INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

MODEL BUIDING AND EVALUATION on SENTIMENT ANALYSIS

By
Ibteda Azeem
M.Balakrishna

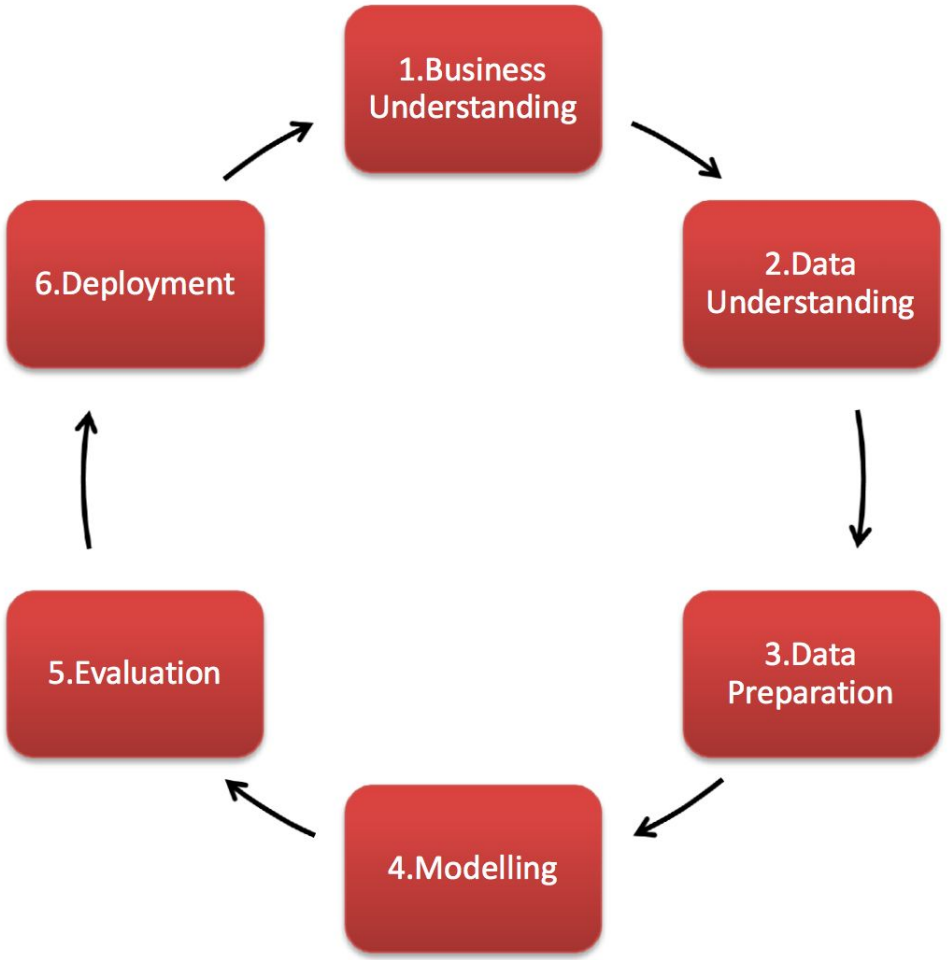
About us

- **Team Member-1**: I am Ibteda Azeem.I have done B.Tech in Electrical and Electronics Engineering from United College of Engineering and Research affiliated to A.P.J Abdul Kalam Technical University from Prayagraj(Allahabad).I want to make my career in the field of Data Science and update myself with the new technology.

- **Team Member-2:** Myself M.Balakrishna. Recently completed my Degree in B.Com in 2020 at MNR Degree College

  I want to learn data science for skills and career transformation. It has great job opportunities.

CRISP DM FRAMEWORK



Objective of the project

- The task here is to transform the given data(i.e. Text files) to tabular format(i.e. csv file).
- Perform data preprocessing on the given text data and convert it into numerical vectors.
- Build models to predict the Score and Sentiment of a given text review.

Dataframe

	RowId	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	ReviewSummary	ReviewText
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00171APVA	A21BT40VZCCYT4	Carol A. Reed	0	0	5	1351209600	Healthy Dog Food	This is a very healthy dog food. Good for thei...
2	3	B0019CW0HE	A2P6ACFZ8FTNVV	Melissa Benjamin	0	1	1	1331164800	Bad	I fed this to my Golden Retriever and he hated...
3	4	B006F2NYI2	A132DJVI37RB4X	Scottdrum	2	5	2	1332374400	Not hot, not habanero	I have to admit, I was a sucker for the large ...
4	5	B000P41A28	A82WIMR4RSVLI	Emrose mom	0	1	4	1337472000	The best weve tried so far	We have a 7 week old... He had gas and constip...

- Our dataframe consists of **568454** rows and **10** Columns/Features

```
df.dtypes
```

```
RowId                     int32
ProductId                object
UserId                   object
ProfileName              object
HelpfulnessNumerator      int64
HelpfulnessDenominator    int64
Score                     int64
Time                      int64
ReviewSummary            object
ReviewText               object
dtype: object
```

INNOMATICS RESEARCH LABS

Data Manipulation

- We created a new column which calculates the helpfulness fraction by dividing the helpfulness numerator by helpfulness denominator.
- We created two more columns and labels according to the value present in those columns.
- For Helpfulness label we utilised the Helpfulness column:

  Values greater than 0.5 as “>0.5”

  Values between 0.5 and 0.25 as”<0.5”

  Values less than 0.25 as “Useless”

- For Sentiment column we utilised Score column as:

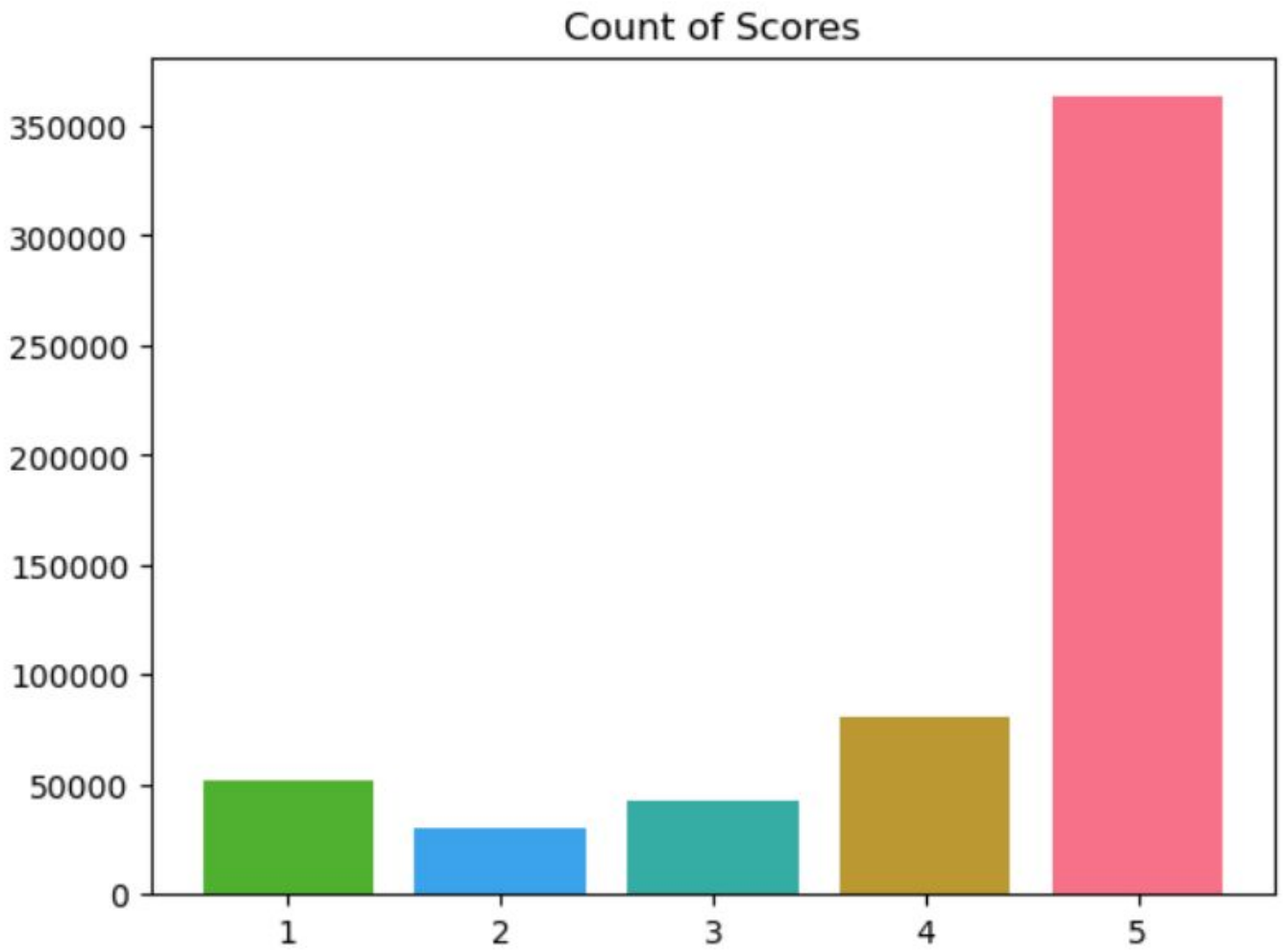
  score=4,5 then sentiment=”positive”

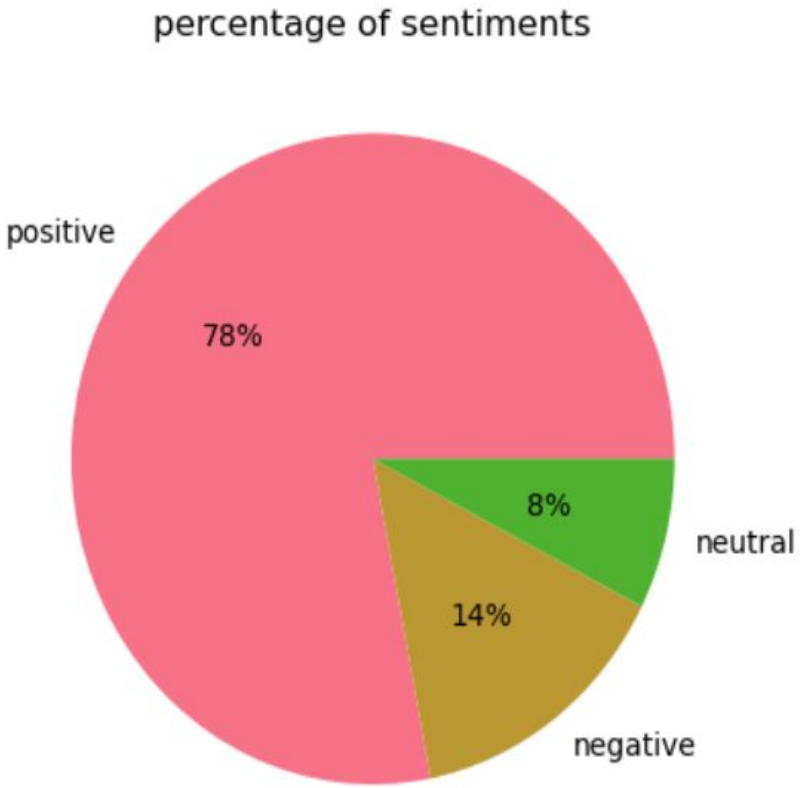
  score=3,then sentiment=”neutral”

  score=1,2,then sentiment=”negative”.

- We added another column which counts the number of words present in ReviewText column.

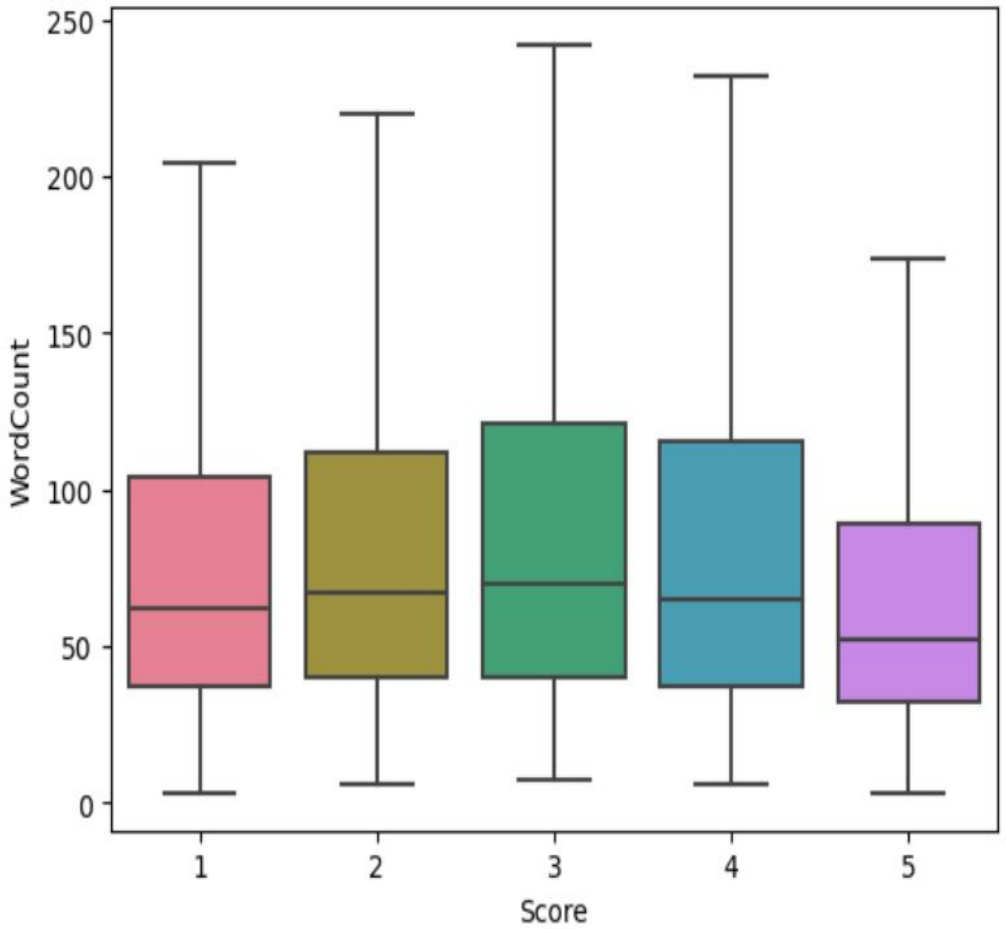
Score counts and sentiments





INNOMATICS RESEARCH LABS

Word Count of Summary Text Based in Score



Score	WordCount
1	87.323697
2	90.040881
3	95.645755
4	91.393156
5	74.167305

Mean

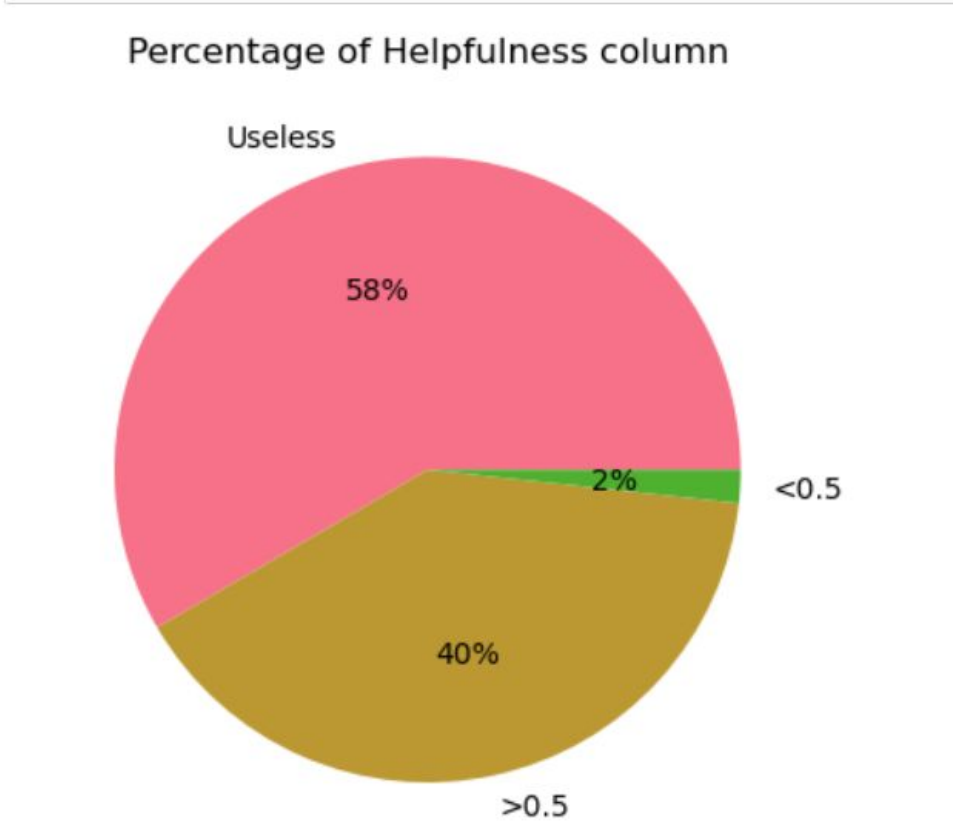
:

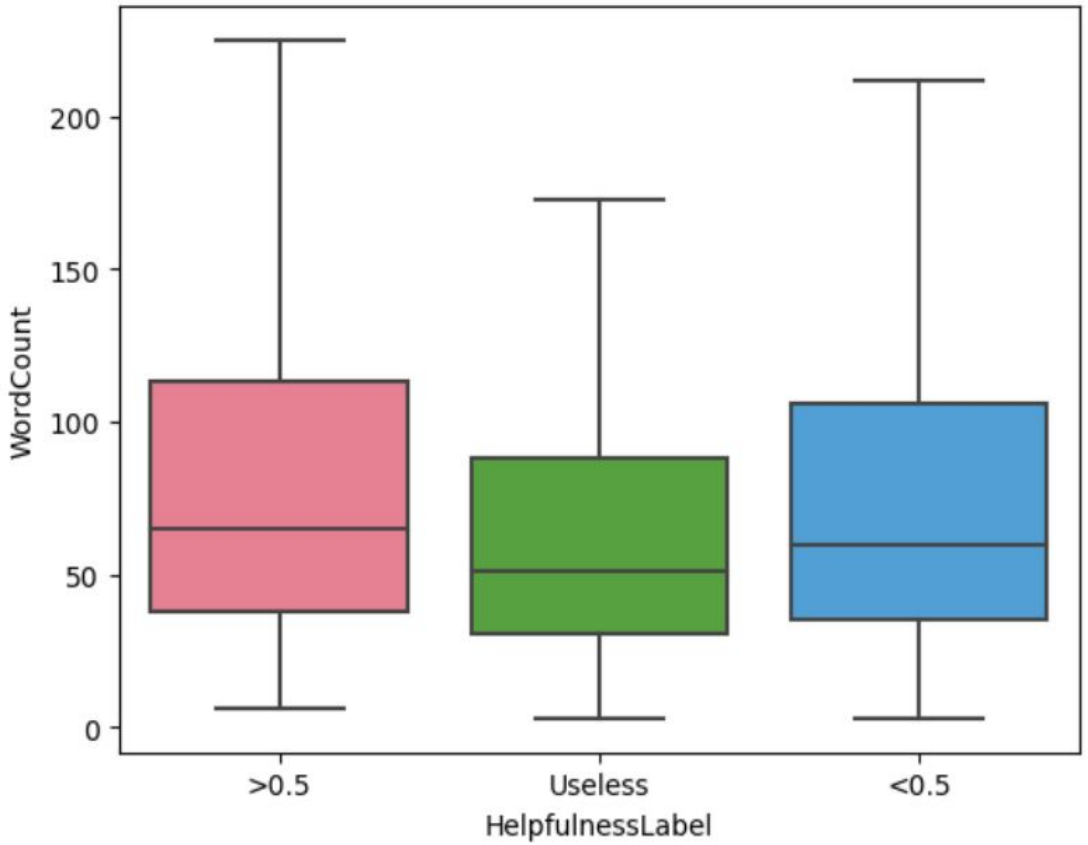
Score	WordCount
1	62.0
2	67.0
3	70.0
4	65.0
5	52.0

Median

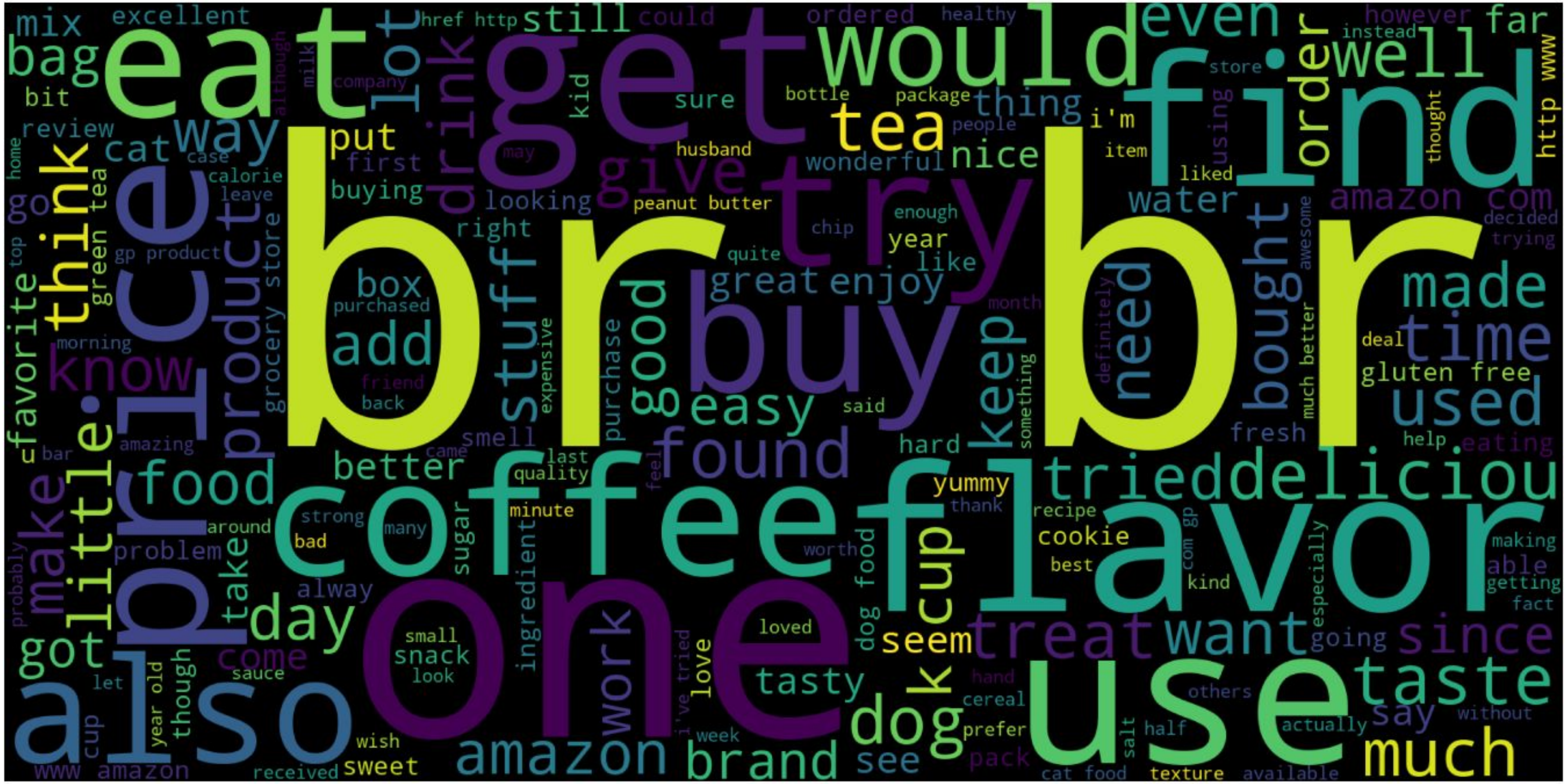
INNOMATICS RESEARCH LABS

Helpfulness Count

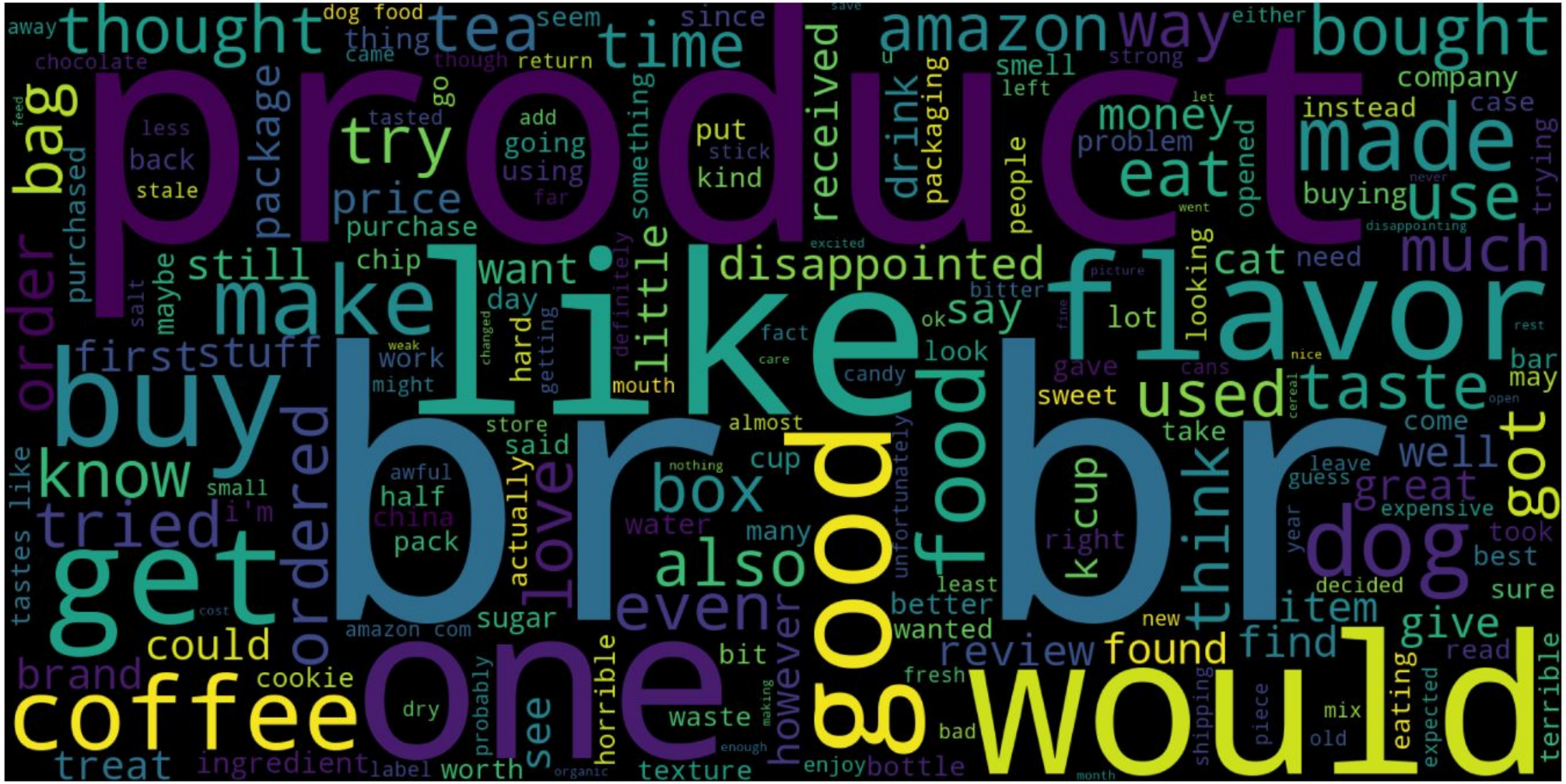




Positive Reviews Word Cloud



Negative Review Word Cloud



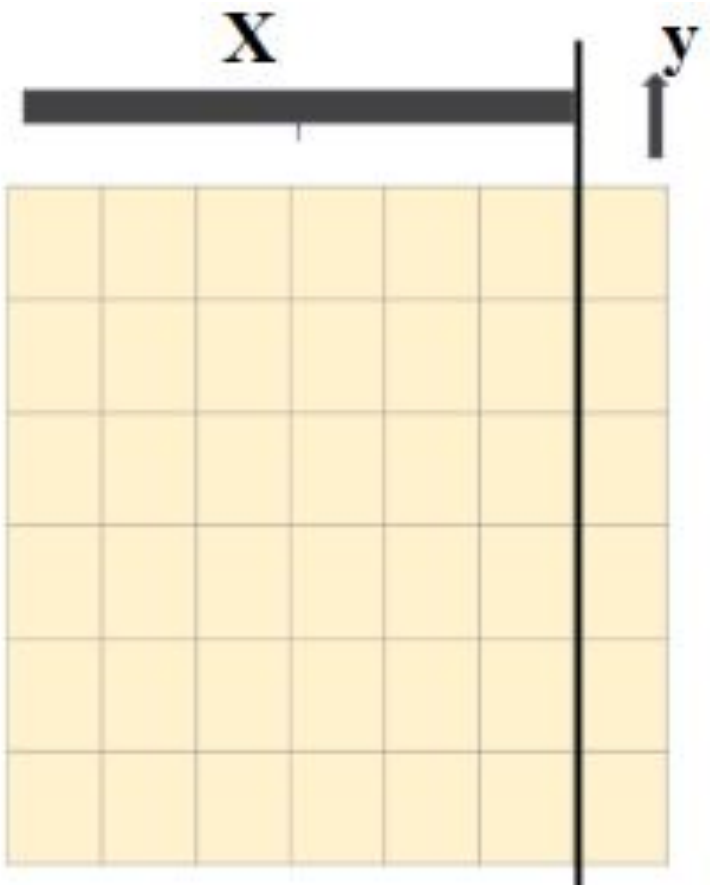
Insights:

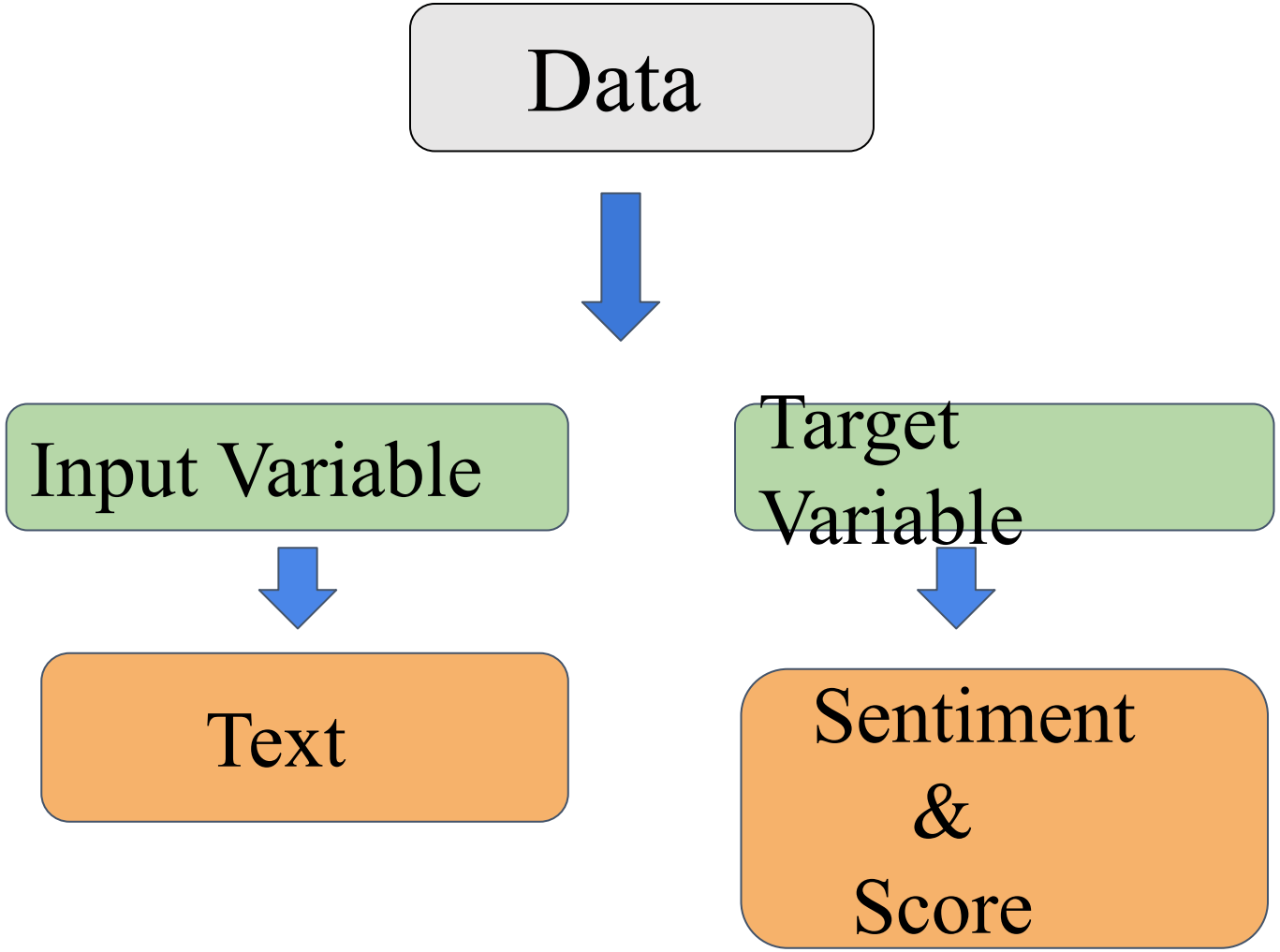
- The number of users who have given reviews are 256047.It means many people have given more than one review.
- The number of unique products are 74,258 shows that many reviews are for same product.
- Score 5 is dominating the score distribution having more than 350000 count.Least count is of score 2 having count around 40000
- 78% of the reviews are having positive sentiment,14% are having negative sentiment while 8% are having neutral sentiment.
- Reviews with score 3 have highest word count average as well as median.Score 5 have the least word count average and median.It shows that people giving very positive reviews are using less words to express their sentiment.
- 58% of reviews have helpfulness less than 25% while 40% of the reviews are having helpfulness greater than 50%.
- Reviews which have helpfulness greater than 50% are hiaving the highest wordcount.It means that more lengthy reviews are more helpful

Machine Learning Framework

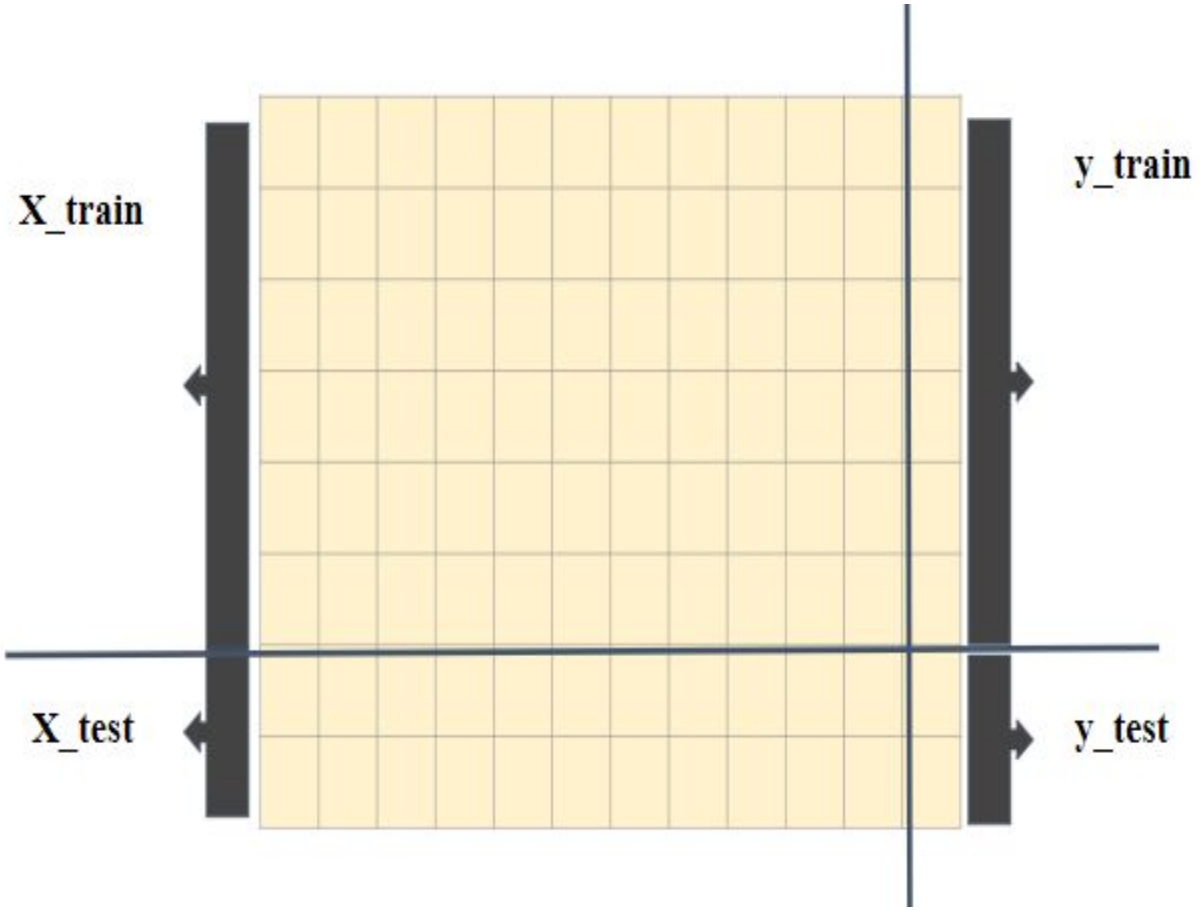
- We will first look at the problem statement and if the target variable is given in the historical data then we are going to use Supervised Machine Learning Algorithm.
- If the target variable is not given then we use Unsupervised Machine Learning Algorithm.
- Identify the input variables and the target or output variable.The input variables helps us decide the data preparation strategies and the target variable helps us decide the type of task and the evaluation metric.
- split the data into train and test.
- data preparation on X_train(fit and transform both)
- Building the model.
- data preparation on X_test(only transformation)
- Prediction on test data.
- Evaluation of model.

Identifying the Input and Target Variables and Separating them:





Train Test Split

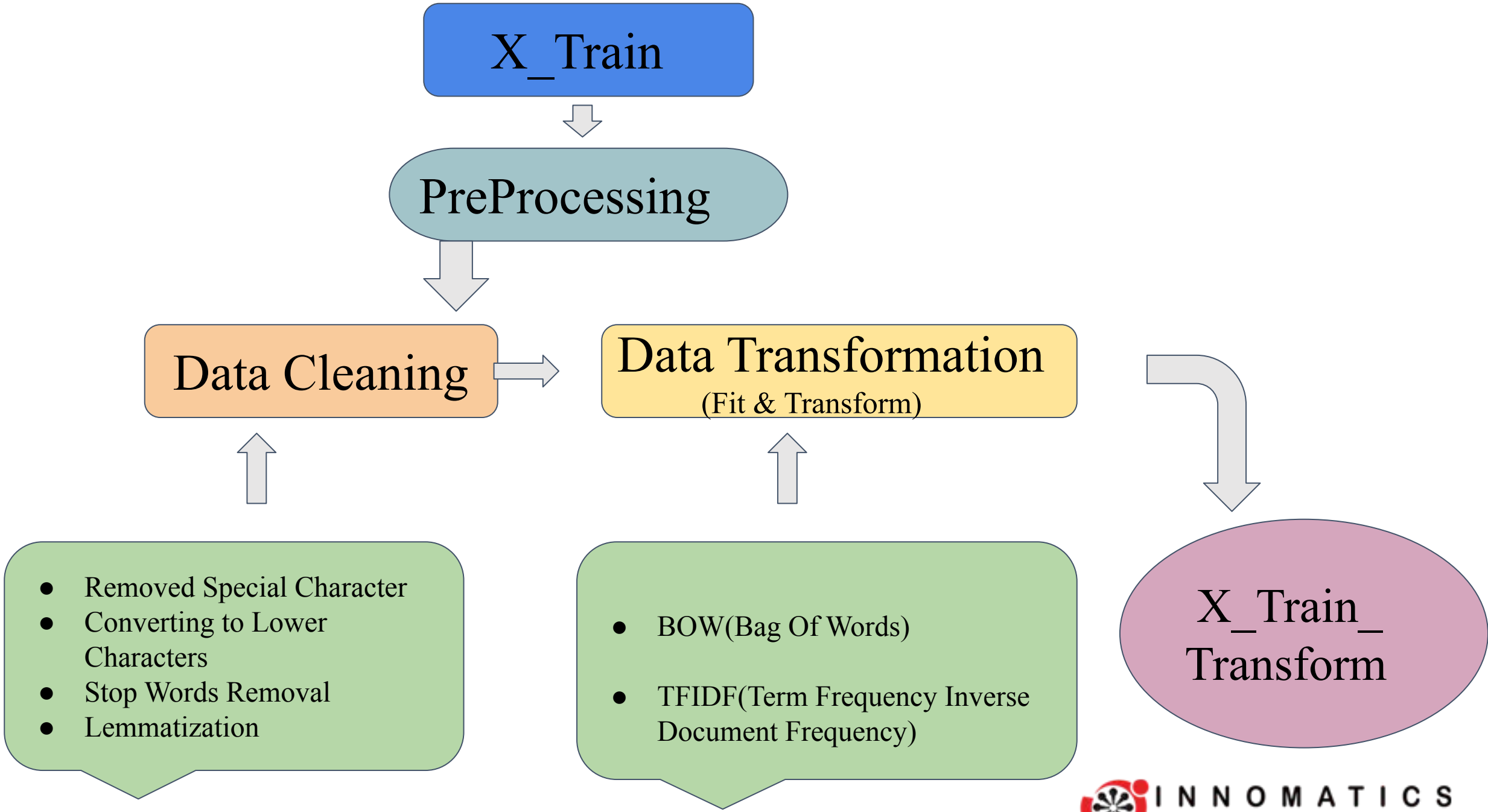


- For sentiment prediction
  We have
  394360 training datapoints
  131454 test datapoints

- For Score prediction
  We have
  426340 training datapoints
  142114 test datapoints.

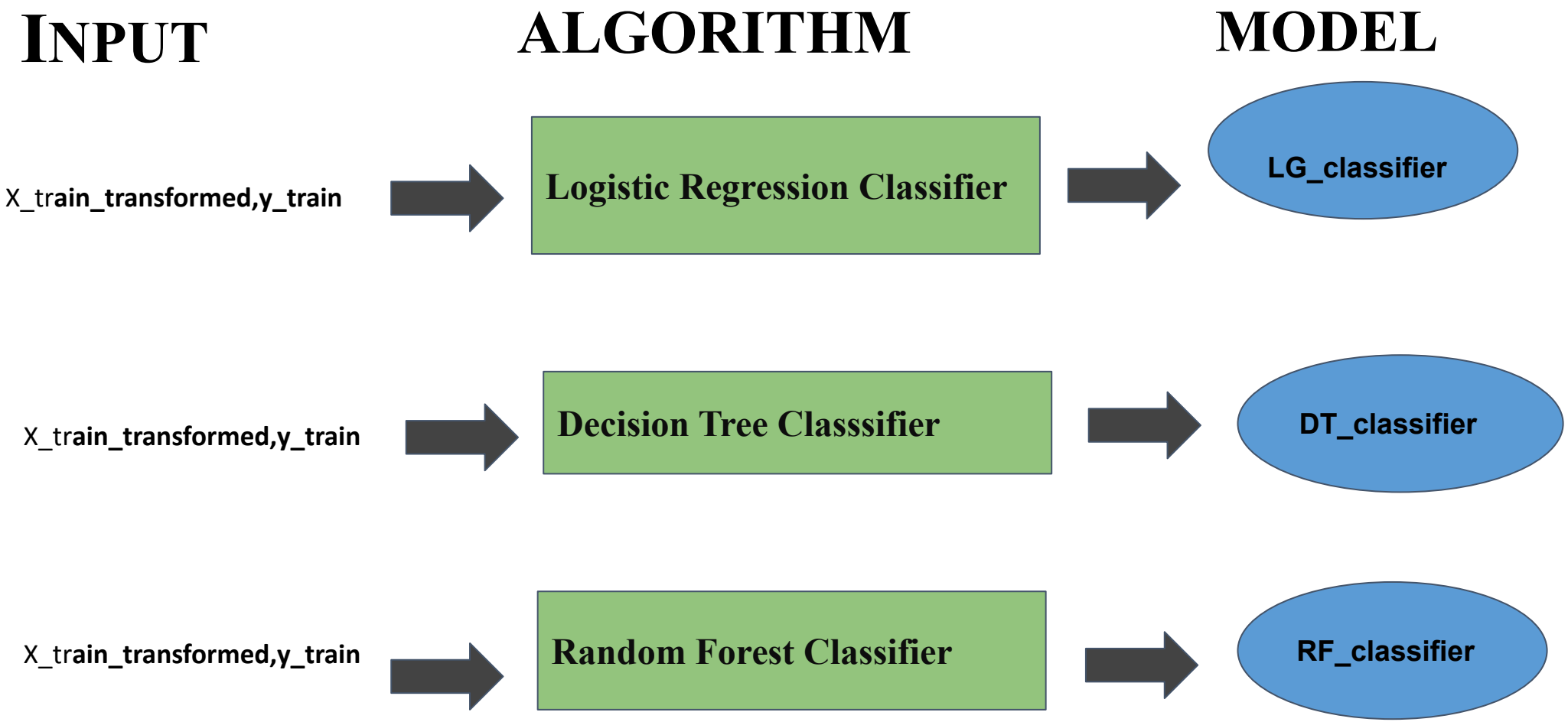
- We have used 75:25 train:test split.

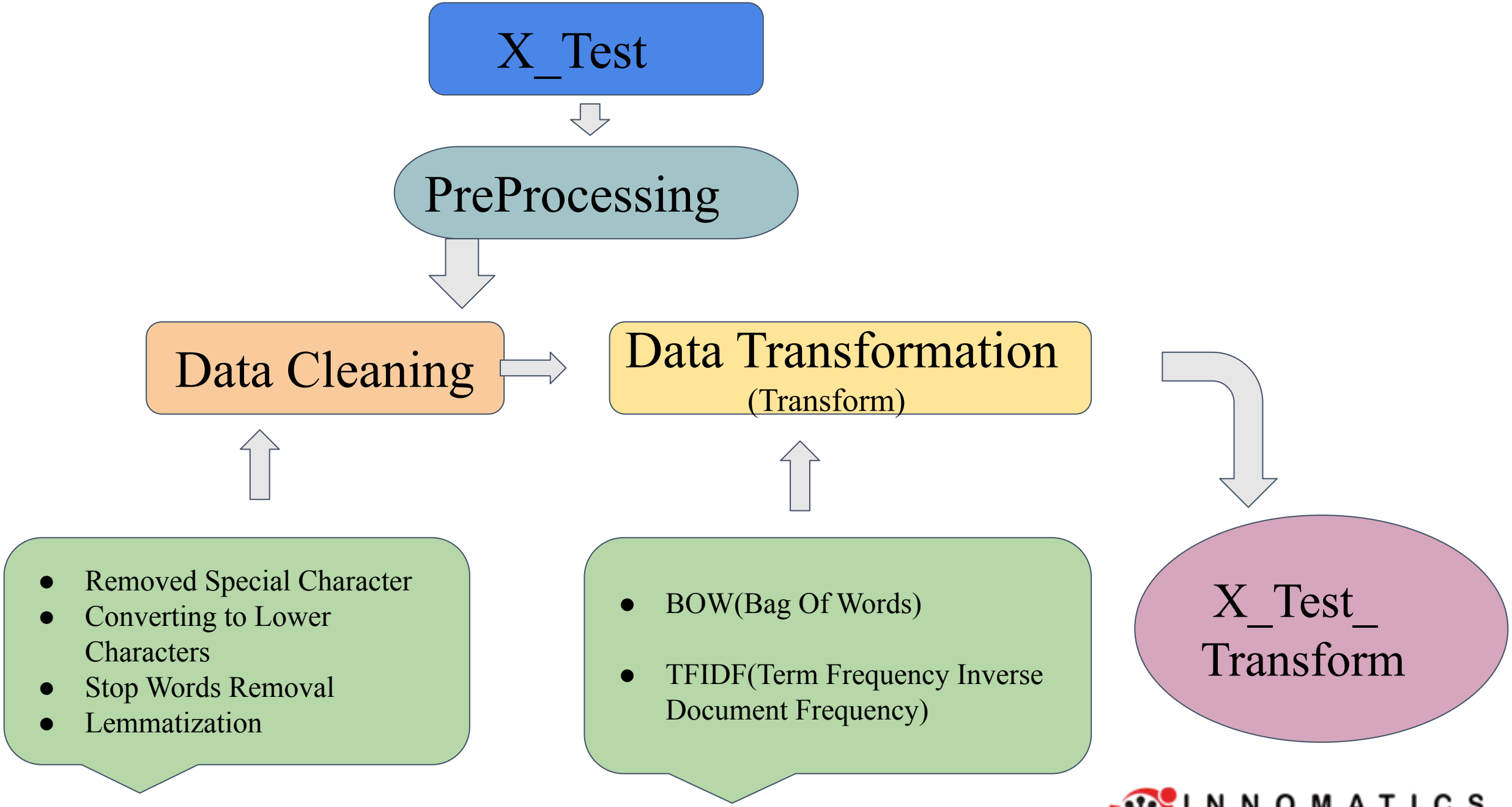
INNOMATICS RESEARCH LABS



INNOMATICS RESEARCH LABS

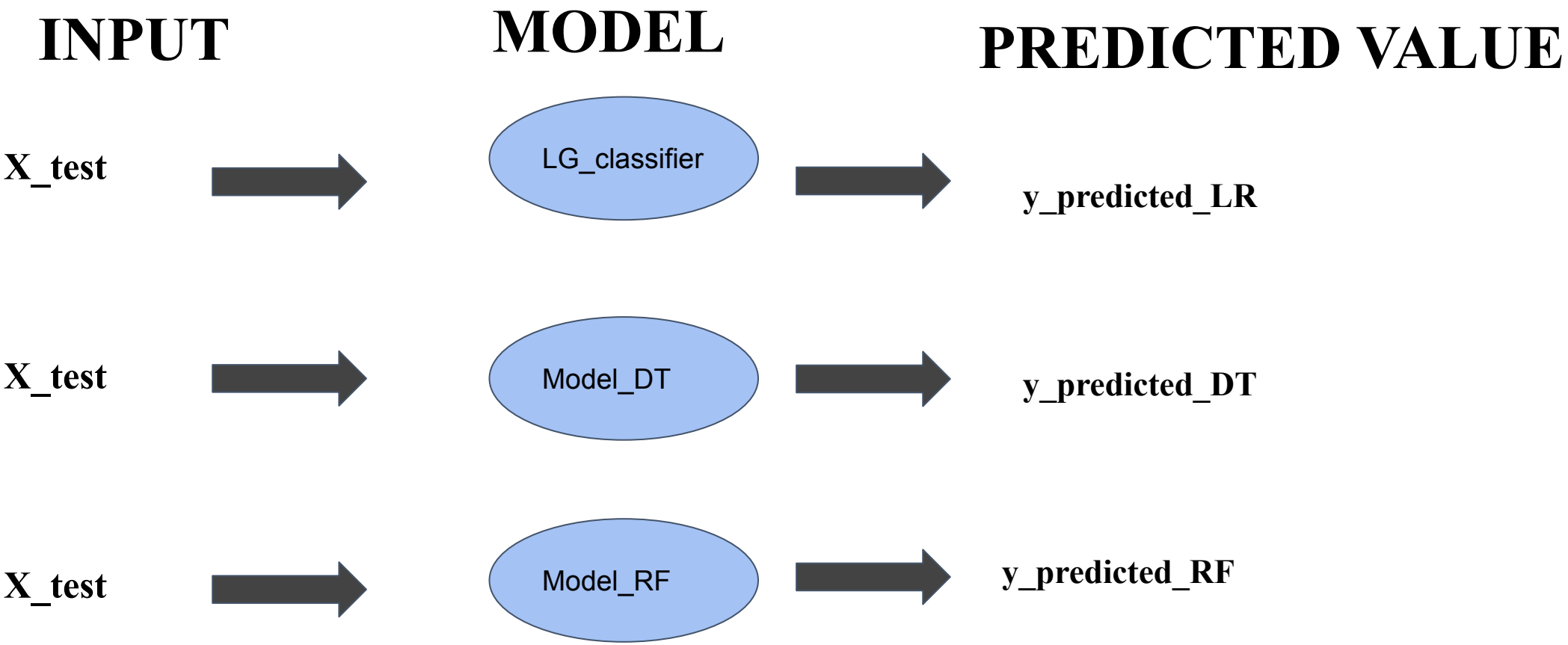
Model Building:(By Using Classification Algorithm)



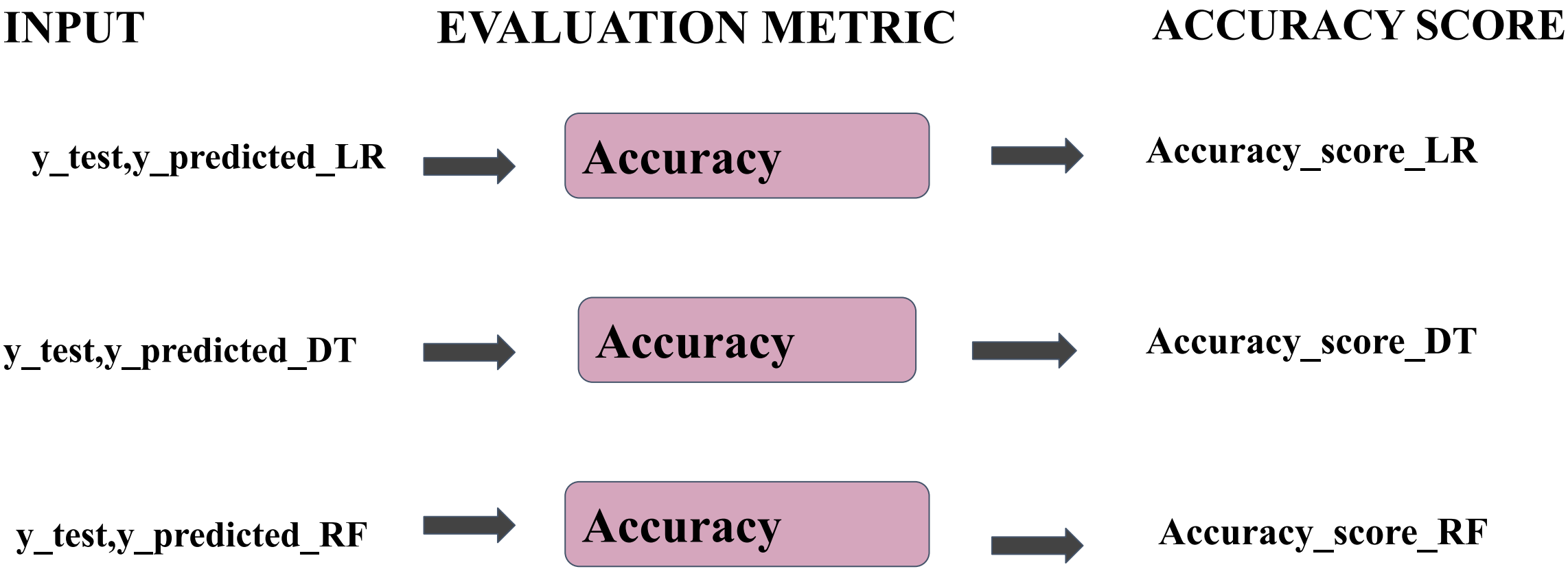


INNOMATICS RESEARCH LABS

Prediction using the Model:

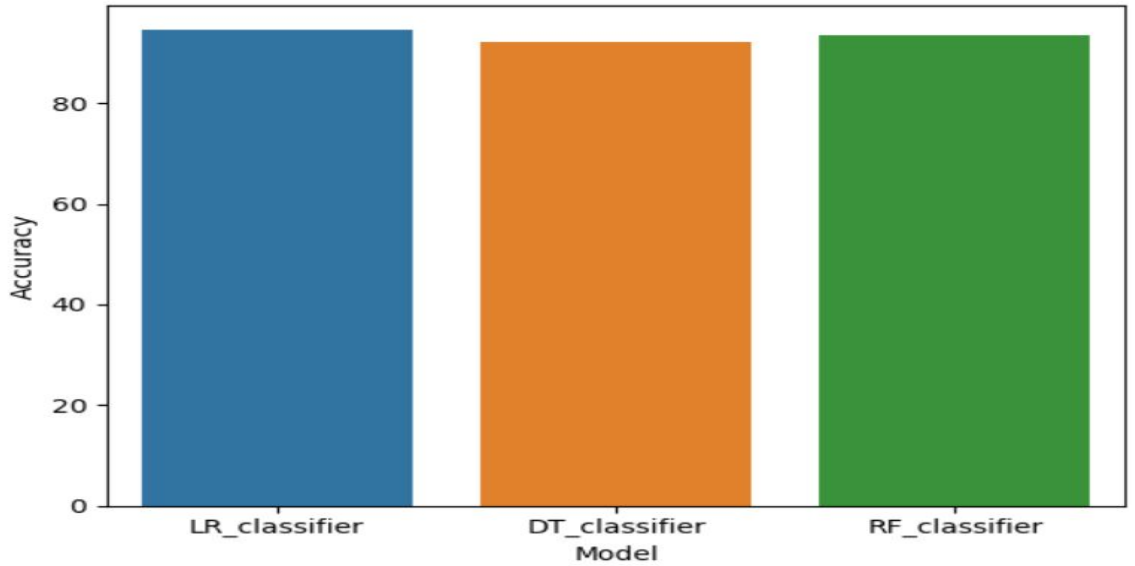


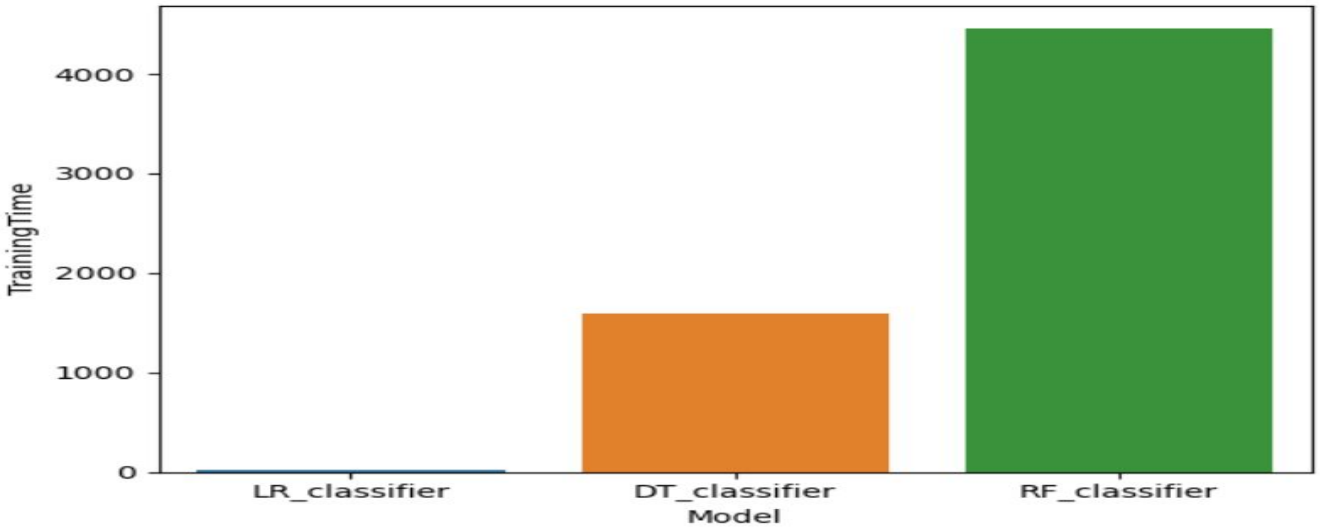
Evaluation Of The Models

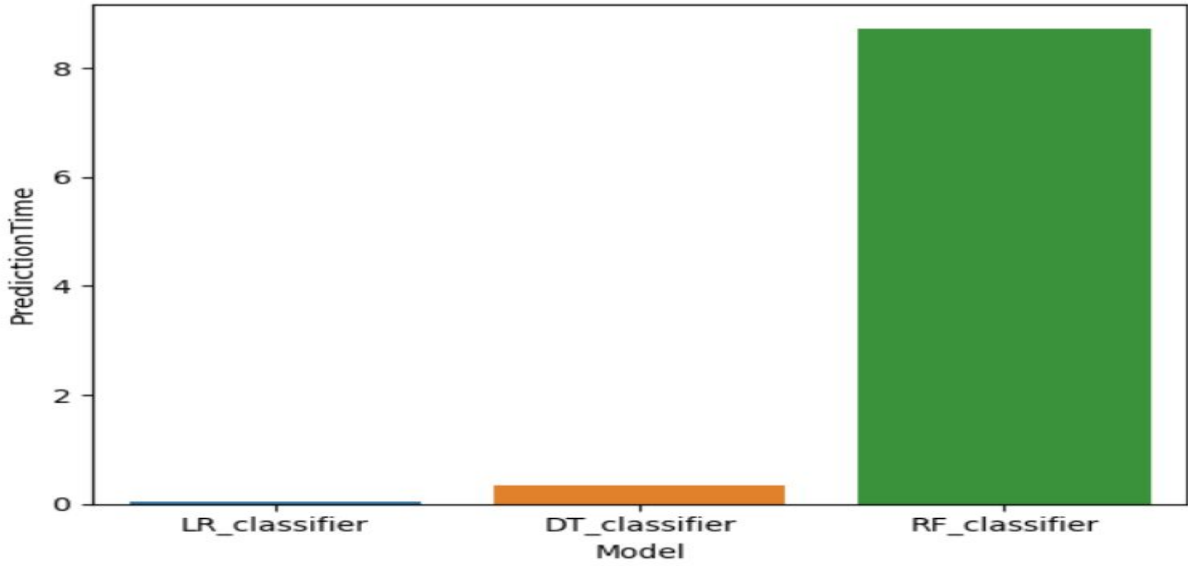


Models Performance for **Sentiment** Prediction (**BOW**)

	Model	Accuracy	TrainingTime	PredictionTime	size(kb)
0	LR_classifier	94.736562	15.4580	0.03193	715
1	DT_classifier	92.295404	1591.9756	0.32884	2956
2	RF_classifier	93.475284	4461.4506	8.73942	985275



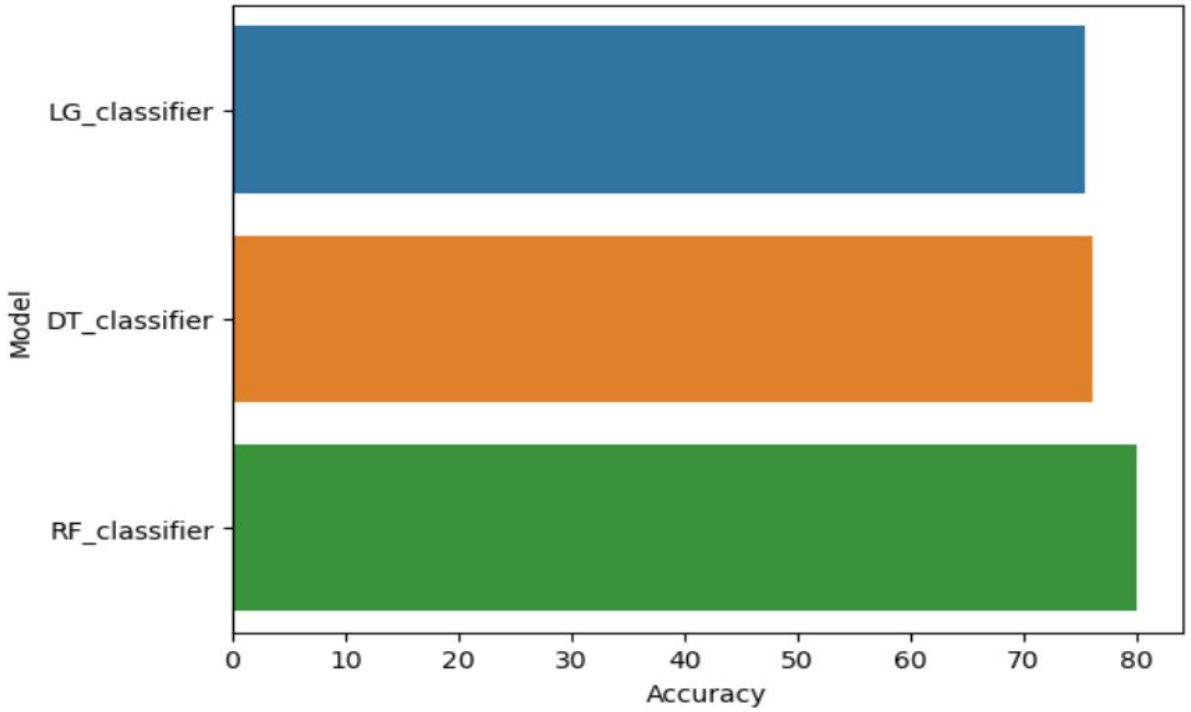


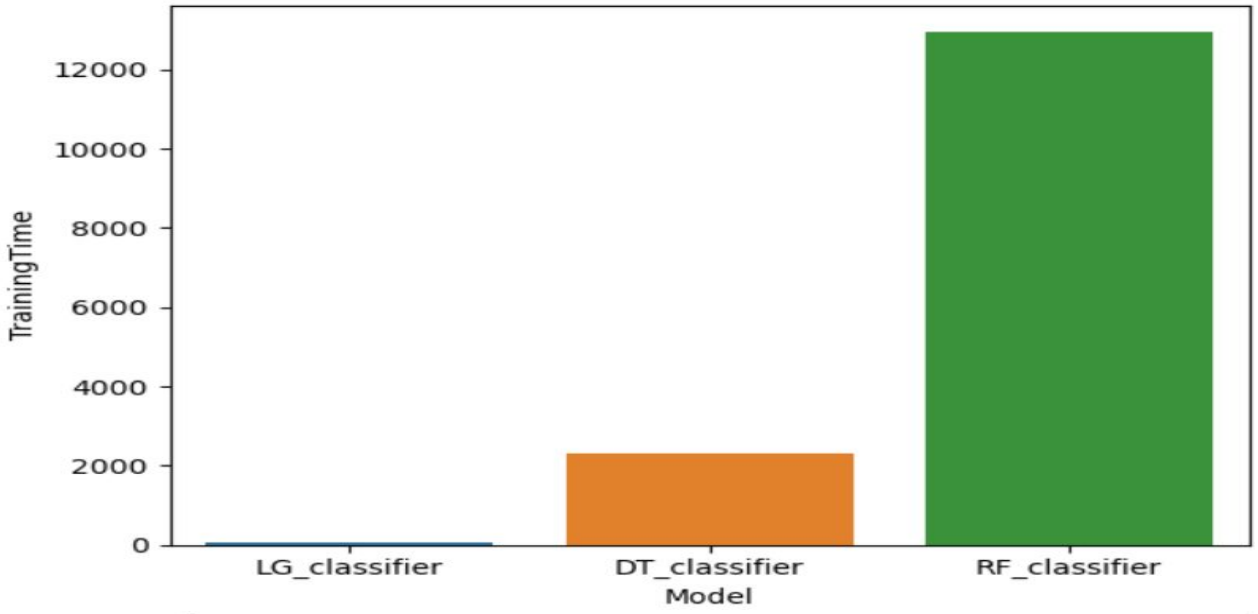


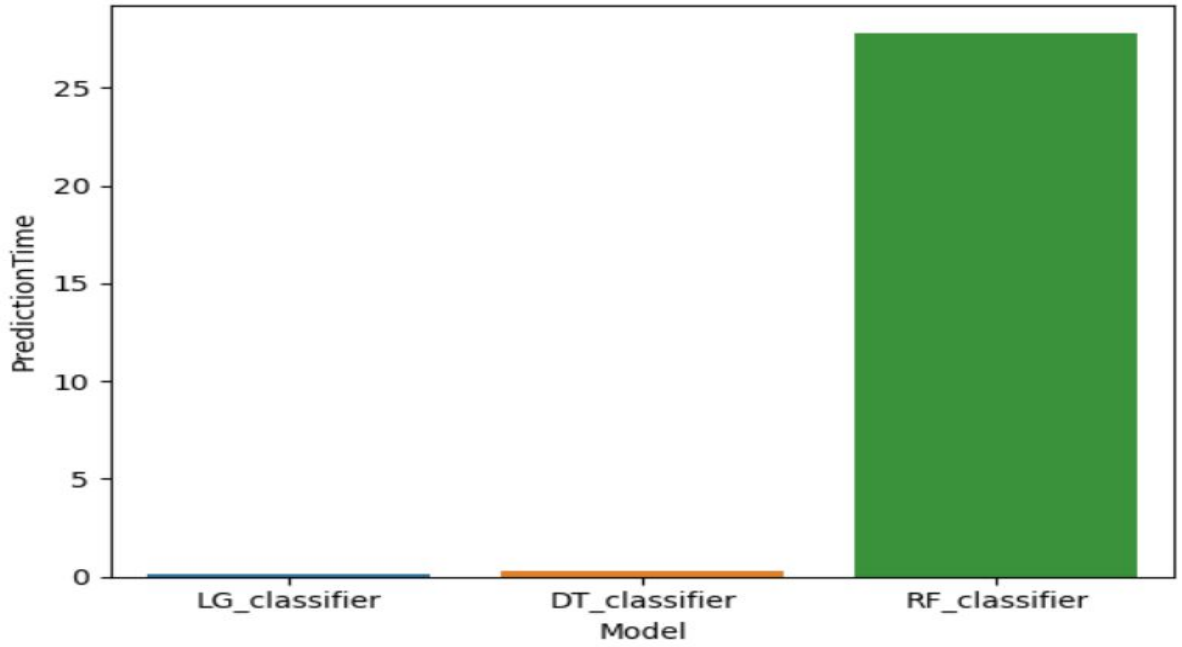
INNOMATICS RESEARCH LABS

Model predictions for **score (BOW)**

	Model	Accuracy	TrainingTime	PredictionTime	Size(kb)
0	LG_classifier	75.492038	60.45	0.1033	3738
1	DT_classifier	76.106338	2289.90	0.2770	13740
2	RF_classifier	80.112305	12970.54	27.8240	2647556



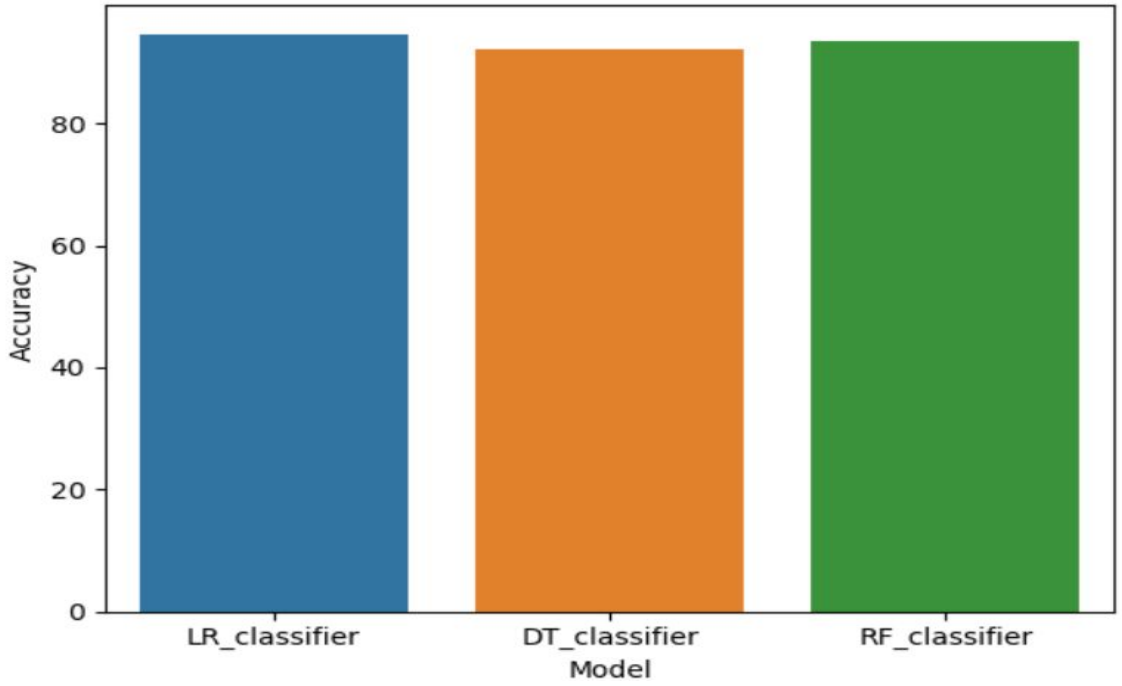


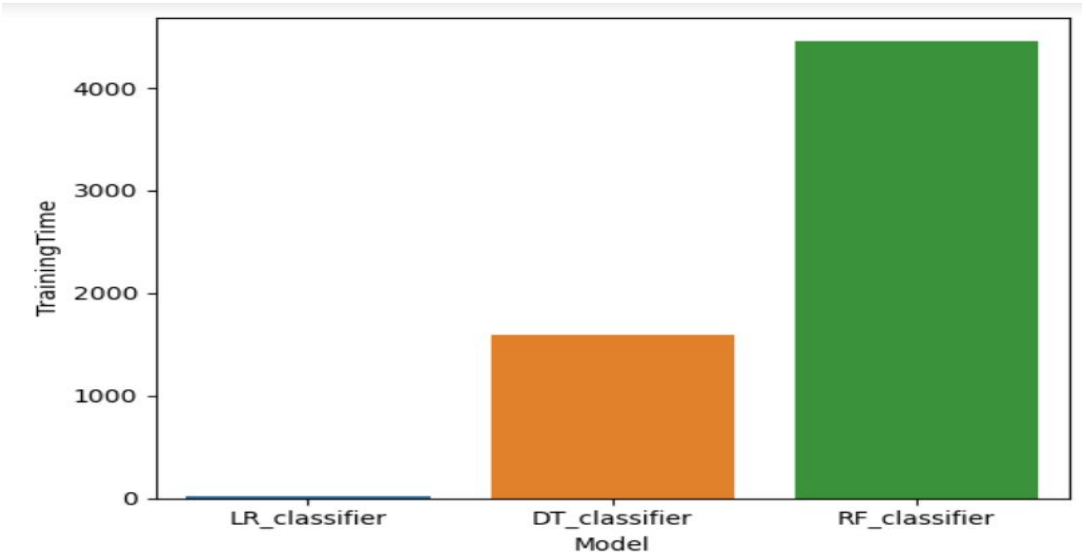


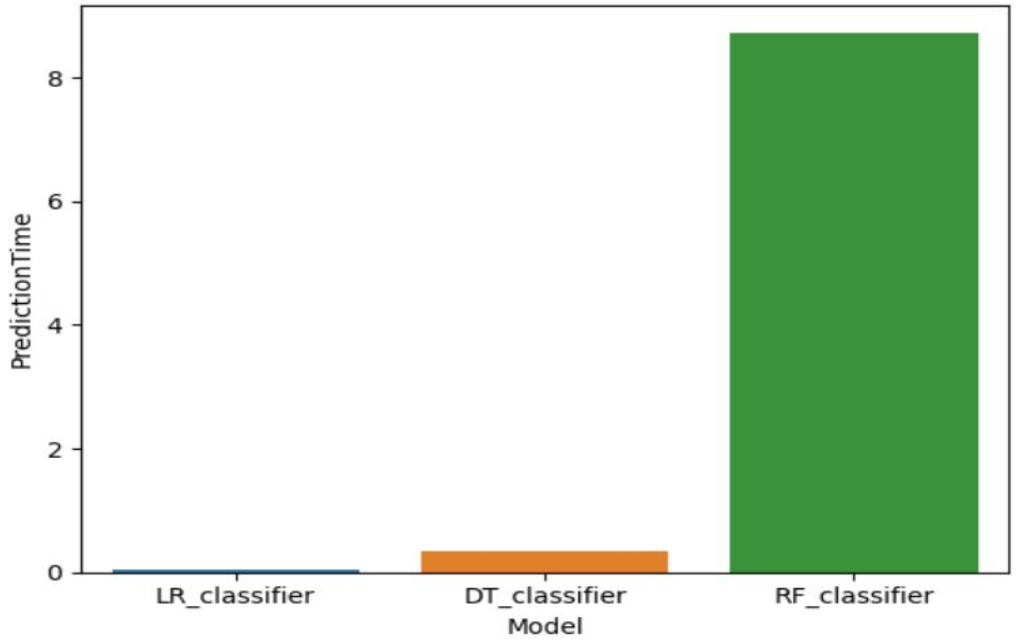
INNOMATICS RESEARCH LABS

Models Performance for **Sentiment** Prediction (**TFIDF**)

	Model	Accuracy	TrainingTime	PredictionTime	size(kb)
0	LR_classifier	94.658968	14.2476	0.02633	715
1	DT_classifier	91.855706	2086.1053	0.38093	2439
2	RF_classifier	93.393126	2630.5660	9.33100	789935



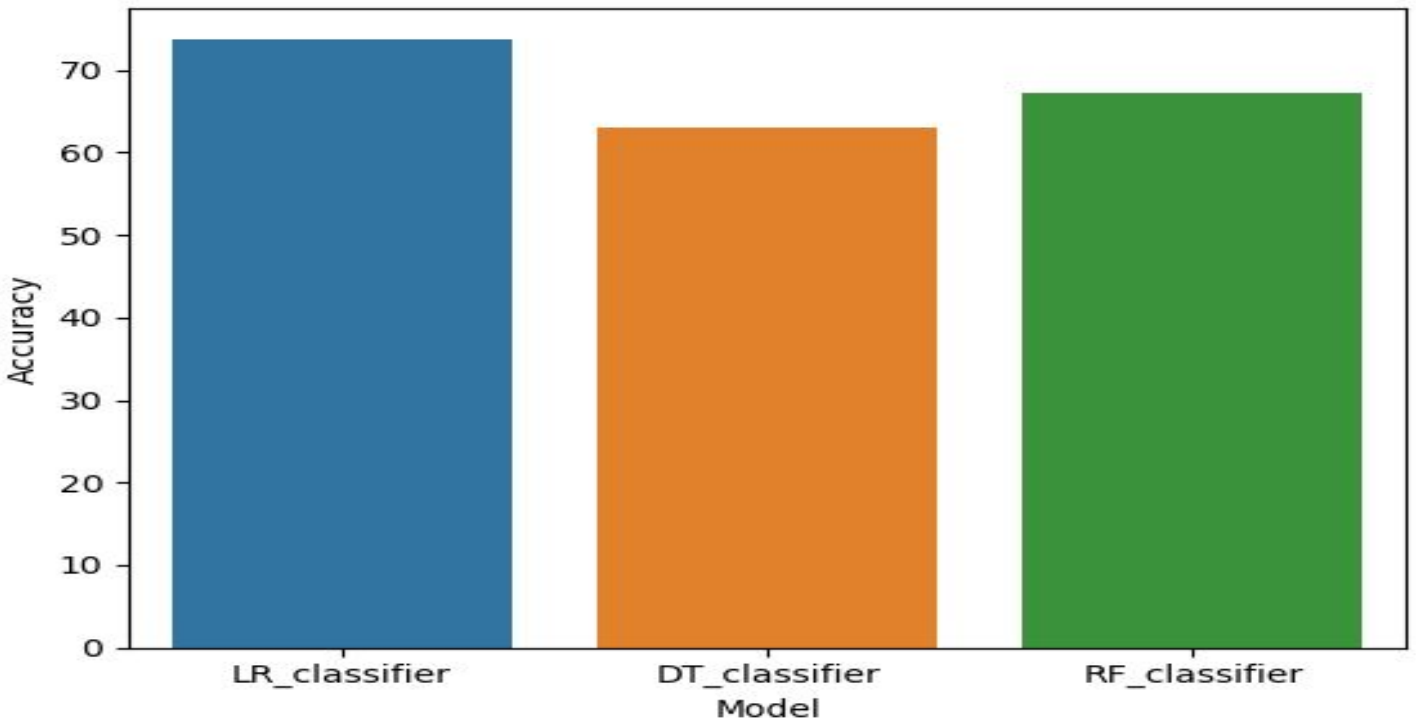


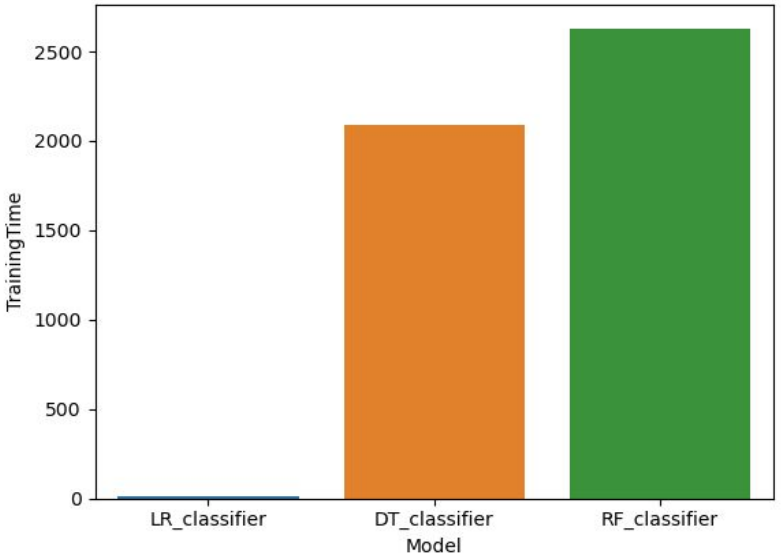


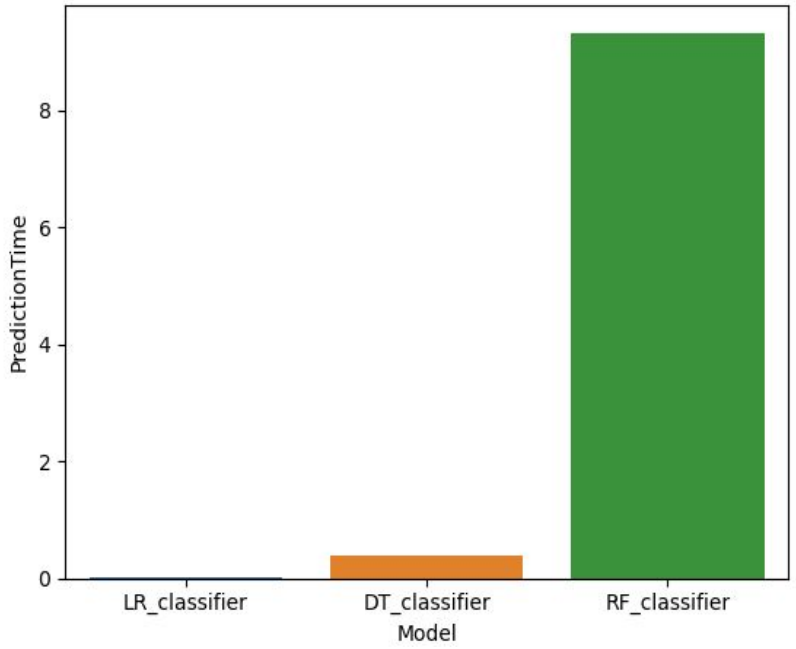
INFORMATICS RESEARCH LABS

Models Performance for **Score** Prediction (**TFIDF**)

	Model	Accuracy	TrainingTime	PredictionTime	size(kb)
0	LR_classifier	73.812573	14.2476	0.02633	4337
1	DT_classifier	63.065567	2086.1053	0.38093	10938
2	RF_classifier	67.154538	2630.5660	9.33100	2229249







INNOMATICS RESEARCH LABS

Conclusion

- For sentiment analysis the logistic regression classifier model was giving the highest accuracy around 95%. On top of that it the model which takes the least amount of time for prediction and occupies lthe least memory. So we deployed this model for prediction of sentiment.
- For the score prediction we saw that the Ranfom Forest classifier model was giving the highest accuracy of about 80% while the Logistic Regression model gave the accuracy of around 75% .We have deployed the Logistic Regression Classifier Model for the score prediction because of its prompt prediction and lightness since Random Forest model takes 2.73 GB of memory space as compared to the Logistic Regression classifier which takes only 3.83MB

Q & A

Thank you!

INNOMATICS RESEARCH LABS