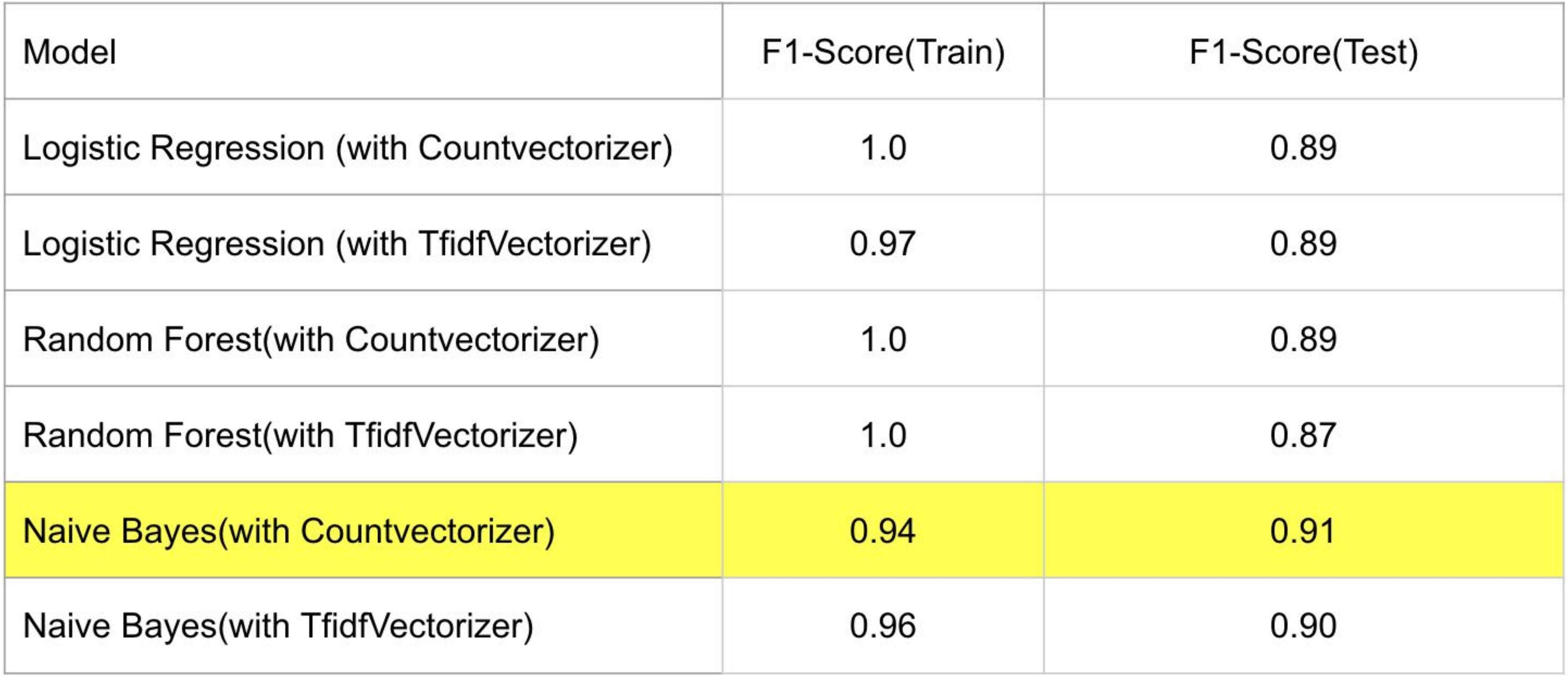
1. Countvectorizer
2. Tfidfvectorizer

Modeling & Evaluation



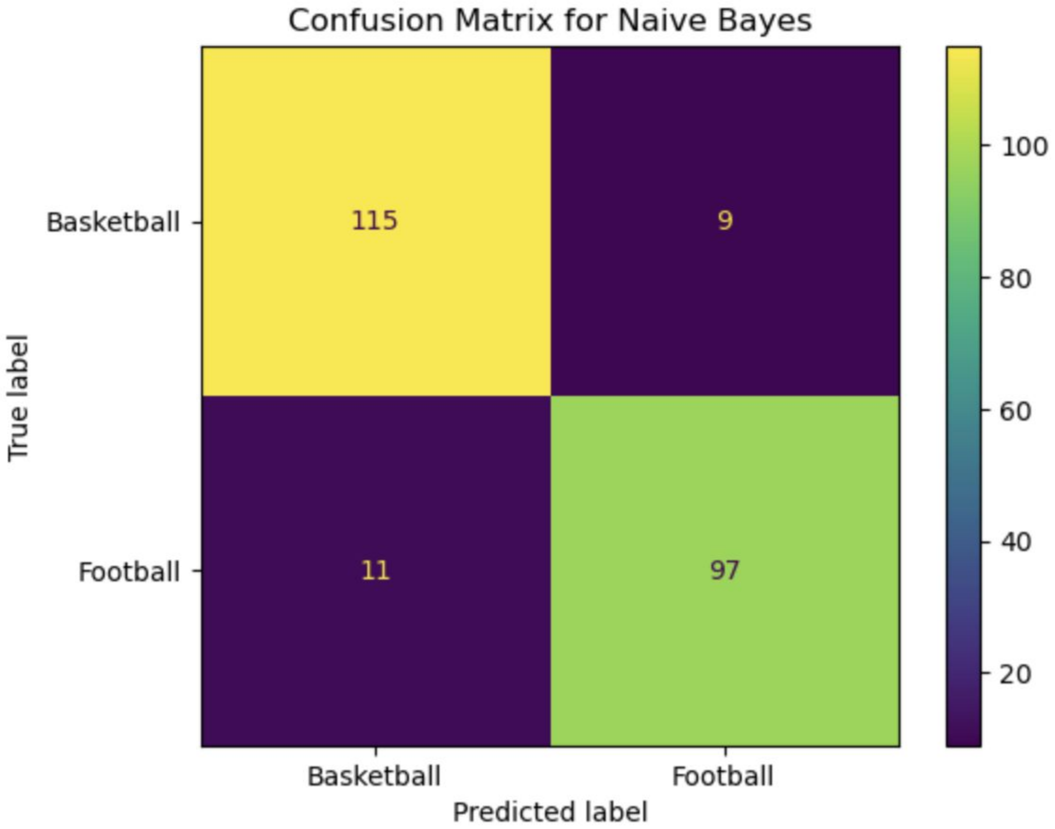
Model(with Countvectorizer)	F1-Score(Train)	F1-Score(Test)
Logistic Regression (X = Title)	0.92	0.77
Logistic Regression(X = Selftext)	0.99	0.87
Logistic Regression (X= Title+Selftext)	1.0	0.89
Random Forest (X = Title)	0.95	0.74
Random Forest (X = Selftext)	1.0	0.84
Random Forest (X= Title+Selftext)	1.0	0.89
Naive Bayes (X = Title)	0.89	0.80
Naive Bayes (X = Selftext)	0.94	0.89
Naive Bayes (X= Title+Selftext)	0.94	0.91

Modeling & Evaluation



Model	F1-Score(Train)	F1-Score(Test)
Logistic Regression (with Countvectorizer)	1.0	0.89
Logistic Regression (with TfidfVectorizer)	0.97	0.89
Random Forest(with Countvectorizer)	1.0	0.89
Random Forest(with TfidfVectorizer)	1.0	0.87
Naive Bayes(with Countvectorizer)	0.94	0.91
Naive Bayes(with TfidfVectorizer)	0.96	0.90

Error Analysis



0 : Basketball
1 : Football

Error Analysis

- True Positive (TP): The model correctly predicted posts related to football 97 posts
- True Negative (TN): The model correctly predicted posts related to basketball. 115 posts
- False Positive (FP): The model predicted posts as football but they aret basketball 9 post
- False Negative (FN): The model predicted posts as basketball but they are football 11 post

Predicted Basketball But Actual Football	Predicted Football but Actual Basketball
ee get picked ever year fantasy he always gambleguy think lonzo truly done	channel similar production quality depthbest channel similar thinking
would rather guy like casemiro guy like busquets team profile likepivot v destroyer type dm prefer	extra ticket duke season need sell wont attending game grad student dm infoselling duke season ticket
small struggle fitnesse shot powershotscasillas weakness	titletell track field coach main focus miss end season

Limitations

1. The limitation in collecting data from the subreddit is that I attempted to gather information from a total of 3,000, 7,000, and 14,000 posts, but it seems that after removing duplicated data, only around 600 posts remain.
2. The model may lack the ability to understand the broader context of a post, leading to misclassifications when posts contain URL or ambiguous language.

Question

1. How can we effectively differentiate between posts related to basketball and football?
2. What features of the text data are most influential in making accurate classifications?
3. How can we ensure the model performs reliably on new, unseen posts?

Conclusion and Recommendation

In evaluating the use of the "Title+Selftext" features, we found that it yielded the best results. The Naive Bayes model with CountVectorizer achieved the highest performance, with an F1-score of 0.94 on the training set and 0.91 on the test set.

The best hyperparameters for this model are

- cvec__max_df : 0.9
- cvec__max_features : 3000
- cvec__min_df: 2
- cvec__ngram_range : (1, 2)
- nb__alpha': 1
- nb__fit_prior : True

These results were obtained by testing the model on the validation set, where it achieved an F1-score of 0.91. In summary, the model demonstrates high efficiency in classifying data in both the training and Validation sets.

Conclusion and Recommendation

1. Use features 'Selftext' and 'Title' with the Naive Bayes model and CountVectorizer to achieve the highest F1-score for categorizing posts into the football and basketball categories.
2. Continuously monitor and update the model with new data to maintain its accuracy and relevance.
3. Explore the possibility of incorporating additional features or contextual information to further enhance classification accuracy.
4. Consider integrating user feedback mechanisms to iteratively improve the model based on community preferences

THANK YOU