Author Identification

Name - Sanket S Houde
Roll no - 18387
Department - Biological Sciences
Course - DSE301

Introduction

- Books have influenced me quite a lot, especially the short stories of these three authors.
- Anton Chekhov was a Russian playwright and one of the most famous short stories' writer of all time. He is considered an important author of the genre - 'Realism'.
- Isaac Asimov was an American writer and a genre defining at that. He has written a lot of seminal works in the genre of 'Science fiction'.
- Luis Borges was an Argentine short story writer, essayist, poet. His most famous short stories are compiled into this one book - 'Ficciones'. His works can be categorised as 'Magical Realism'.
- These three authors have works in distinct genres.

About the dataset

- Roughly 20,000 words' corpus was compiled for each of these authors taken from various resources across the internet. This was taken as the training dataset.
- Another 2000 words' dataset was created for testing the models from the same sources.
- The corpus was then divided into chunks of individual sentences for extracting features.
- The dataframe consisted of three columns - 'text' containing the sentence extracts, 'author' containing the author of the extract and a vestigial column 'id' (can be ignored).

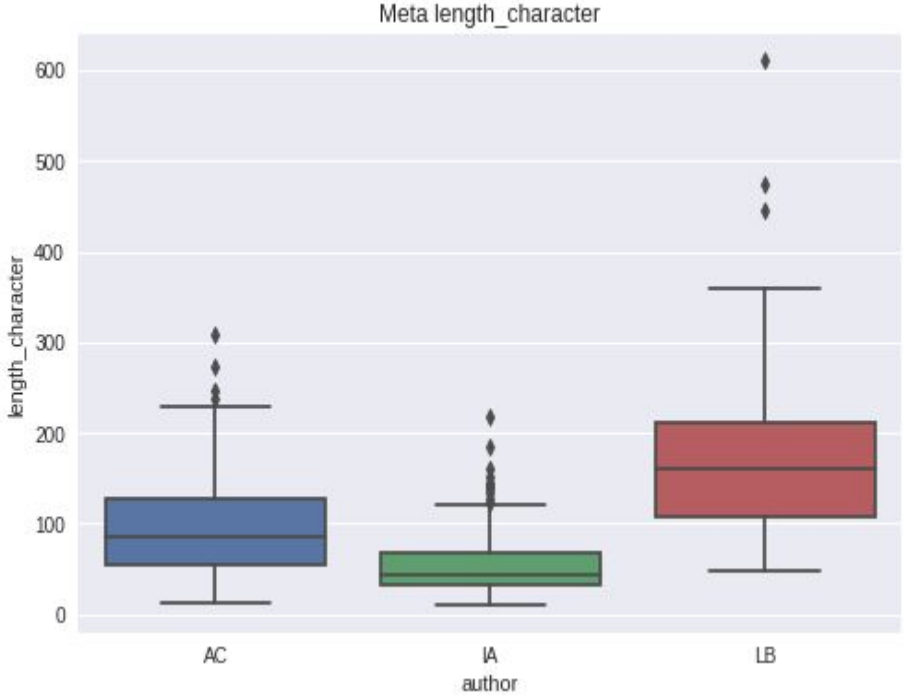
Dataframe sample

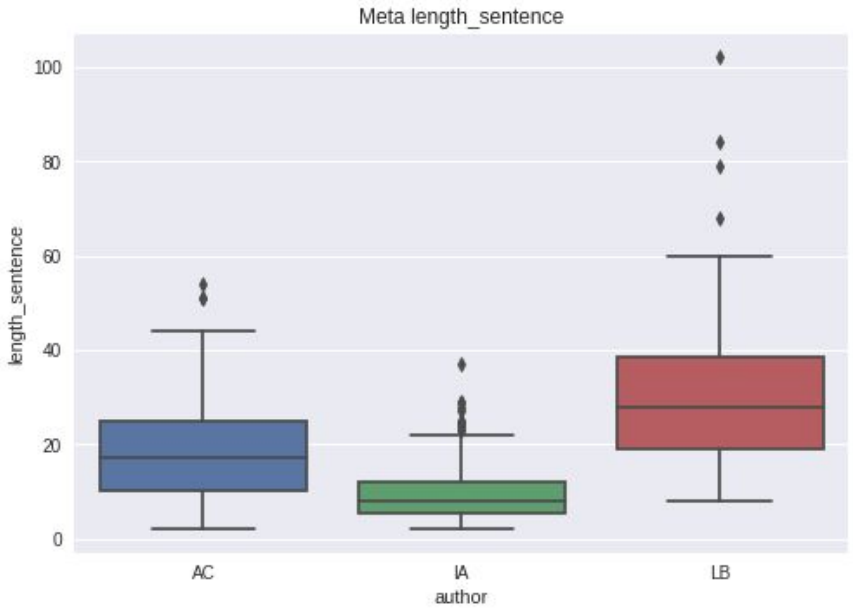
	text	author	id
0	"You are mad, and gone the wrong way.	AC	dummy
1	You would marvel if suddenly apple and orange ...	AC	dummy
2	So do I marvel at you, who have bartered heave...	AC	dummy
3	"That I may show you in deed my contempt for t...	AC	dummy
4	That I may deprive myself of my right to them,...	AC	dummy

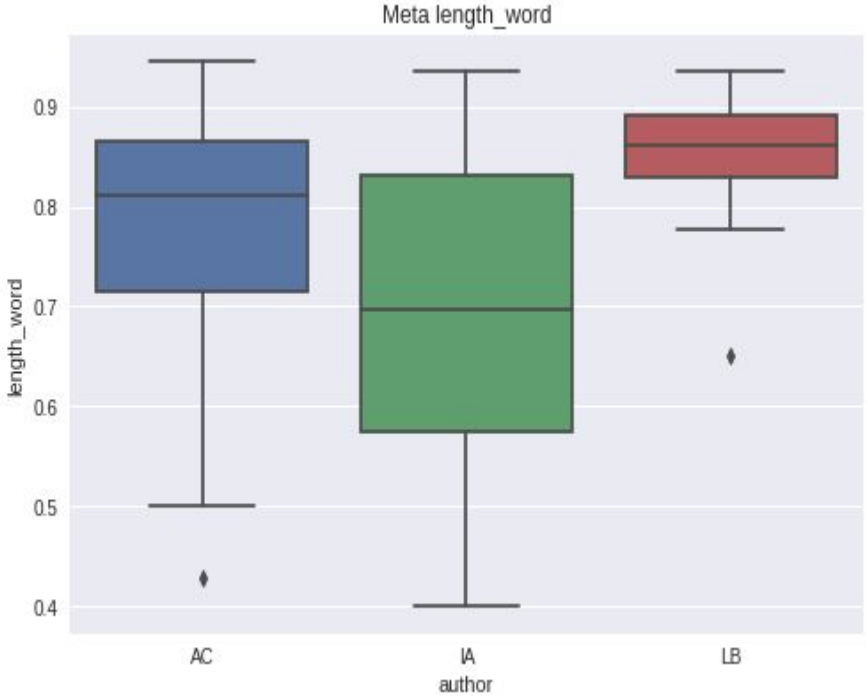
Features extracted from the dataset

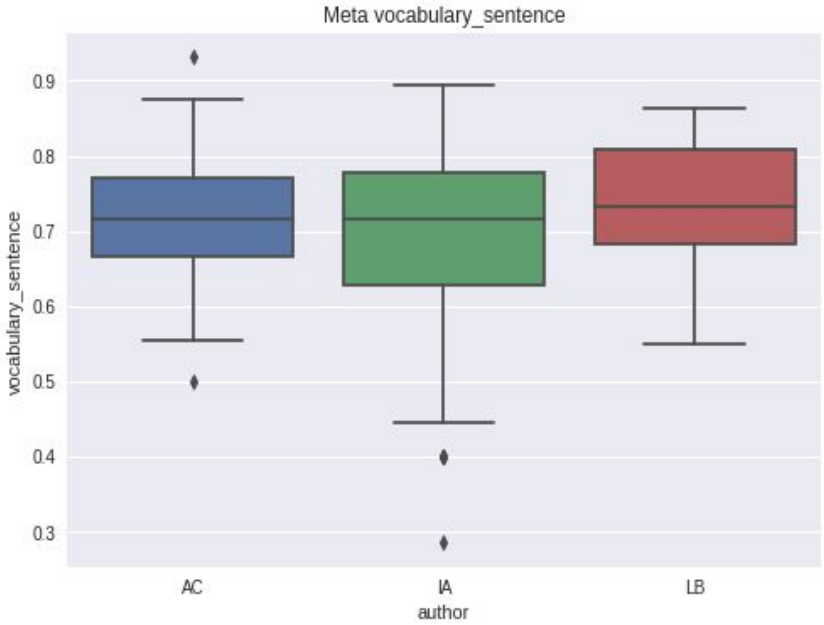
Meta Features	Text Features
1. Sentence length (characters & words) 2. Word length 3. Punctuation density 4. Percentage of unique words 5. Stopword count 6. Noun/adjective/verb density 7. Adjective to noun ratio 8. Emphases on words or phrases 9. Dialogue density 10. Feminine to masculine words ratio 11. Use of foreign languages	1. POS tag of first/last word of a sentence 2. Emotions (NRC data), positive/negative 3. TF-IDF (words n-grams): degree to which an author uses a word more than the two other authors 4. TF-IDF (characters n-grams) 5. TF-IDF (POS tags n-grams)

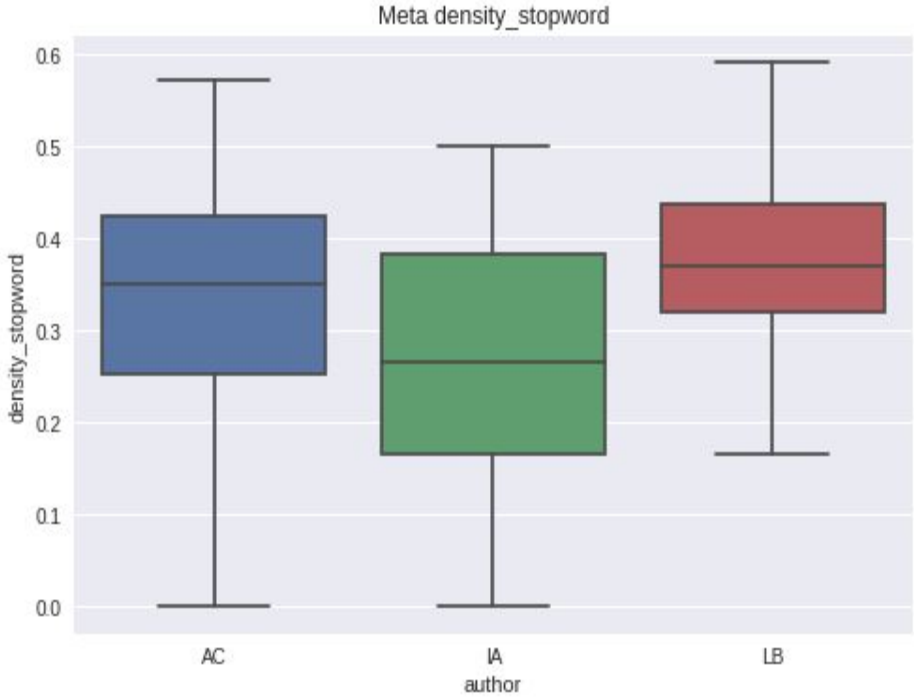
- A total of **1829** features were extracted from the training dataset.

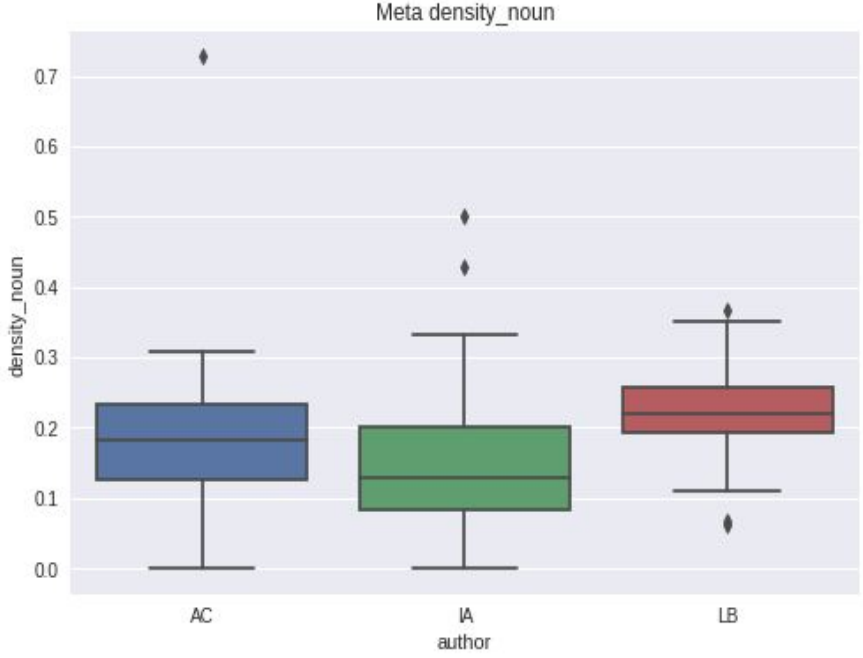

Meta length_character
length_character
0
100
200
300
400
500
600
AC
IA
LB
author

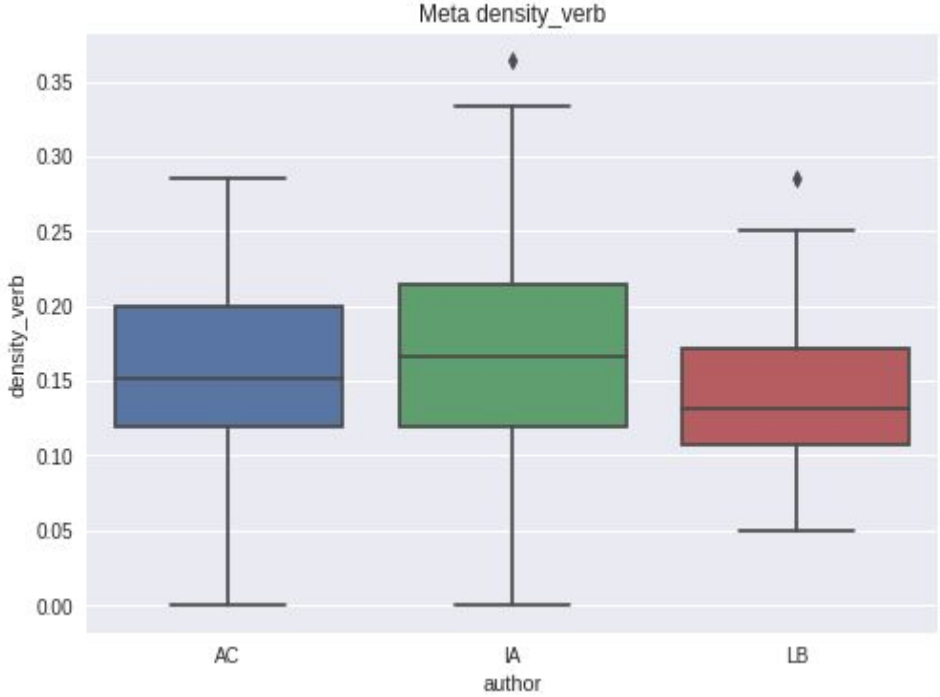

Meta length_sentence
length_sentence
0
20
40
60
80
100
AC
IA
LB
author

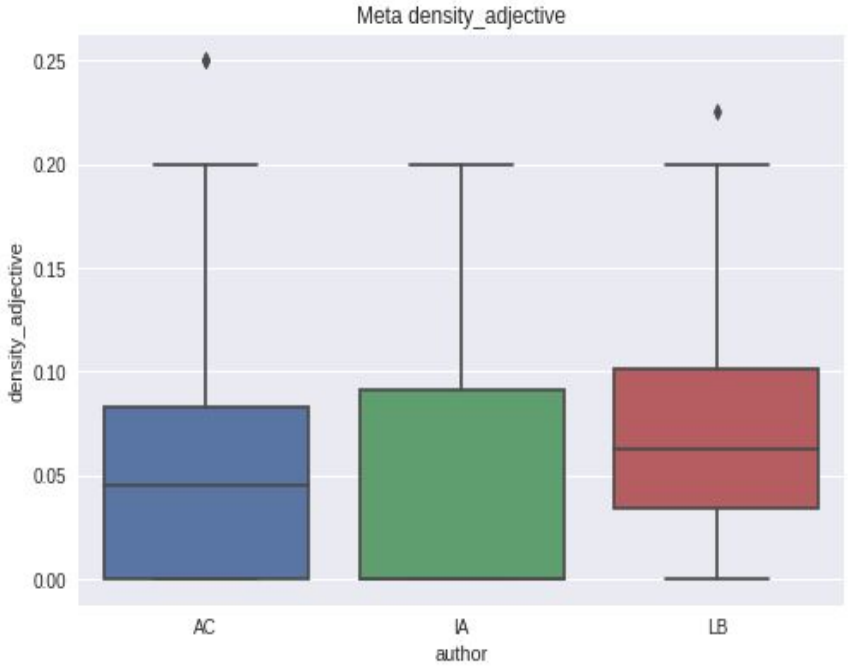

Meta length_word
length_word
0.9
0.8
0.7
0.6
0.5
0.4
AC
IA
LB
author

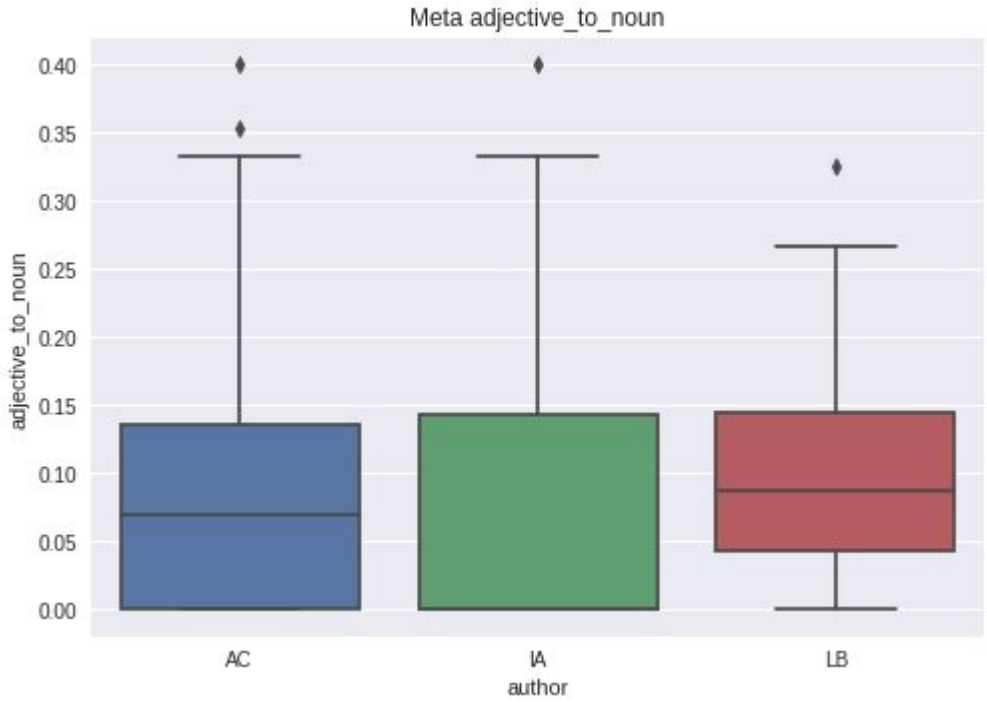

Meta vocabulary_sentence
vocabulary_sentence
0.9
0.8
0.7
0.6
0.5
0.4
0.3
AC
IA
LB
author


Meta density_stopword
density_stopword
0.0
0.1
0.2
0.3
0.4
0.5
0.6
AC
IA
LB
author

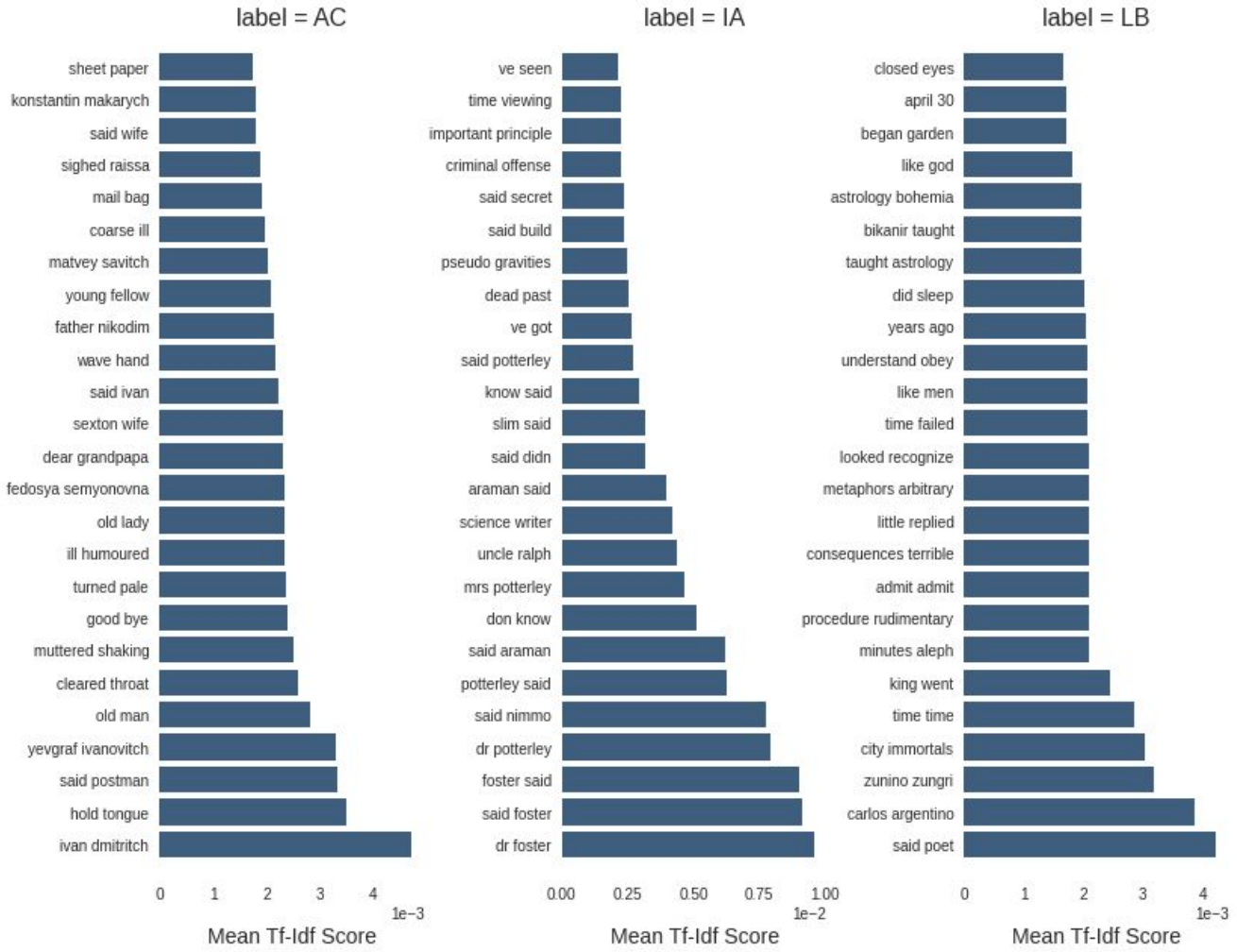

Meta density_noun
density_noun
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
AC
IA
LB
author


Meta density_verb
density_verb
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
AC
IA
LB
author


Meta density_adjective
density_adjective
0.25
0.20
0.15
0.10
0.05
0.00
AC
IA
LB
author


Meta adjective_to_noun
adjective_to_noun
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
AC
IA
LB
author

TF-IDF bigrams



Feature Selection methods

Two kinds of feature selection methods were used :-

- ➔ *Univariate Feature Selection* : Features were selected based on the ANOVA-F scores of the features.
- ➔ *Principal Component Analysis* ; It's a dimensionality reduction method which decomposes a large number of feature set to a smaller one which would still contain all the information from the larger set.

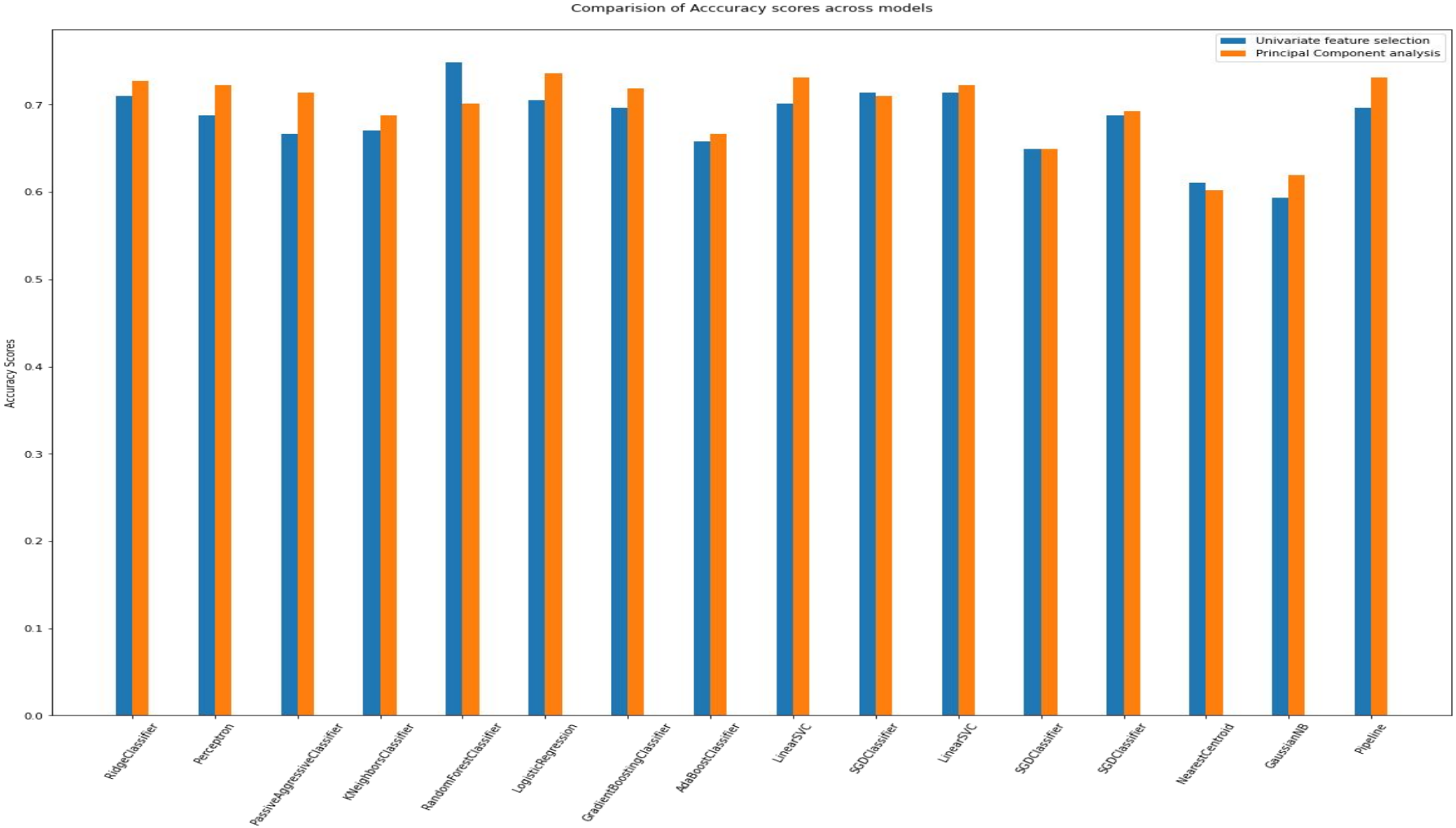
After tinkering with the number of desirable features, **500** was the number I struck upon.

Machine Learning Models

15 conventional Machine Learning models were used in combination with both of these feature selection methods, making the total number of models to be **30** in this project.

I. Gaussian Naive Bayes Classifier
II. Nearest Centroid Classifier
III. SGD Classifier (with ‘l1’ , ‘l2’ and ‘elasticnet’ penalties separately)
IV. Linear Support vector classifier (with ‘l1’ and ‘l2’ penalties separately)
V. AdaBoost Classifier
VI. Gradient Boosting Classifier
VII. Logistic Regression
VIII. Passive Aggressive Classifier
IX. Perceptron Classifier
X. Ridge Classifier
XI. Random Forest Classifier
XII. K-Nearest Neighbours Classifier

Results

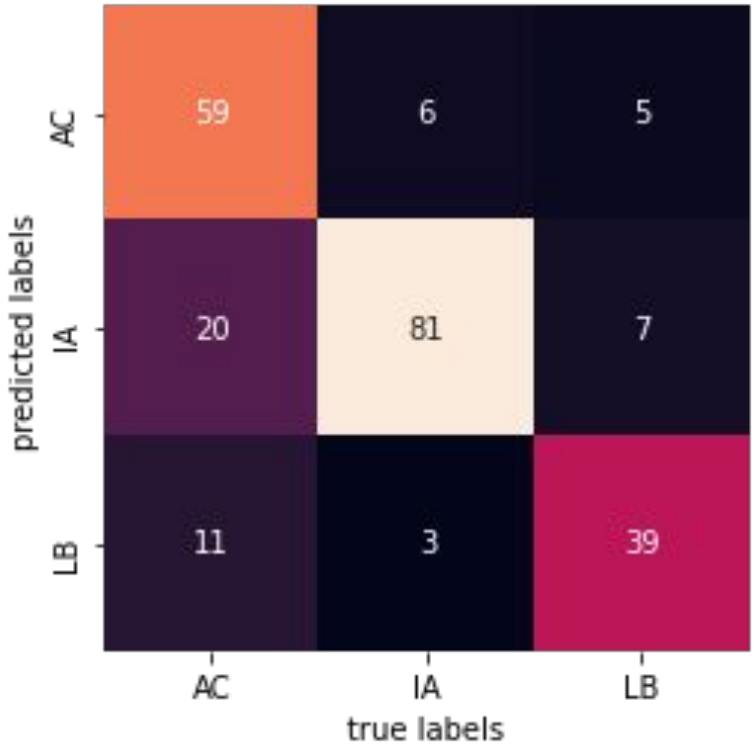

Comparision of Acccuracy scores across models
Univariate feature selection
Principal Component analysis
Accuracy Scores
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
RidgeClassifier
Perceptron
PassiveAggressiveClassifier
KNeighborsClassifier
RandomForestClassifier
LogisticRegression
GradientBoostingClassifier
AdaBoostClassifier
LinearSVC
SGDClassifier
LinearSVC
SGDClassifier
SGDClassifier
NearestCentroid
GaussianNB
Pipeline

- The most accurate model was the Random Forest Classifier with univariate feature selection . It had the highest accuracy score of **78.79%.**
- Most of the models showed *improved accuracies when Principal Component Analysis was used* for feature selection. This is mostly because, unlike univariate feature selection, PCA actually retains some information of all the extracted features instead of discarding features below some specified criterion.
- I tried using LSTM for training, however due to shortage of time for model optimization and execution the idea was scrapped. The model which was run had very low accuracy scores between **0.4-0.5,** even after multiple epochs.

Random Forest Classifier (Univariate Feature Selection) Metrics

Metrics :

- Accuracy Score : 78.79%
- F1 Score : 78.39%



Future Prospects/Improvements

- Advanced feature extraction methods like GloVe vectors, word2vec embeddings etc. can be used to improve model accuracies.
- LSTMs and Transformers are shown to provide excellent text classification models. However, the restraining factors here are computational power and time. This was the main reason that I had to abandon these options in this project.
- Increasing the dataset always helps, however the need for computational power increases.
- A alternative future prospect for this project would be to classify various works into genres.

THANK YOU !!