DSML - Assessment: 4 Project Report

Classification of Fiction and Nonfiction Genre Using Machine Learning

Piyush Paliwal, 17180

BS-MS 2017,

Department of Chemistry,

IISER Bhopal.

Pooja Singh, 2010214

PhD 2020,

Department of Chemistry,

IISER Bhopal.

Python Code:

Link of Google Colab Notebook Code:

Please click on above link for code in google colab. It is easier to run code directly on colab and see the result. Please create the copy of the code and the run it online to check the results used in analysis.

(Analysis Done By Piyush)

Introduction to Data set:

We have used Brown corpus data set for training machine learning models. Brown corpus data set contains 500 files information along with five features like no. of verb, pronoun, noun, adverb, adjectives present in each text file and its category (total 15 categories like: Adventure, belles' letters, editorial, fiction, hobbies, reviews, government, hobbies, humour, learned, lore, mystery, news, religion, romance, science fiction). Out of these categories:

Fiction category contains: fiction, mystery, romance, science fiction, adventure

Nonfiction category contains: News, hobbies, government, reviews, learned

Neither fiction nor nonfiction: lore, belles' letters, humour, religion, editorial.

We need to clean files form data set that do not belong to any of the category. After removing these, we are left with around 333 files whose data will be used for training and testing the ML models.

Analysis of Data set:

After classifying data into fiction (category = 1) and nonfiction (category = 0), analysis of additional features from given features (noun, pronoun, verb, adjectives, adverb) is done. These additional features are combinations of ratios primary features. It will help to figures out which features contributes most in solving classification problem. Train and test split is 40%.

Table of features analysed

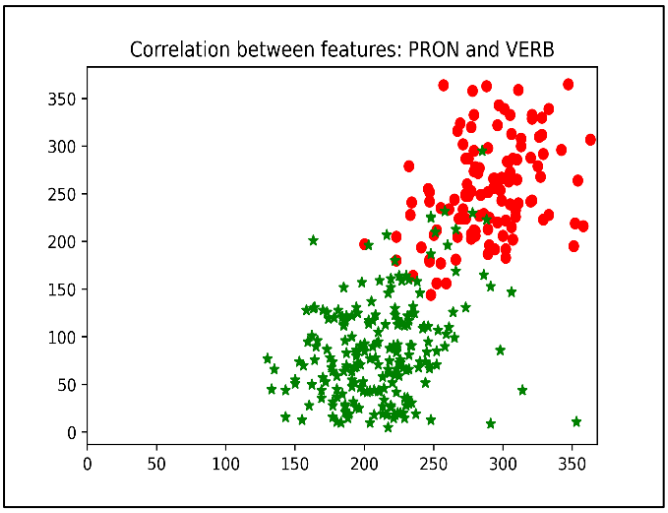
Verb	Noun / Pronoun	Adverb/ Pronoun	Verb + Adjective
Adverb	Noun/Adverb	Adjective/verb	Verb + Noun
Noun	Noun/ Verb	Adjective/Noun	Verb + Pronoun
Pronoun	Adverb/Adjective	Adjective/Pronoun	Verb + Adverb
Adjectives	Adverb/ Verb	Verb/Pronoun	Noun + Pronoun
Noun-Pronoun	Adjective+ Verb	Adjective + Noun	Noun + Adverb
Noun-Adverb	Noun-Verb	Adjective+ Pronoun	Adverb +Adjective
Adverb-Adjective	Adverb-Verb	Adverb-Pronoun	Adjective-Verb

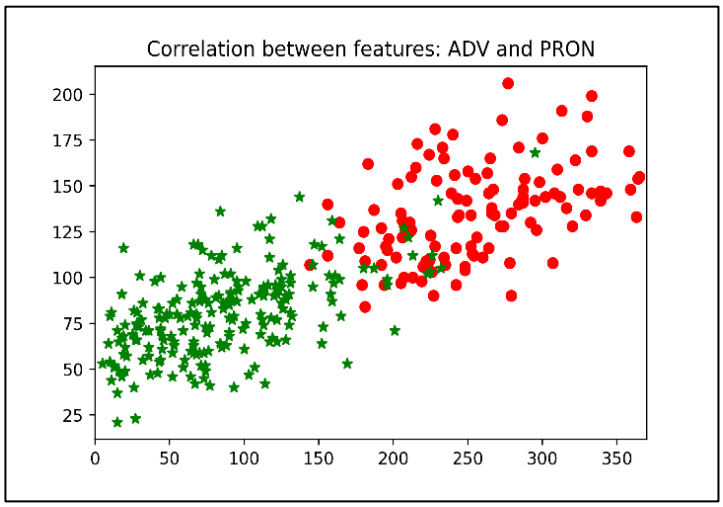
Adjective-Noun	Adjective-Pronoun	Verb-Pronoun	Noun*Pronoun
Noun*Adverb	Noun*Verb	Adverb*Adjective	Adverb*Verb
Adjective*Verb	Adjective*Noun	Adjective*Pronoun	Verb*Pronoun
Adverb*Pronoun			

Note: All the analysis is made on the basis of training model on Brown corpus data set and testing accuracy on different pairs of features. All the data are generated in code and only conclusions made from training and testing of various features are written.

(1) Linear Regression Analysis:

We have performed data analysis of various combinations of features and derived features present in Brown corpus data set. For the all set of





combinations between primary features, we analysed results:

(a) Primary features analysis: (Verb, Noun, Pronoun, Adjectives, Adverbs)

~70 % Test accuracy

~70 % Test accuracy

(All images of combinations are not shown in this report. They are used for analysis purpose and Important images are only presented in this report)

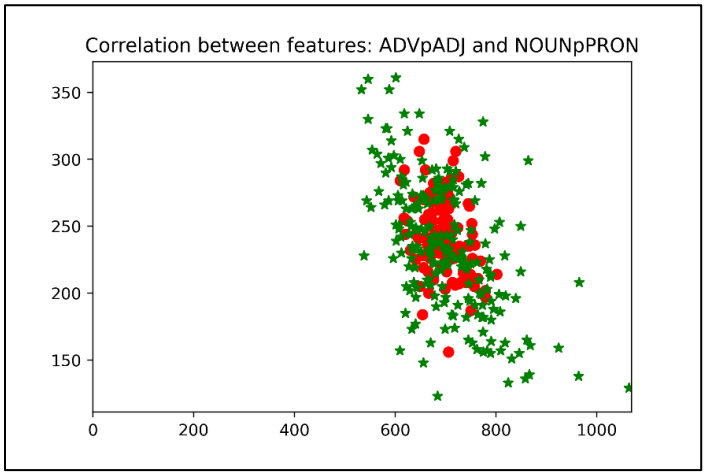
Linear regression model is trained on these 10 combinations ($^5C_2$) and above results were obtained on training and testing.

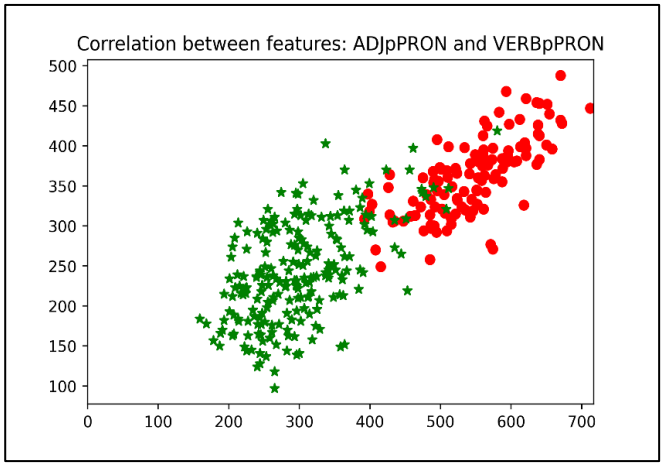
This analysis shows good linear separation boundary between two classes when: Pronoun is strongly correlated with Verb, Adjectives and Adverbs while it is poorly correlated with noun. All other combinations are also poorly correlated. The best accuracy we get on selection on dominating features is around 71%. From linear regression, ***Verb and Pronoun*** turned out to be important features for classifications while adverb and noun is least important.

(b) Analysis of ratios of features: (10 ratios of features)

From the table of analysis of features ratios on linear regression model, maximum accuracy, we got around 55% that is not good. This shows that ratios of features are not linearly classifiable. There exist some nonlinear boundaries between different feature ratios. [Adjective/Verb] and [Adverb/Noun] shows comparatively good results than other features ratios.

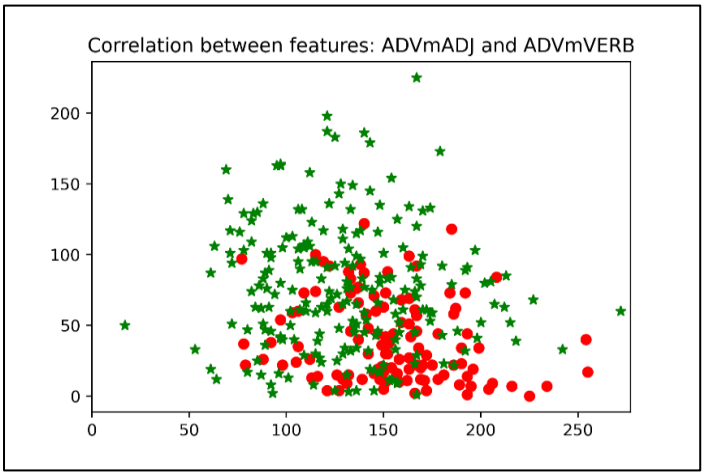
(c) *Analysis of combinations of sum of features: (10 derived features)*

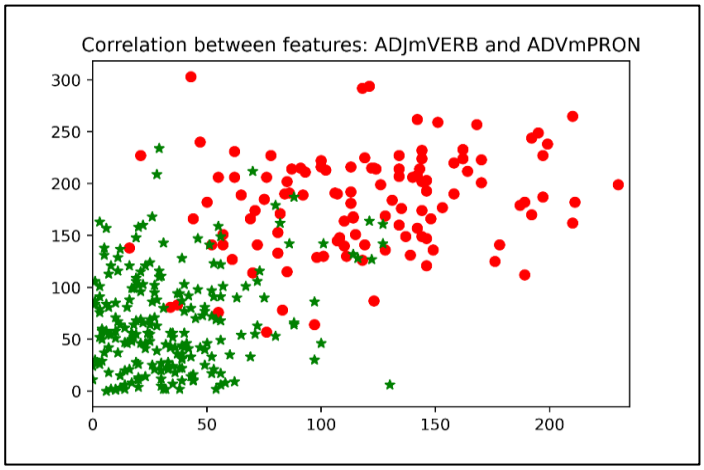




Earlier we have noticed that Pronoun shows strong correlation with Adjective, adverb and Verb and poor with Noun. Here by addition pronoun to different features create good separation boundary between two categories. Although all pairs have test accuracy in the range of 60%.

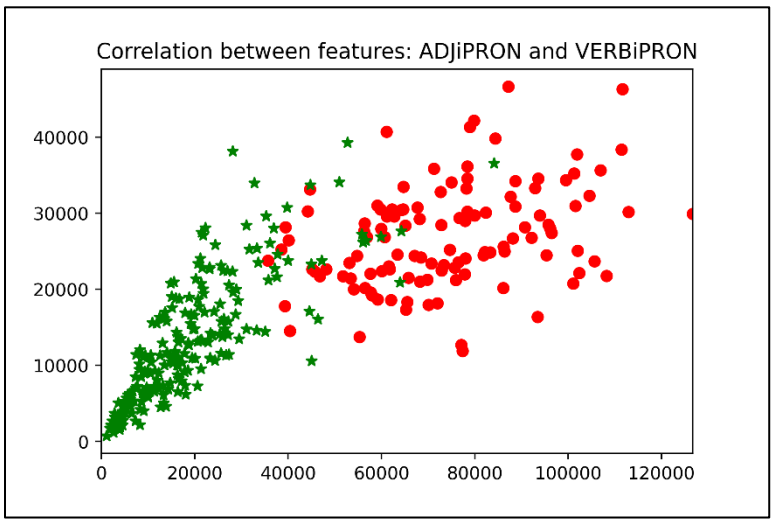
(d) Analysis of absolute difference of combinations features:

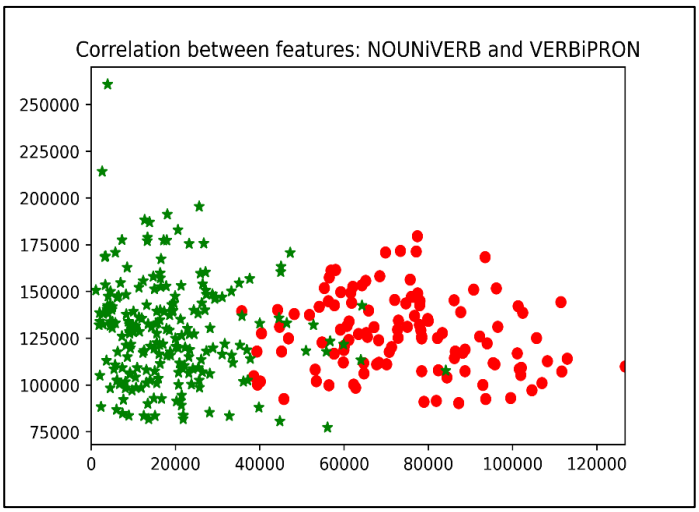




Analysis shows that there are lot of combinations possible that can give accuracy more than 60%. Features that are giving high accuracy include difference with pronoun common while those of features showing poor accuracy includes difference between adjectives and adverbs. From the figures shown at the beginning of this section, we can say that (Adjective-Adverb) and (Adverb – Verb) leads to mixing of data and poor separation boundary formed.

(e) Multiplication of combinations of features: (10 features):





Multiplying pronoun with other features is helping in creating separation boundaries. Given two plots above give best accuracy around 70%.

Verb and Pronoun multiplication turned out to be very important feature that gave above 70 % accuracy with every combination of multiplied pairs while that of Noun* Verb gives poor accuracy with all pairs.

	Sr.No	Feature_1	Feature_2	Train_Accuracy	Test_Accuracy	Overall_Accuracy
0	1	ADJ	ADV	63.788330	53.335146	59.464573
1	2	ADJ	NOUN	64.680463	51.427006	59.330509
2	3	ADJ	VERB	55.251270	55.359881	55.402479
3	4	ADJ	PRON	72.161130	68.515099	70.731446
4	5	ADV	NOUN	60.818754	47.495534	55.391776
5	6	ADV	VERB	68.662647	60.867732	65.545237
6	7	ADV	PRON	72.769954	67.131199	70.506051
7	8	NOUN	VERB	64.944969	58.786653	62.629481
8	9	NOUN	PRON	71.494428	66.004301	69.313253
9	10	VERB	PRON	74.294388	71.173788	73.122638

(Table for features and accuracy on Linear regression model)

(2) Logistic Regression analysis:

	Sr.No	Feature_1	Feature_2	Train_Accuracy	Test_Accuracy	Overall_Accuracy
0	1	ADJ	ADV	92.462312	86.567164	89.189189
1	2	ADJ	NOUN	92.462312	89.552239	91.291291
2	3	ADJ	VERB	86.934673	89.552239	89.489489
3	4	ADJ	PRON	95.979899	94.029851	94.294294
4	5	ADV	NOUN	90.452261	82.835821	88.288288
5	6	ADV	VERB	92.462312	90.298507	92.492492
6	7	ADV	PRON	96.984925	91.791045	94.294294
7	8	NOUN	VERB	94.974874	88.059701	92.192192
8	9	NOUN	PRON	94.974874	91.044776	92.792793
9	10	VERB	PRON	93.467337	91.791045	93.093093

(a) Training any two Primary features:

out of 5 features, 10 combinations are trained on logistic regression model. Following features pairs showed accuracy max. test score: Adverb and

Pronoun, Verb and Pronoun, Adjective and Pronoun, noun and pronoun. All these pairs have accuracy above 90%. That's is better than linear classifier.

Adverb Vs Pronoun has shown 92.34% accuracy.

Pronoun is showing strong correlation with all the other features.

(b) Training any two Ratios of primary features:

All the combinations of features ratios show good accuracy above 70% and max. accuracy 93% (Noun/Verb and Verb/Pronoun, Adverb/Verb and Noun/Pronoun, Adverb/Adjective and Verb/Pronoun, Adverb/Adjective and Adjective/Pronoun)

- Pair of features having Noun/Pronoun ratios common shows good performance of logistic model.
- This shows that features ratios are can be treated as potential features for classification.

(c) Training any two primary features sum pairs:

- In this case accuracy of 45 pairs of features vary from 65% (Adj. + Verb and Noun + Pronoun) to 94% (Adj. + Verb and Verb + Pronoun).
- Sum of these combinations are giving good accuracy: Adjective, Verb, Adverb and pronoun.
- Presence of Noun in one pair and presence of Adjective in another pair gives poor accuracy results (~65%)
- Since more than one combination are giving results above 90%, It is difficult to choose any one pair.

(d) Training any two absolute differences in features pairs:

- Accuracy of 45 pairs of features vary from 68% (Adv – Adjective and Adv - Verb) to 94% (Adjective - Verb and Verb - Pronoun).
- Adv – Pronoun and Adv – Noun shows almost same accuracy (~85%) with all pairs.

(e) Training any two multiplication of features pairs:

- Accuracy of 56 pairs vary from 47% (Adverb*Adjective and Noun*Verb) to 95% (Adjective*Verb and Noun*Pronoun).

- Adjective and Noun feature gives accuracy 89% (from linear regression) that is also reflected on multiplying them with pronoun.
- It is observed that presence of “pronoun” in any multiplication feature pair results into good classification.
- Noun*Pronoun is very important feature and it is giving classification accuracy (above 90%) with all multiplication pairs.

(3) SVM analysis:

	Sr.No	Feature_1	Feature_2	Train_Accuracy	Test_Accuracy	Overall_Accuracy
0	1	ADJ	ADV	92.964824	86.567164	89.489489
1	2	ADJ	NOUN	92.964824	89.552239	91.291291
2	3	ADJ	VERB	87.939698	90.298507	89.489489
3	4	ADJ	PRON	95.979899	93.283582	95.195195
4	5	ADV	NOUN	91.457286	82.835821	87.687688
5	6	ADV	VERB	92.462312	86.567164	91.591592
6	7	ADV	PRON	95.477387	91.791045	94.294294
7	8	NOUN	VERB	94.472362	88.059701	92.192192
8	9	NOUN	PRON	94.472362	94.029851	94.294294
9	10	VERB	PRON	94.472362	92.537313	92.792793

(a) Training on pair of primary features:

- Accuracy range is of the same as the order of logistic regression.
- Best classification accuracy is obtained on ‘Noun’ and ‘Pronoun’ features set (~94%).
- All the features forming pair with pronoun showed good classification accuracy. This analysis is same for Logistic regression and linear regression as well.
- While presence of Noun is showing lowest classification accuracy with all pairs. It makes noun as least important feature among all.

(b) Training any two Ratios of primary features:

- Its prediction accuracy is less than logistic regression.

- Since most of the feature pairs accuracy is lying in same range (80% to ~90%), it is difficult to choose any one pair.
- Adjective/Noun and Adverb/Verb show poor performance on model. These ratios must be excluded.

(c) Training any two primary features sum pairs:

- Pair with Adjective + Verb and Noun + Pronoun does not give good classification boundary (~62%) while adjective + adverb and Verb + Pronoun give good classification boundary (~92%) amongst all 45 pairs.
- It is because sum of Noun and Pronoun for all the documents is close to 700 to 800. This is due to increase in Noun on moving from category 1 to 0 and decrease in Pronoun. So, all data points will be close. Same is the case for Verb + Adjectives.
- But in case of Verb + Pronoun and Adjective + Pronoun, the number of Adjective and Pronoun vary by large number in both categories.

(d) Training any two absolute differences in features pairs:

- All pair of features showed accuracy in range 65% (Adverb – Adjective and Adverb – Verb) to 92% (Adjective-Verb and Verb - Pronoun).
- **Verb-Pronoun** can be treated as important features because this difference is less in category-1 while high in category-0 due to lower pronouns. Similar is the case for Noun-Verb but Noun is more in category-1 than catgory-2.
- Adjective – Noun also showed good classification results.

(e) Training any two multiplication of features pair:

- Multiplying Verb with adjective and Pronouns results in important features showing good testing accuracy (~94%).
- All the feature pairs containing Pronoun in multiplication showed good accuracy (>90%).
- Since the pronoun changes by large factor in both categories, it is dominating feature than others.
- Product of Adverb and Pronoun shows good classification accuracy with all the other feature multiplication pairs.

(4) Random forest Analysis:

	Sr.No	Feature_1	Feature_2	Train_Accuracy	Test_Accuracy	Overall_Accuracy
0	1	ADJ	ADV	100.000000	88.805970	92.792793
1	2	ADJ	NOUN	100.000000	89.552239	94.294294
2	3	ADJ	VERB	100.000000	85.074627	88.588589
3	4	ADJ	PRON	100.000000	91.791045	96.396396
4	5	ADV	NOUN	100.000000	81.343284	89.489489
5	6	ADV	VERB	99.497487	88.805970	93.393393
6	7	ADV	PRON	100.000000	89.552239	94.294294
7	8	NOUN	VERB	100.000000	87.313433	92.792793
8	9	NOUN	PRON	100.000000	88.059701	91.291291
9	10	VERB	PRON	100.000000	91.044776	93.093093

(a) Primary features:

- This classifier shows overfitting with training data (~100%) and good accuracy on testing data (81% to 92%).
- Pairs with Verb, Pronoun and Adjectives shows best classification accuracy (>91%).

(b) Ratios of pair of features:

- Adj/Pron and Adj/Noun, Adj/Pron and Noun/Pron show best classification on this model (~94% accuracy).
- Noun/Pronoun ratios shows best accuracy with all the ratios.
- Table of Ratios of features creates well separated accuracies of all the features so taking important features becomes easier.

(c) Sum of pairs of features:

- These features turned out to be less important than ratios of features because it gives lots of pairs with similar range of accuracies.
- It makes difficult to choose any two important pairs

(d) Difference of features:

- Difference of Adjective-Noun and Noun-Pronoun gives good separation boundary (~92%).
- Difference of Adjective-Adverb perform poorly with all the pairs of differences.

(e) Multiplication of features:

- **Verb and pronoun multiplication pair** with all the other features multiplication showed good accuracy result. Hence it can be taken as potential feature

	Sr.No	Feature_1	Feature_2	Score	Accuracy
0	1	ADJ	ADV	0.348412	0.850746
1	2	ADJ	NOUN	0.378363	0.873134
2	3	ADJ	VERB	0.272824	0.895522
3	4	ADJ	PRON	0.388280	0.917910
4	5	ADV	NOUN	0.371714	0.813433
5	6	ADV	VERB	0.301172	0.850746
6	7	ADV	PRON	0.228218	0.910448
7	8	NOUN	VERB	0.319764	0.880597
8	9	NOUN	PRON	0.233652	0.925373
9	10	VERB	PRON	0.188111	0.925373

- Adv * Adj and Noun*Verb performs poorly on this model.
- Out of 45 features combinations pairs, almost 40 features sets showed good accuracy. It also implies that that this model is taking care of nonlinear boundaries present between two categories in well manner.

(5) Neural Network Analysis:

(a) On Training pair of primary features on Neural Network, Noun and Pronoun & Verb and Pronoun shows good results.

(b) Ratios of features:

- Out of 45 pairs of combinations, very few pairs give poor accuracy (~65%) and other pairs lies in similar range of accuracy. So, it is difficult in case of Neural network to pick particular pair of ratios for classification purpose.

(c) Addition of features:

- Addition of pronoun to adjective and adverb created good separation boundary.
- These features are well separated as compared to ratios on wide range so selection particular feature for analysis is easier.

(d) Difference of features:

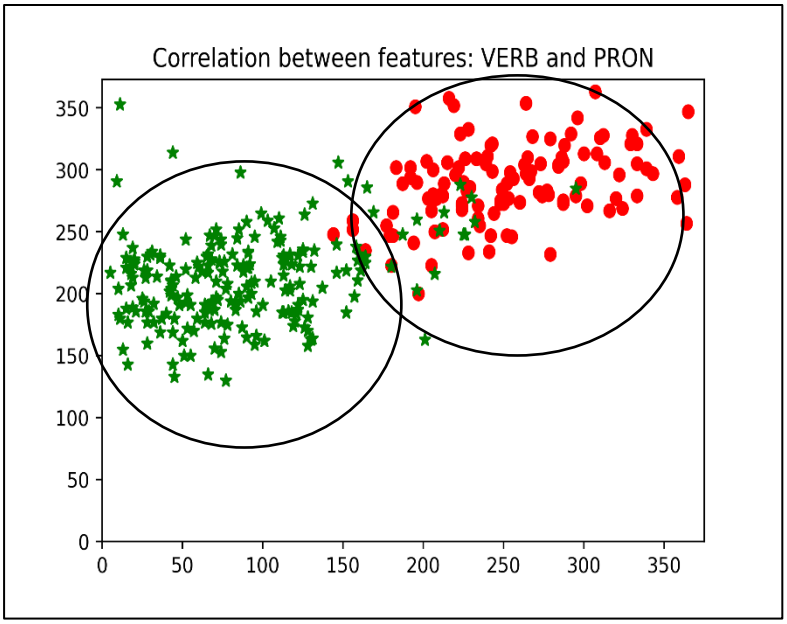
- Adjective - Verb and Verb – Pronoun is a pair showing good classification. Since all the pairs lies in same accuracy range, it is difficult to pick particular pair.

(e) Multiplication of features:

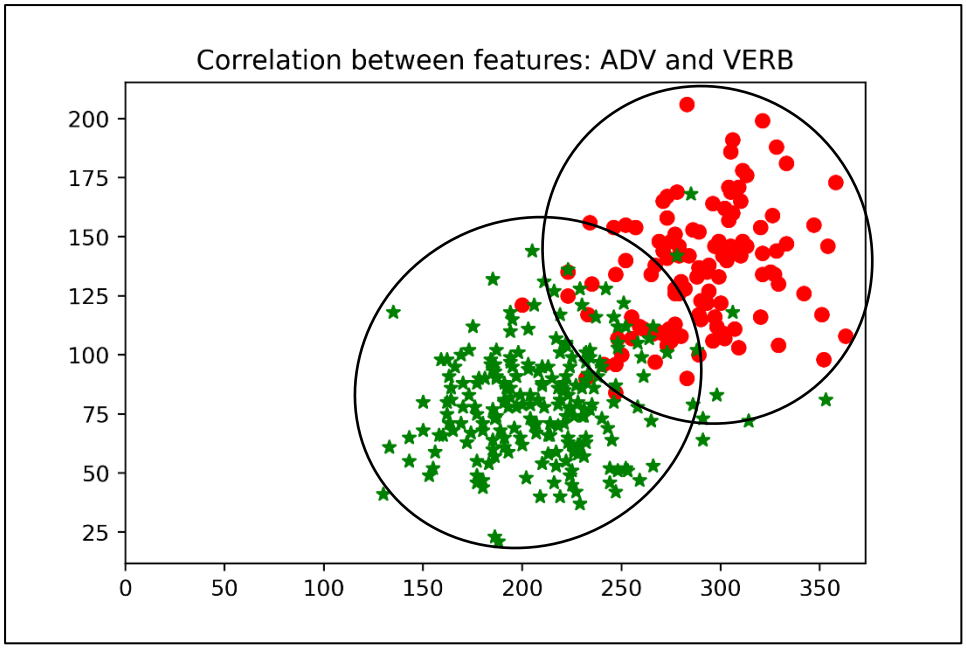
- Adverb*Verb and Noun* Pronoun shows good accuracy but since all the multiplication features lies in the same range, picking one feature is difficult.

(6) Unsupervised learning Analysis:

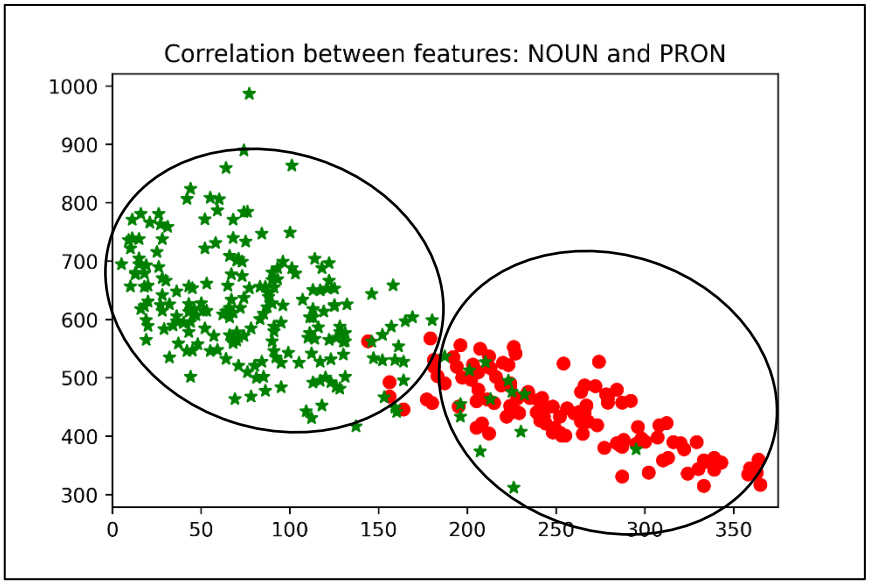
	Sr.No	Feature_1	Feature_2	Accuracy
0	1	ADJ	ADV	88.288288
1	2	ADJ	NOUN	89.489489
2	3	ADJ	VERB	18.318318
3	4	ADJ	PRON	5.705706
4	5	ADV	NOUN	86.786787
5	6	ADV	VERB	89.489489
6	7	ADV	PRON	92.792793
7	8	NOUN	VERB	10.510511
8	9	NOUN	PRON	91.891892
9	10	VERB	PRON	6.906907



(a) Primary feature analysis: Verb and Noun ratio is less scattered. Hence showing good classification accuracy while adverb and verb shows poor boundary.



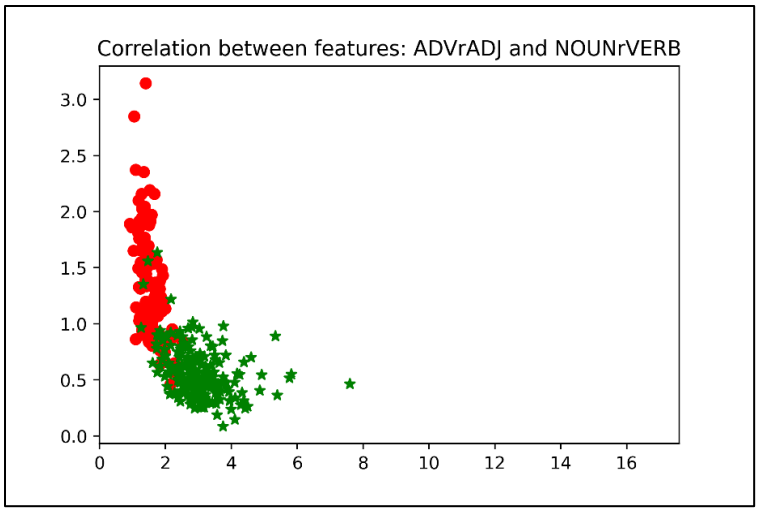
Although in unsupervised learning classification is done without using training labels, we have just shown possible boundary created by K-means clustering algorithm that may not be the case in reality. However, this analysis may change on changing train – test split. In unsupervised learning, after splitting data into train and test, Label remain of no use because there is no training required. We just compared the predicted output with actual label of the data and counted the correct label predicted by this algorithm to predict the accuracy. We don't need to split data into train and test in this case. We have directly divided data into features and classes. Then tested accuracy by comparing predicted labels with actual labels.

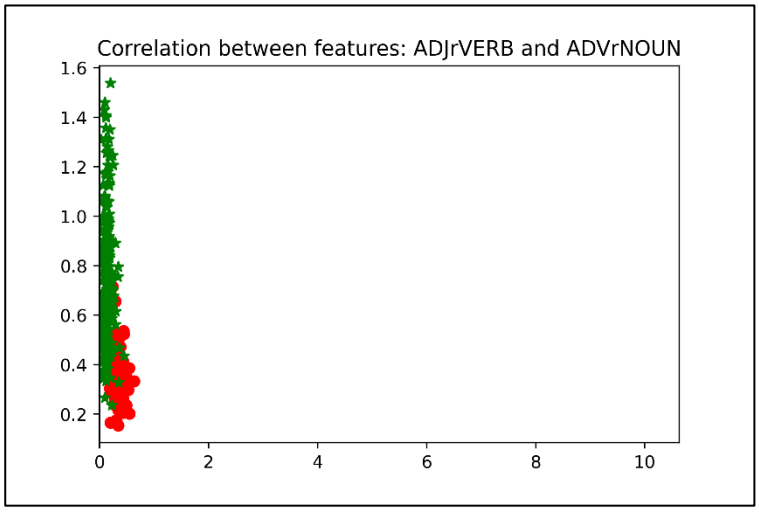


Good separation between features is possible in this case.

(b) Ratio of features:

- Poor accuracy in separating most of the features (out of 45 combinations) points into correct category.
- Adverb/adjective and Noun/verb turned out to be important feature for this algorithm while Adjective/Verb and Adverb/Noun are least significant amongst all.





(c) Addition of features:

- Adverb + Verb has plays good role in forming clusters into two categories with all other sum of features (all accuracies >90%).
- Adverb + Pronoun and noun+ pronoun result in mixing of both clusters. Hence poor accuracy is observed. This is because no of pronouns, nouns and adverbs are decreasing on moving from category 1 to 0. So, data points are merging.

(d) Difference of features:

- Best accuracy in cluster separation in observed in case of Adjective-Verb and Noun – Pronoun. It is due to decrease in Noun, pronoun and adjective in moving from category 1 to 0 while Verb remains almost of Same order for both categories. So, one feature is increasing and other is decreasing. That result in formation of clusters that are well separated.

(e) Multiplication of features:

- Adjective*Verb and Verb*Pronoun is giving good classification (94%). Reasons being increase in adjective on moving from category-1 to category-0 while decrease in pronoun while verb is of same order.

Comparative analysis of all ML methods:

Primary Features:

	Best Feature_1	Best Feature_2	Accuracy (%)
Linear Regression	Verb	Pronoun	~71
Logistic Regression	**Adjective**	**Pronoun**	**~94**
SVM	**Noun**	**Pronoun**	**~94**
Random forest	Verb	Pronoun	~92
ANN	Verb	Pronoun	~92
Unsupervised	Adjective	Pronoun	~93

Ratios of features pairs:

	Best Feature_1	Best Feature_2	Accuracy (%)
Linear Regression	Adj / Verb	Adv / Noun	~57
Logistic Regression	Adv/Pron	Noun/Pronoun	~91
SVM	**Adv/Adj**	**Adv/Pronoun**	**~93**
Random forest	**Adj/Pronoun**	**Adj/noun**	**~94**
ANN	**Adj/Pronoun**	**Adj/noun**	**~94**
Unsupervised	**Adv/Adj**	**Noun/Verb**	**~93**

Sum of Features pairs:

	Best Feature_1	Best Feature_2	Accuracy (%)
Linear Regression	Adj + Pronoun	Verb+ Pronoun	~71
Logistic Regression	Adj + Pronoun	**Verb+Pronoun**	~93
SVM	Adj + Pronoun	**Verb+Pronoun**	~93
Random forest	Adj + Verb	Verb + Pronoun	~93
ANN	Adj + Pronoun	**verb + Pronoun**	~93
Unsupervised	Adv + Adj	**Verb + Pronoun**	~93

Difference of features pairs:

	Best Feature_1	Best Feature_2	Accuracy (%)
Linear Regression	Adj - Verb	Verb - Pronoun	~66
Logistic Regression	**Adj - Verb**	**Adv -Pronoun**	~91
SVM	Adj – Verb	Verb -Pronoun	~92
Random forest	Adj - Nou	noun - Pronoun	~92
ANN	Adv - Verb	Noun-Pronoun	~92
Unsupervised	**Adj - verb**	Verb -Pronoun	~93

	Best Feature_1	Best Feature_2	Accuracy (%)
Linear Regression	Adj *Pronoun	Verb*Pronoun	~72
Logistic Regression	Noun*Verb	Noun*Pronoun	~94
SVM	**Adj *Pronoun**	**Noun*Pronoun**	~95
Random forest	**Adj *Verb**	**noun *Pronoun**	~94
ANN	Adv *Verb	Noun*Pronoun	~93
Unsupervised	**Noun*Verb**	**Verb*Pronoun**	~94

This comparative analysis shows that derived features improve accuracy over primary features. In all the cases, Nouns has also showed important role in classification.

(Analysis Done by Pooja)

Objective:

In this study, we classify fiction or non-fiction genres by performing different mathematical operations on different features available in the brown corpus dataset. We apply the different algorithms of Machine learning to see how these operations affect the distribution of dataset and accuracy for classification. We thoroughly check that is it part of speech that matters the more the different mathematical operations we perform on that matter for better classification. This not only helps in classification using machine learning but also provides useful linguistic insights.

INTRODUCTION:

Literature, in its broadest sense, is any written work. Etymologically, the term derives from Latin *litaritura/litteratura* "writing formed with letters," although some definitions include spoken or sung texts. More restrictively, it is writing that possesses literary merit Texts written in any human language can be classified in various ways, one of them being fiction and non-fiction genres.
The work done by Qureshi *et.al.* indicate the role played by salient features (adverb/adjective and adjective/pronoun ratios) in classifying fiction and non-fiction genres
Similarly, we tried to classify between fiction and non-fiction by applying mathematical operation like addition(p), subtraction(m), multiplication(i), and division(r) on the various part of speech available as different features in the dataset we have used, also calculated the accuracy by different machine learning algorithms for the same.

Data set and Method Used:

For training the Machine learning model we have used the Brown corpus data set. The nature of the distribution of texts in the Brown corpus helps us to conduct our studies conveniently.
The Corpus consists of 500 samples along with five features like no. of verb, pronoun, noun, adverb, adjectives present in each text file distributed across 15 categories (i.e Adventure, belles' letters, editorial, fiction, hobbies, reviews, government, hobbies, humor, learned, lore, mystery, news, religion, romance, science fiction).
Out of these categories:
Fiction category contains: fiction, mystery, romance, science fiction, adventure

Nonfiction category contains: News, hobbies, government, reviews, learned

Neither fiction nor nonfiction: lore, belles' letters, humor, religion, editorial.

We need to clean files from the data set that do not belong to any of the categories. After removing these, we are left with around 333 files whose data will be used for training and testing the ML models.

Analysis of Data set:

After classifying data into fiction (category = 1) and nonfiction(category = 0), analysis of additional features from given features (noun, pronoun, verb, adjectives, adverb) is done. These additional features are combinations of ratios

primary features. It will help to figures out which features contribute most to solving a classification problem.

Table of features Analyzed

Features available in Brown corpus dataset	Derived features obtain by performing a mathematical operation (multiplication)	(division)
Verb	Noun*Pronoun	Noun / Pronoun
Adverb	Noun*Adverb	Noun/Adverb
Noun	Noun*Verb	Noun/ Verb
Pronoun	Adverb*Adjective	Adverb/Adjective
Adjectives	Adverb*Verb	Adverb/ Verb
	Adverb*Pronoun	Adverb/ Pronoun
	Adjective*Verb	Adjective/verb
	Adjective*Noun	Adjective/Noun
	Adjective*Pronoun	Adjective/Pronoun
	Verb*Pronoun	Verb/Pronoun

For calculating, the accuracy of the classification we have used the following supervised and unsupervised machine learning algorithms.

1. Linear regression
2. Logistic regression
3. SVM
4. Random forest
5. ANN
6. K-Means clustering

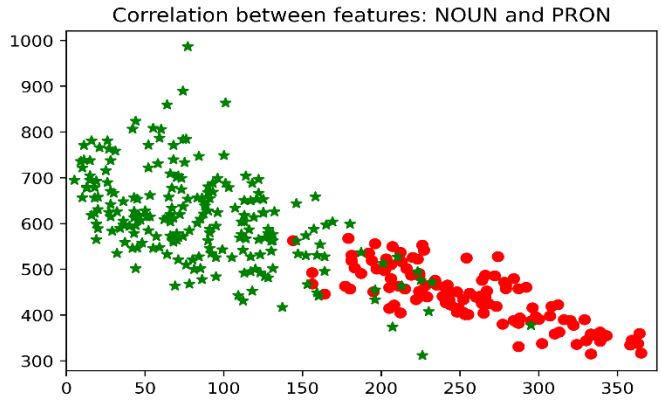
Result and discussion:

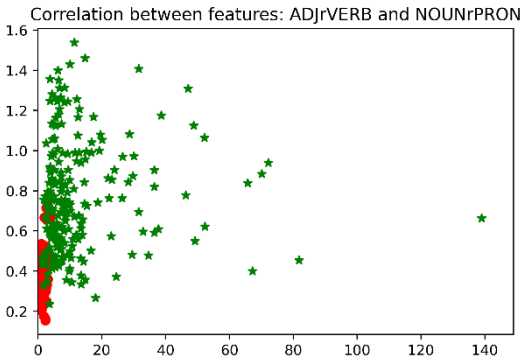
This section describes our experiments aimed to classify texts into fictional and non-fictional genres using machine learning. The next subsection describes various linguistic features we deploy in detail and the use of feature selection to identify the most useful feature.

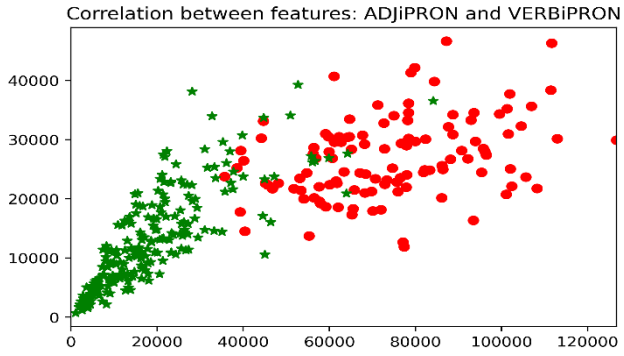
1. Linear Regression Analysis:

In the below one can see the classification between fictional (red) and non-fictional genre (green) done with the help of a correlation plot. From the

correlation plot, one can see that a linear classification boundary is easily obtained for the correlation between a primary feature and derives feature obtain by multiplication of primary feature.







In this Linear Regression analysis, a total $^{5}C_{2}$ combination is possible out of which we observed that for the primary feature in the dataset when the pronoun is taken as one of the feature to correlates with other primary feature it gives maximum accuracy (above 65%) .

Feature_1	Feature_2	Test_Accuracy
ADJ	PRON	68.515099
ADV	PRON	67.131199
NOUN	PRON	66.004301
VERB	PRON	71.173788

In the derived feature total $^{10}C_2$ i.e 45 .Among all these correlation when we division as mathematical operation for the deriving a feature from primary features correlate it with one another its accuracy is decreased irrespective of any feature being used whereas when derived feature is obtained from pronoun in the multiplication case accuracy is slightly increase or remain the same as compare to primary features.

Feature_1	Feature_2	Test_Accuracy
ADJrVERB	ADVrNOUN	57.976084
ADJrVERB	NOUNrVERB	53.015764
ADJrNOUN	ADVrNOUN	55.530897
ADJrNOUN	NOUNrVERB	53.240742
ADVrADJ	ADVrNOUN	53.224045
ADVrADJ	NOUNrVERB	55.121916
ADVrNOUN	ADVrVERB	56.900787
ADVrNOUN	NOUNrVERB	53.274249

Feature_1	Feature_2	Test_Accuracy
ADJiPRON	VERBiPRON	72.705533
ADJiVERB	VERBiPRON	72.381375
ADJiNOUN	VERBiPRON	71.44006
ADViADJ	VERBiPRON	72.647446
ADViNOUN	VERBiPRON	73.130525
ADViPRON	NOUNiVERB	70.779213
ADViPRON	NOUNiPRON	70.061693
ADViPRON	VERBiPRON	72.400823
ADViVERB	VERBiPRON	72.432053
NOUNiVERB	VERBiPRON	72.628073
NOUNiPRON	VERBiPRON	72.619864

All the correlation plot and test accuracy result is used for analysis purpose and is not shown here for the sake of better understanding of the obtained result.

Logistic regression:

In this Logistic Regression analysis, out o 10 possible combinations we observed high accuracy(above 85%) for almost all correlation among the primary features specifically a higher accuracy(appox. above 90%) for the pronoun when used as one of the primary feature.

Feature_1	Feature_2	Test_Accuracy
ADJ	ADV	86.567164
ADJ	NOUN	89.552239
ADJ	VERB	89.552239
ADJ	PRON	94.029851
ADV	NOUN	82.835821

ADV	VERB	90.298507
ADV	PRON	91.791045
NOUN	VERB	88.059701
NOUN	PRON	91.044776

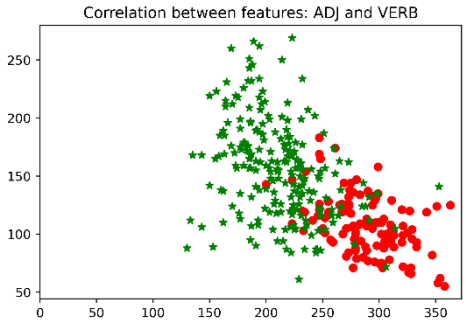
In the derived feature out of 45 combinations, accuracy remain same or above 90% when taken ratio of different primary feature on taking correlate with one another. Whereas when derived feature is obtained from the multiplication case accuracy is slightly increase almost for the cases (as compare to primary features) specifically when one of the derived feature contain multiplication of pronoun.

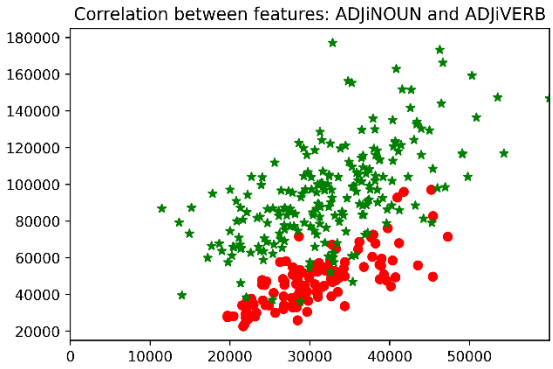
Feature_1	Feature_2	Test_Accuracy
ADJrPRON	ADJrNOUN	91.791045
ADJrPRON	ADVrPRON	91.044776
ADJrPRON	NOUNrPRON	91.044776
ADJrPRON	VERBrPRON	91.044776
ADJrVERB	ADVrNOUN	90.298507
ADJrNOUN	NOUNrPRON	91.791045
ADJrNOUN	VERBrPRON	90.298507
ADVrADJ	ADVrPRON	91.044776
ADVrADJ	NOUNrVERB	90.298507
ADVrPRON	NOUNrPRON	91.791045
NOUNrPRON	VERBrPRON	91.044776

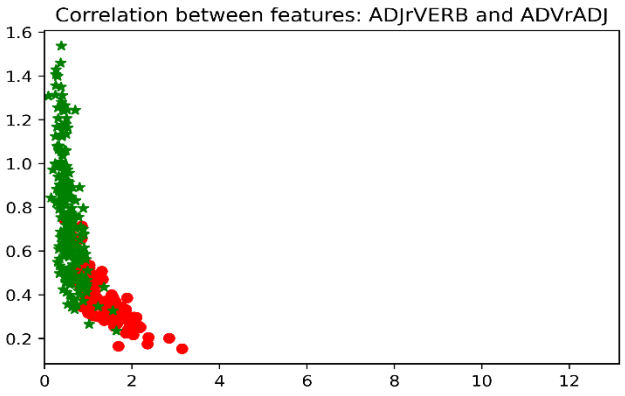
Feature_1	Feature_2	Test_Accuracy
ADJiPRON	ADJiNOUN	90.298507
ADJiPRON	ADViPRON	91.044776
ADJiPRON	ADViVERB	91.044776
ADJiPRON	NOUNiPRON	93.283582
ADJiPRON	VERBiPRON	92.537313
ADJiVERB	ADViPRON	91.791045
ADJiVERB	ADViVERB	91.044776
ADJiVERB	NOUNiPRON	95.522388
ADJiVERB	VERBiPRON	94.029851
ADJiNOUN	ADViPRON	91.791045
ADJiNOUN	ADViVERB	90.298507
ADJiNOUN	NOUNiPRON	93.283582
ADJiNOUN	VERBiPRON	92.537313
ADViADJ	ADViPRON	92.537313
ADViADJ	ADViVERB	92.537313
ADViADJ	NOUNiPRON	94.029851
ADViADJ	VERBiPRON	92.537313
ADViNOUN	ADViPRON	91.044776
ADViNOUN	ADViVERB	91.044776
ADViNOUN	NOUNiPRON	94.029851
ADViNOUN	VERBiPRON	94.029851
ADViPRON	ADViVERB	91.791045
ADViPRON	NOUNiVERB	92.537313
ADViPRON	NOUNiPRON	92.537313

ADViPRON	VERBiPRON	91.791045
ADViVERB	NOUNiPRON	94.029851
ADViVERB	VERBiPRON	92.537313
NOUNiVERB	NOUNiPRON	95.522388
NOUNiVERB	VERBiPRON	91.791045
NOUNiPRON	VERBiPRON	94.029851

In the below one can see the classification between fictional (red) and non-fictional genre(green) done with the help of correlation plot. From the correlation plot one can see that linear classification boundary is easy obtained for the correlation between primary feature and derives feature obtain by multiplication of primary feature







SVM:

In this SVM analysis, out o 10 possible combinations we observed high accuracy of classification between fiction and non-fiction genre above 90% specifically for pronoun when correlated with other primary feature , also one case for above 90% is when adjective is correlated with verb.

Feature_1	Feature_2	Test_Accuracy
ADJ	VERB	90.298507
ADJ	PRON	93.283582
ADV	PRON	91.791045
NOUN	PRON	94.029851
VERB	PRON	92.537313

In the derived feature out of 45 combinations, accuracy is above 90% for some of the derived features for the case of multiplication whereas when taken ratio of different primary feature on taking correlate with one another accuracy is high for most of the derived feature as data shown in tables.

Feature_1	Feature_2	Test_Accuracy
ADJrPRON	ADVrADJ	91.044776

ADJrPRON	NOUNrPRON	90.298507
ADJrVERB	ADVrNOUN	91.044776
ADJrVERB	NOUNrPRON	90.298507
ADVrADJ	ADVrPRON	92.537313
ADVrADJ	NOUNrVERB	90.298507
ADVrADJ	NOUNrPRON	90.298507

Feature_1	Feature_2	Test_Accuracy
ADJiPRON	ADJiNOUN	91.044776
ADJiPRON	ADViPRON	91.791045
ADJiPRON	VERBiPRON	91.791045
ADJiVERB	ADViPRON	91.791045
ADJiVERB	ADViVERB	90.298507
ADJiVERB	NOUNiPRON	95.522388
ADJiVERB	VERBiPRON	94.029851
ADJiNOUN	ADViPRON	93.283582
ADJiNOUN	ADViVERB	90.298507
ADJiNOUN	NOUNiPRON	91.791045
ADJiNOUN	VERBiPRON	92.537313
ADViADJ	ADViPRON	93.283582
ADViADJ	ADViVERB	91.791045
ADViADJ	NOUNiPRON	93.283582
ADViADJ	VERBiPRON	93.283582
ADViNOUN	ADViPRON	91.044776
ADViNOUN	NOUNiPRON	94.029851
ADViNOUN	VERBiPRON	94.029851
ADViPRON	ADViVERB	90.298507
ADViPRON	NOUNiVERB	90.298507
ADViPRON	NOUNiPRON	91.791045
ADViPRON	VERBiPRON	92.537313
ADViVERB	NOUNiPRON	94.776119
ADViVERB	VERBiPRON	91.791045
NOUNiVERB	NOUNiPRON	94.776119
NOUNiVERB	VERBiPRON	93.283582
NOUNiPRON	VERBiPRON	91.044776

Random Forest:

In this Random forest analysis, out of 10 possible combinations we observed an accuracy of 64 % for different combination of primary feature for classification between fiction and non-fiction genre .

Feature_1	Feature_2	Test_Accuracy
ADJ	VERB	64.179104
ADJ	PRON	64.179104
ADV	NOUN	64.179104
ADV	VERB	64.179104
ADV	PRON	64.179104
NOUN	VERB	64.179104
NOUN	PRON	64.179104
VERB	PRON	64.179104

Feature_1	Feature_2	Test_Accuracy
ADJmPRON	ADJmVERB	64.179104
ADJmPRON	ADJmNOUN	64.179104
ADJmPRON	ADVmADJ	64.179104
ADJmPRON	ADVmNOUN	64.179104
ADJmPRON	ADVmPRON	64.179104
ADJmPRON	ADVmVERB	64.179104
ADJmPRON	NOUNmPRON	64.179104

In the derived feature out of 45 combinations, accuracy is increased when adjective or adverb is used for the ratio for derive the feature. Whereas surprisingly when the derived feature is obtained from the multiplication case accuracy remains same for most of the cases (as compared to primary features) even when adjective and adverb is used for deriving the feature.

Feature_1	Feature_2	Test_Accuracy
ADJrVERB	ADVrADJ	82.835821
ADJrVERB	ADVrNOUN	83.58209
ADJrNOUN	ADVrADJ	77.61194
ADJrNOUN	ADVrNOUN	85.074627
ADVrADJ	ADVrNOUN	85.074627
ADVrADJ	ADVrPRON	84.328358
ADVrADJ	ADVrVERB	80.597015
ADVrADJ	NOUNrPRON	88.059701
ADVrADJ	VERBrPRON	84.328358
ADVrNOUN	ADVrPRON	82.089552
ADVrNOUN	ADVrVERB	83.58209
ADVrNOUN	VERBrPRON	82.089552
ADJmPRON	VERBmPRON	64.179104
ADJmVERB	ADJmNOUN	64.179104
ADJmVERB	ADVmADJ	64.179104
ADJmVERB	ADVmNOUN	64.179104
ADJmVERB	ADVmPRON	64.179104
ADJmVERB	ADVmVERB	64.179104
ADJmVERB	NOUNmVERB	64.179104
ADJmVERB	NOUNmPRON	64.179104
ADJmVERB	VERBmPRON	64.179104
ADJmNOUN	ADVmPRON	64.179104
ADJmNOUN	ADVmVERB	64.179104
ADVmADJ	ADVmPRON	64.179104
ADVmADJ	ADVmVERB	64.179104
ADVmNOUN	ADVmPRON	64.179104
ADVmNOUN	ADVmVERB	64.179104
ADVmPRON	ADVmVERB	64.179104
ADVmPRON	VERBmPRON	64.179104

ANN:

In this ANN analysis, out of 10 possible combinations we observed accuracy is above 85% specifically for adjective and adverb is used as a feature to classify between fiction and non-fiction genres .

Feature_1	Feature_2	Test_Accuracy
ADJ	ADV	85.8209
ADJ	NOUN	87.3134
ADJ	VERB	87.3134
ADJ	PRON	93.2836
ADV	NOUN	82.8358
ADV	VERB	87.3134
ADV	PRON	91.0448
NOUN	VERB	88.0597
NOUN	PRON	91.0448
VERB	PRON	92.5373

Feature_1	Feature_2	Test_Accuracy
ADJiPRON	ADJiNOUN	90.2985
ADJiPRON	ADViPRON	91.791
ADJiPRON	NOUNiPRON	90.2985
ADJiPRON	VERBiPRON	91.791
ADJiVERB	ADViPRON	91.791
ADJiVERB	NOUNiPRON	95.5224
ADJiVERB	VERBiPRON	91.791
ADJiNOUN	ADViPRON	93.2836
ADJiNOUN	NOUNiPRON	91.0448
ADJiNOUN	VERBiPRON	91.791
ADViADJ	ADViPRON	93.2836
ADViADJ	NOUNiPRON	90.2985
ADViADJ	VERBiPRON	93.2836
ADViNOUN	ADViPRON	91.0448
ADViNOUN	NOUNiPRON	90.2985
ADViNOUN	VERBiPRON	91.0448
ADViPRON	ADViVERB	90.2985
ADViPRON	NOUNiVERB	90.2985
ADViPRON	NOUNiPRON	92.5373
ADViPRON	VERBiPRON	92.5373
ADViVERB	NOUNiPRON	93.2836
ADViVERB	VERBiPRON	91.0448
NOUNiVERB	NOUNiPRON	93.2836
NOUNiVERB	VERBiPRON	94.0298
NOUNiPRON	VERBiPRON	92.5373

In the derived feature out of 45 combinations, accuracy is slightly increased when adjective or adverb is used for the ratio for deriving the feature. Same with the case when a derived feature is obtained from the multiplication case accuracy remain same or increased slightly same for most of the cases (as compare to primary features) even when adjective and adverb is used for deriving the feature

Feature_1	Feature_2	Test_Accuracy
ADJrPRON	ADJrNOUN	94.0298
ADJrPRON	ADVrADJ	91.0448
ADJrPRON	ADVrPRON	91.791
ADJrPRON	ADVrVERB	92.5373
ADJrPRON	NOUNrPRON	93.2836
ADJrPRON	VERBrPRON	91.791
ADJrVERB	NOUNrPRON	92.5373
ADJrVERB	VERBrPRON	91.791
ADJrNOUN	NOUNrPRON	94.0298
ADJrNOUN	VERBrPRON	90.2985
ADVrADJ	ADVrPRON	92.5373
ADVrADJ	NOUNrPRON	91.0448
ADVrPRON	NOUNrPRON	92.5373
ADVrPRON	VERBrPRON	90.2985
ADVrVERB	NOUNrPRON	91.0448
NOUNrPRON	VERBrPRON	91.0448

K means clustering (unsupervised Machine learning):

This is an unsupervised learning algorithm, In this analysis, out of 10 possible combinations we observed high accuracy of classification between fiction and non-fiction genre above 85% specifically for Adjective adverb is used for along with the other three as shown in the table correlated with the primary feature, also one case for above 90% is when an adjective is correlated with verb.

Feature_1	Feature_2	Test_Accuracy
ADJ	ADV	88.288288
ADJ	NOUN	89.489489
ADJ	VERB	81.681682
ADJ	PRON	94.294294
ADV	VERB	89.78979
ADV	PRON	92.792793
NOUN	VERB	89.489489
NOUN	PRON	91.891892
VERB	PRON	92.792793

For the derived feature out of 45 combinations, accuracy is slightly increased when adjective or adverb is used for the ratio for deriving the feature. As shown in tabl . Same with the case when the derived feature is obtained from the and division adjective and adverb play an important role in deciding the accuracy of the divided feature itself.

Feature_1	Feature_2	Test_Accuracy
ADJrPRON	ADVrADJ	89.78979
ADJrPRON	ADVrNOUN	89.489489
ADJrVERB	ADVrADJ	88.588589
ADJrVERB	ADVrNOUN	92.792793
ADVrADJ	ADVrNOUN	87.687688
ADVrADJ	ADVrPRON	89.489489
ADVrADJ	NOUNrVERB	93.093093
ADVrADJ	NOUNrPRON	89.78979
ADVrADJ	VERBrPRON	88.888889
ADVrNOUN	ADVrPRON	89.489489
ADVrNOUN	NOUNrPRON	89.489489
ADVrNOUN	VERBrPRON	87.987988

Feature_1	Feature_2	Test_Accuracy
ADJiPRON	ADJiNOUN	89.489489
ADJiPRON	ADViVERB	90.990991
ADJiPRON	VERBiPRON	89.78979
ADJiVERB	ADViPRON	92.192192
ADJiNOUN	ADViVERB	93.093093
ADJiNOUN	NOUNiPRON	90.09009
ADViNOUN	NOUNiPRON	83.483483
ADViPRON	NOUNiPRON	92.192192
ADViVERB	NOUNiPRON	94.894895
NOUNiVERB	NOUNiPRON	84.684685
NOUNiPRON	VERBiPRON	92.492492

Conclusion:

From the above study, we concluded that primary features in the brown corpus dataset, which the part of speech in this data set play a significant role in deciding the accuracy obtained for classification between the fiction and nonfiction genre. However, for the larger data set, one can expect to get better accuracy from the same dataset on applying different mathematical operations on the dataset. We perform different algorithm for the above purpose we observed that logistic regression gives the maximum accuracy for the classification followed by ANN and SVM (in Supervised machine learning algorithm) and K means clustering (unsupervised machine learning algorithm). Also, Random forest and Linear regression algorithm is not the good choice for the classification however Random forest gives better accuracy when division is used on its primary feature and Linear regression give better accuracy on derived features when multiplication operation is used on primary features.