Text Classification

Abstract

Text classification is one of the major tasks of AI in general, NLP in particular. Having five Gutenberg books, this report discusses the methodologies and models with different transformation techniques that have been applied to reach the best accuracy that the champion model achieves by correctly classifying unseen text to the corresponding book.

Business Case

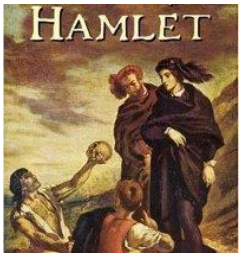
Since it's challenging to predict the name of a British author based on their writing style, due to the many styles and genres which can greatly similar within British literature, by training a model on these styles and themes, it can make educated guesses about the authorship of a work based on its content and style.

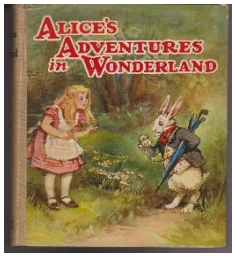
So, we started our project by choosing five books by different authors belongs to the same category (history) and applied the following steps

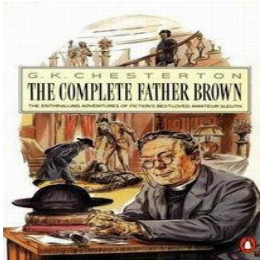
Dataset

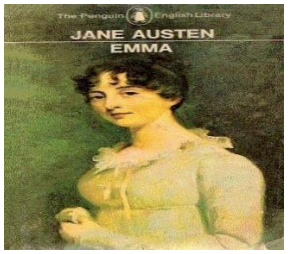
The Gutenberg dataset represents a corpus of over 60,000 book texts, their authors and titles. The data has been scraped from the Project Gutenberg website using a custom script to parse all bookshelves. we have taken five different samples of Gutenberg digital books that are of five different authors, that we think are of the history same genre and are semantically the same.

Books









Data Preparation, Preprocessing and, Cleaning:

- Listing all the books in Gutenberg's library
- Choose five different books by five different authors belong to the same category (History)
- Data preparation:
    I. Removing stop words
    II. Converting all words to the lower case
    III. Tokenize the text
    IV. Lemmatization is the next step that reduces a word to its base form
- Data Partitioning: partition each book into 200 documents, each document is a 100 word record

- Data labeling as follows:
    I. austen-emma→ a
    II. chesterton-thursday→ b
    III. shakespeare-hamlet→ c
    IV. chesterton-ball→ d
    V. carroll-alice→ e

	Partitions	Book Name	Book Label	Book Author
0	applied pile whitebait gravest sort enjoyment ...	Brown.txt	a	G. K. Chesterton
1	condition gravel grass major strolled unobtrus...	Brown.txt	a	G. K. Chesterton
2	hawker hears love insulted strike till got sof...	Brown.txt	a	G. K. Chesterton
3	really got cut fence somehow spoiling plant on...	Brown.txt	a	G. K. Chesterton
4	admiral let talk anything enough say whenever ...	Brown.txt	a	G. K. Chesterton
...	...	...	...	...
995	top bow arrow suddenly music heard crowd silen...	Parents.txt	e	Maria Edgeworth
996	manner attention caught shrill sound scolding ...	Parents.txt	e	Maria Edgeworth
997	everything look cheerful usual felt still mise...	Parents.txt	e	Maria Edgeworth
998	word luck bent bow string broke two bow fell h...	Parents.txt	e	Maria Edgeworth
999	put head strangest oddest fancy consequence bu...	Parents.txt	e	Maria Edgeworth

1000 rows × 4 columns

Feature Engineering:

- Encoding

-Text transformation using BOW, TF-IDF, N-gram

- **BOW:** It represents the occurrence of words within a document, it involves two things:
    - A vocabulary of known words.
    - A measure of the presence of known words.

	abated	abatement	abating	abbey	abbot	abhor	abhorred	abhorring	abide	ability	...	youngest	youth	zeal	zealand	zealous	zigzag	zodiac	zone	zooks	zoroaster
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
996	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
997	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
998	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
999	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1000 rows × 12415 columns

- **TF-IDF:** a technique to quantify words in a set of documents. We compute a score for each word to signify its importance in the document and corpus.

	abated	abatement	abating	abbey	abbot	abhor	abhorred	abhorring	abide	ability	...	youngest	youth	zeal	zealand	zealous	zigzag	zodiac	zone	zooks	zoroaster
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1000 rows × 12415 columns

- **N-gram**

	abated speed	abated violence	abatement necessary	abating flurry	abating speed	abbey bec	abbey chapter	abbey curious	abbey fish	abbey formed	...	zealand right	zealous activity	zigzag going	zodiac deadly	zone grub	zone sum	zone surface	zone zone	zooks argufy	zoroaster died
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
996	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
997	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
998	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
999	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1000 rows × 90538 columns

	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93	94	95	96	97	98	99
0	-0.348371	0.326178	0.006444	0.262098	0.048852	-0.696415	0.203450	0.717862	-0.371506	-0.124657	...	0.437340	0.096238	0.128555	0.084469	0.561675	0.223830	0.111275	-0.414582	0.001379	-0.144002
1	-0.338416	0.313164	0.007488	0.254864	0.057260	-0.661923	0.194433	0.689337	-0.361836	-0.109520	...	0.420784	0.084961	0.119660	0.082081	0.538196	0.210460	0.100850	-0.400739	0.001405	-0.137687
2	-0.392249	0.365394	0.009648	0.294911	0.055493	-0.777858	0.227235	0.804982	-0.420055	-0.137058	...	0.491922	0.105781	0.141921	0.093749	0.629756	0.252223	0.124004	-0.465505	0.001606	-0.161290
3	-0.334916	0.311954	0.009306	0.254428	0.059161	-0.654177	0.193049	0.686311	-0.359835	-0.105494	...	0.418775	0.081670	0.116946	0.082716	0.533545	0.204885	0.097108	-0.398415	0.000435	-0.133580
4	-0.348971	0.325244	0.010132	0.262504	0.049769	-0.692729	0.202746	0.716756	-0.371423	-0.120495	...	0.437658	0.092787	0.127722	0.084701	0.560120	0.223546	0.108703	-0.413971	0.002566	-0.143849
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	-0.413564	0.385488	0.011181	0.313044	0.063654	-0.816571	0.243643	0.851557	-0.447614	-0.136791	...	0.519136	0.107665	0.148760	0.102563	0.664842	0.260275	0.122940	-0.493227	0.003003	-0.168489
996	-0.361694	0.339031	0.012096	0.274095	0.053436	-0.721602	0.212108	0.750786	-0.389331	-0.123838	...	0.456220	0.094778	0.130786	0.088776	0.586264	0.230114	0.112589	-0.432724	0.003183	-0.148158
997	-0.378176	0.351748	0.010986	0.287396	0.058281	-0.748632	0.222798	0.782595	-0.407355	-0.126168	...	0.475775	0.097030	0.137380	0.092147	0.607785	0.238801	0.112108	-0.452931	0.002932	-0.155075
998	-0.373131	0.349757	0.014627	0.286373	0.051756	-0.745582	0.223637	0.778618	-0.402109	-0.128373	...	0.472510	0.101825	0.138065	0.093920	0.607127	0.238799	0.114435	-0.447612	0.003918	-0.155691
999	-0.381755	0.357342	0.013903	0.285736	0.056172	-0.764032	0.224080	0.796462	-0.408899	-0.133586	...	0.482717	0.101563	0.139405	0.094980	0.621441	0.245578	0.122525	-0.457950	0.004416	-0.155539

1000 rows × 100 columns

Modelling:

For each technique of the above, these following models are trained and tested.

1. SVM
2. Random Forest
3. Gaussian Naive Bayes
4. Bernoulli Naive Bayes
5. K Nearest Neighbors

6. XGB Extreme X Gradient Boosting
7. Stochastic Gradient Descent SGD
8. Logistic Regression
9. Decision Tree Classifier
10. AdaBoost

And here is the accuracy resulting for each model with different transformations

1. BOW

	Model	Accuracy_Score
0	Gaussian Naive Bayes	100.00
1	Bernoulli Naive Bayes	100.00
2	Support Vector Machine SVM	98.67
3	Random Forest	98.67
4	XGB Extreme X Gradient Boosting	98.67
5	Stochastic Gradient Descent SGD	97.33
6	Logistic Regression	97.33
7	Catboost	96.00
8	K Nearest Neighbors	92.00
9	Decision Tree Classifier	92.00
10	AdaBoost	32.00

Result samples of BOW & three types of modeling

Random Forest

```
Cross_validation Accuracy for Random Forest :
 [0.98823529 0.98823529 1.         1.         1.         0.98823529
 1.         0.97647059 0.97647059 0.98823529]
Random Forest


Confusion Matrix :
[[28  0  0  1  0]
 [ 0 27  0  0  0]
 [ 0  0 34  0  0]
 [ 1  0  0 25  0]
 [ 0  0  0  0 34]]

Classification Report :
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        29
           1       1.00      1.00      1.00        27
           2       1.00      1.00      1.00        34
           3       0.96      0.96      0.96        26
           4       1.00      1.00      1.00        34

    accuracy                           0.99       150
   macro avg       0.99      0.99      0.99       150
weighted avg       0.99      0.99      0.99       150


Testing Accuracy of Random Forest is 98.67 %
Average bias: 0.000
Average variance: 1.994
```

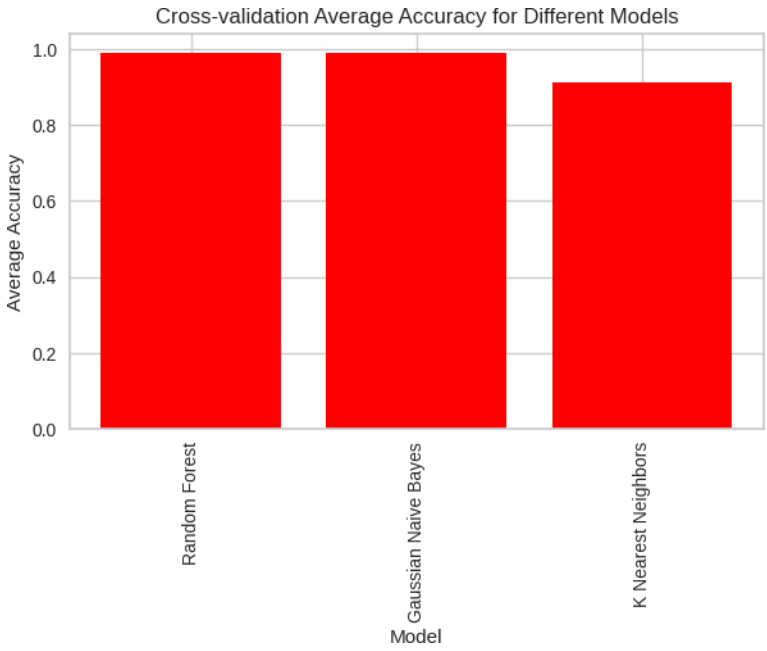
Naïve Bayes

```
Cross_validation Accuracy for Gaussian Naive Bayes :
 [0.97647059 1.         1.         1.         0.97647059 0.96470588
 1.         0.98823529 1.         1.        ]
Gaussian Naive Bayes


Confusion Matrix :
[[29  0  0  0  0]
 [ 0 27  0  0  0]
 [ 0  0 34  0  0]
 [ 0  0  0 26  0]
 [ 0  0  0  0 34]]

Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        29
           1       1.00      1.00      1.00        27
           2       1.00      1.00      1.00        34
           3       1.00      1.00      1.00        26
           4       1.00      1.00      1.00        34

    accuracy                           1.00       150
   macro avg       1.00      1.00      1.00       150
weighted avg       1.00      1.00      1.00       150


Testing Accuracy of Gaussian Naive Bayes is 100.0 %
Average bias: 0.000
Average variance: 1.994
```

K-Neatest Neighbours

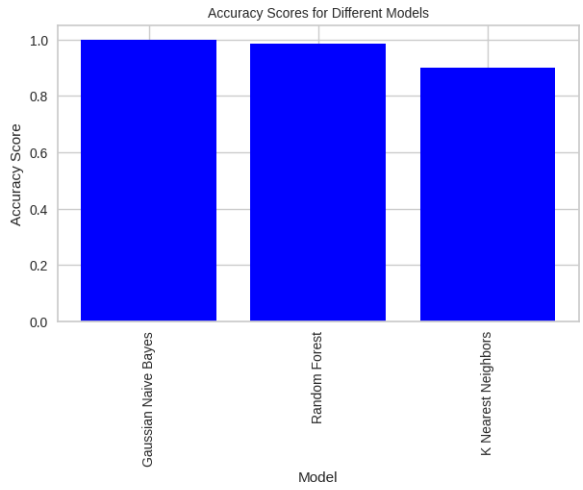
```
Cross_validation Accuracy for K Nearest Neighbors :
 [0.90588235 0.90588235 0.91764706 0.89411765 0.92941176 0.88235294
 0.90588235 0.91764706 0.92941176 0.92941176]
K Nearest Neighbors


Confusion Matrix :
[[28  1  0  0  0]
 [ 0 27  0  0  0]
 [ 0  0 34  0  0]
 [ 3  6  4 13  0]
 [ 0  1  0  0 33]]

Classification Report :
              precision    recall  f1-score   support

           0       0.90      0.97      0.93        29
           1       0.77      1.00      0.87        27
           2       0.89      1.00      0.94        34
           3       1.00      0.50      0.67        26
           4       1.00      0.97      0.99        34

    accuracy                           0.90       150
   macro avg       0.91      0.89      0.88       150
weighted avg       0.92      0.90      0.89       150


Testing Accuracy of K Nearest Neighbors is 90.0 %
Average bias: 0.176
Average variance: 2.028
```

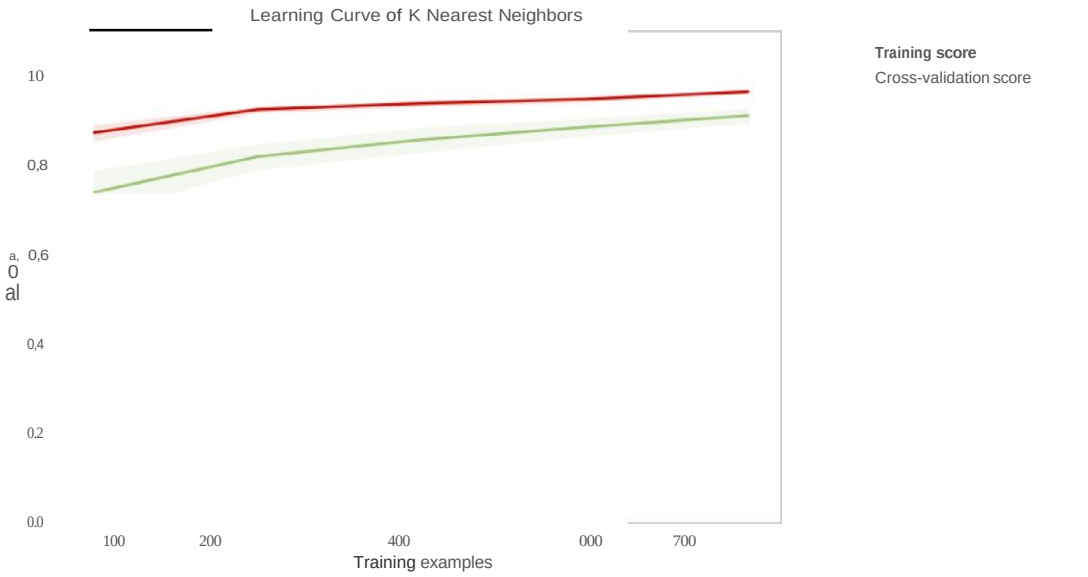
Cross-Validation for the classifications models

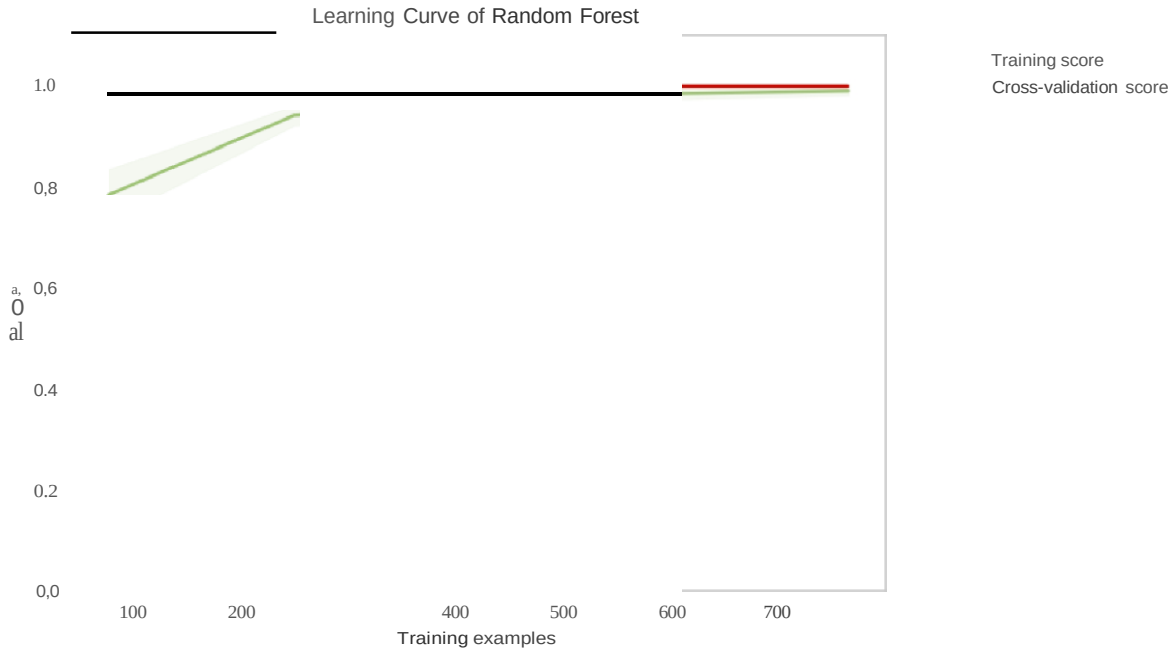


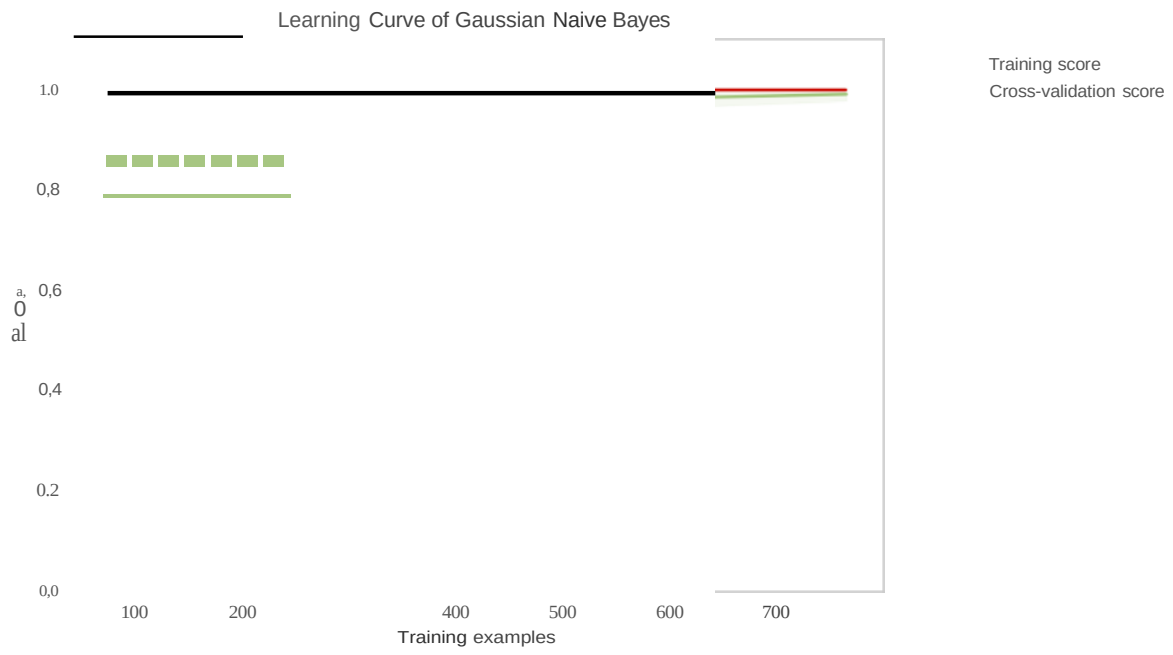
Accuracy Scores for Models

	Model	Accuracy_Score
0	Gaussian Naive Bayes	1.000000
1	Random Forest	0.986667
2	K Nearest Neighbors	0.900000




Learning Curve of K Nearest Neighbors
Training score
Cross-validation score
10
0,8
0,6
0,4
0.2
0.0
100
200
400
000
700
Training examples


Learning Curve of Random Forest
Training score
Cross-validation score
1.0
0,8
0,6
0.4
0.2
0,0
100
200
400
500
600
700
Training examples


Learning Curve of Gaussian Naive Bayes
Training score
Cross-validation score
1.0
0,8
0,6
0,4
0.2
0,0
100
200
400
500
600
700
Training examples

2. TF-IDF

	Model	Accuracy_Score
0	Support Vector Machine SVM	100.00
1	Gaussian Naive Bayes	100.00
2	Bernoulli Naive Bayes	100.00
3	Stochastic Gradient Descent SGD	100.00
4	Logistic Regression	100.00
5	Random Forest	98.67
6	XGB Extreme X Gradient Boosting	98.67
7	Catboost	98.67
8	K Nearest Neighbors	97.33
9	Decision Tree Classifier	93.33
10	AdaBoost	32.00

Result samples of TF-IDF & three types of modeling

Random Forest

```
Cross_validation Accuracy for Random Forest :
 [0.94117647 1.         1.         0.97647059 0.97647059 0.96470588
 0.98823529 0.98823529 0.95294118 1.        ]
Random Forest


Confusion Matrix :
[[27  0  0  2  0]
 [ 0 27  0  0  0]
 [ 0  0 34  0  0]
 [ 1  1  1 23  0]
 [ 0  0  0  0 34]]

Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.93      0.95        29
           1       0.96      1.00      0.98        27
           2       0.97      1.00      0.99        34
           3       0.92      0.88      0.90        26
           4       1.00      1.00      1.00        34

    accuracy                           0.97       150
   macro avg       0.96      0.96      0.96       150
weighted avg       0.97      0.97      0.97       150


Testing Accuracy of Random Forest is 96.67 %
Average bias: 0.000
Average variance: 1.994
```

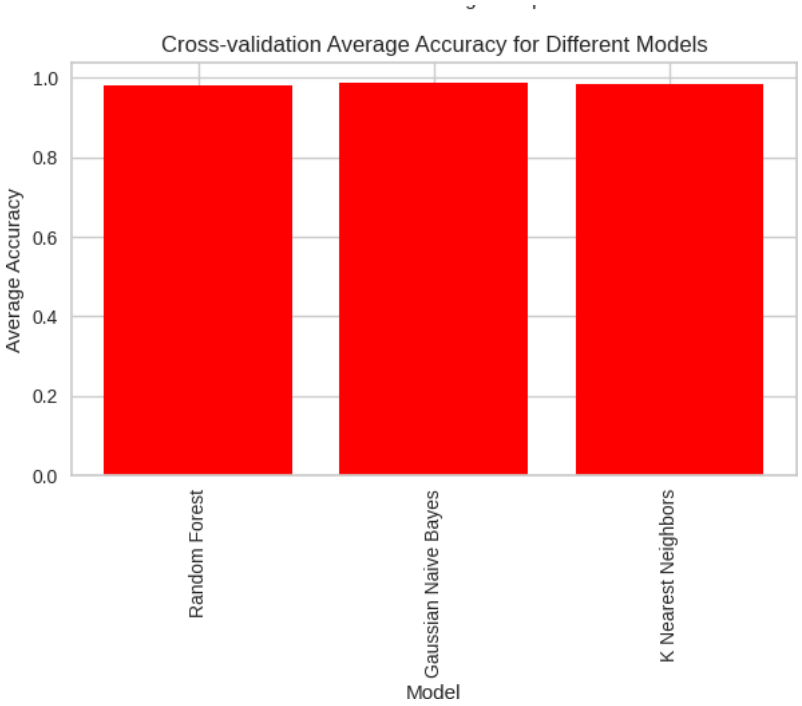
Naïve Bayes

```
Cross_validation Accuracy for Gaussian Naive Bayes :
 [0.97647059 1.         1.         1.         0.96470588 0.96470588
 1.         0.97647059 1.         1.        ]
Gaussian Naive Bayes


Confusion Matrix :
[[29  0  0  0  0]
 [ 0 25  0  2  0]
 [ 0  0 34  0  0]
 [ 0  0  0 26  0]
 [ 0  0  0  0 34]]

Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        29
           1       1.00      0.93      0.96        27
           2       1.00      1.00      1.00        34
           3       0.93      1.00      0.96        26
           4       1.00      1.00      1.00        34

    accuracy                           0.99       150
   macro avg       0.99      0.99      0.98       150
weighted avg       0.99      0.99      0.99       150


Testing Accuracy of Gaussian Naive Bayes is 98.67 %
Average bias: 0.000
Average variance: 1.994
```

K-Nearest Negibour

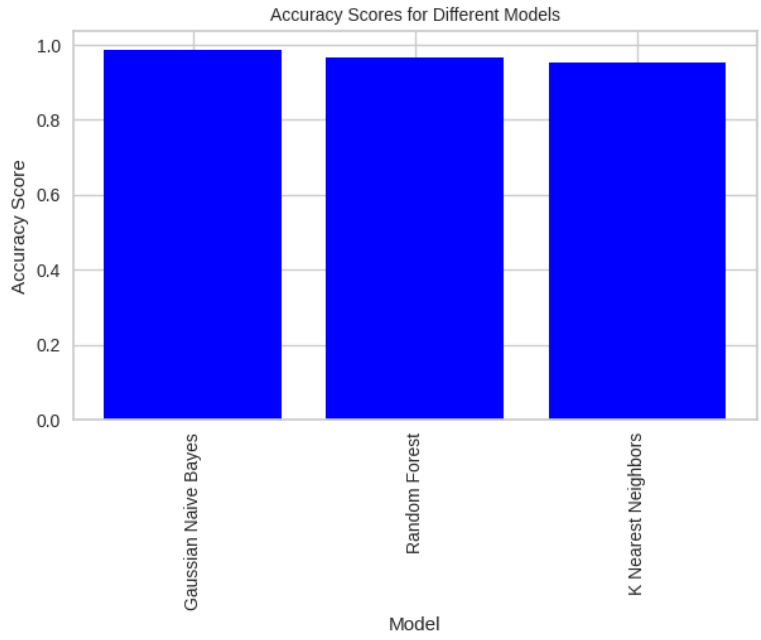
```
Cross_validation Accuracy for K Nearest Neighbors :
 [0.98823529 0.96470588 0.98823529 0.98823529 0.98823529 0.97647059
 0.98823529 0.98823529 0.97647059 0.97647059]
K Nearest Neighbors


Confusion Matrix :
[[28  0  0  1  0]
 [ 2 24  1  0  0]
 [ 0  0 34  0  0]
 [ 0  2  1 23  0]
 [ 0  0  0  0 34]]

Classification Report :
              precision    recall  f1-score   support

           0       0.93      0.97      0.95        29
           1       0.92      0.89      0.91        27
           2       0.94      1.00      0.97        34
           3       0.96      0.88      0.92        26
           4       1.00      1.00      1.00        34

    accuracy                           0.95       150
   macro avg       0.95      0.95      0.95       150
weighted avg       0.95      0.95      0.95       150


Testing Accuracy of K Nearest Neighbors is 95.33 %
Average bias: 0.006
Average variance: 1.993
```

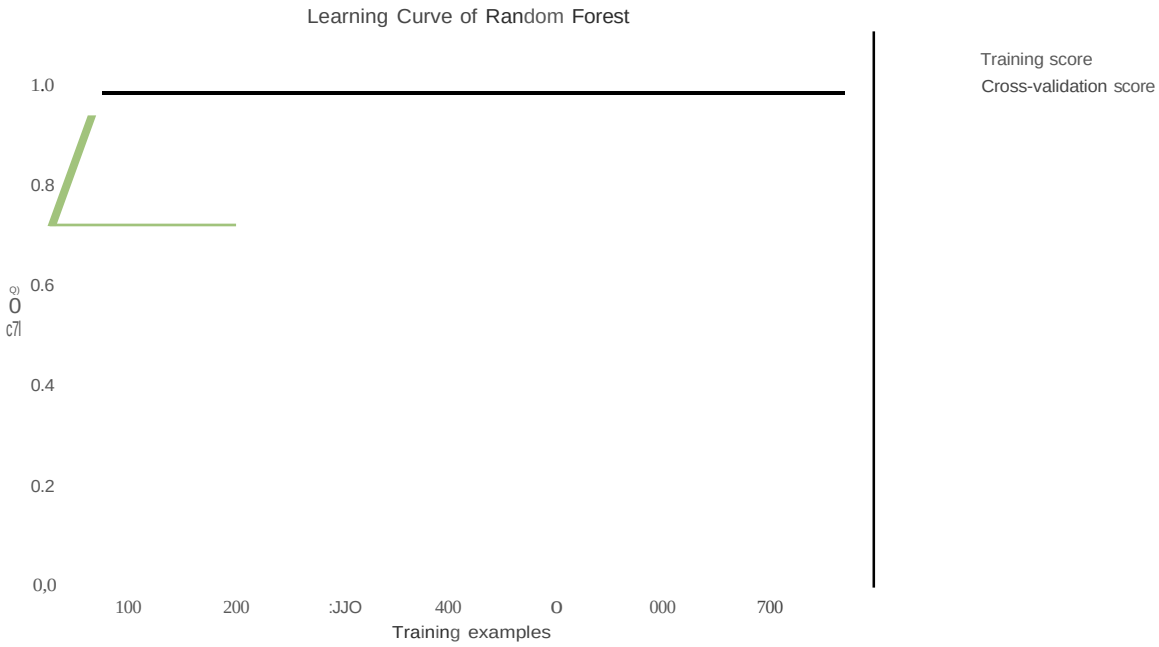
Cross-Validation for the classifications models

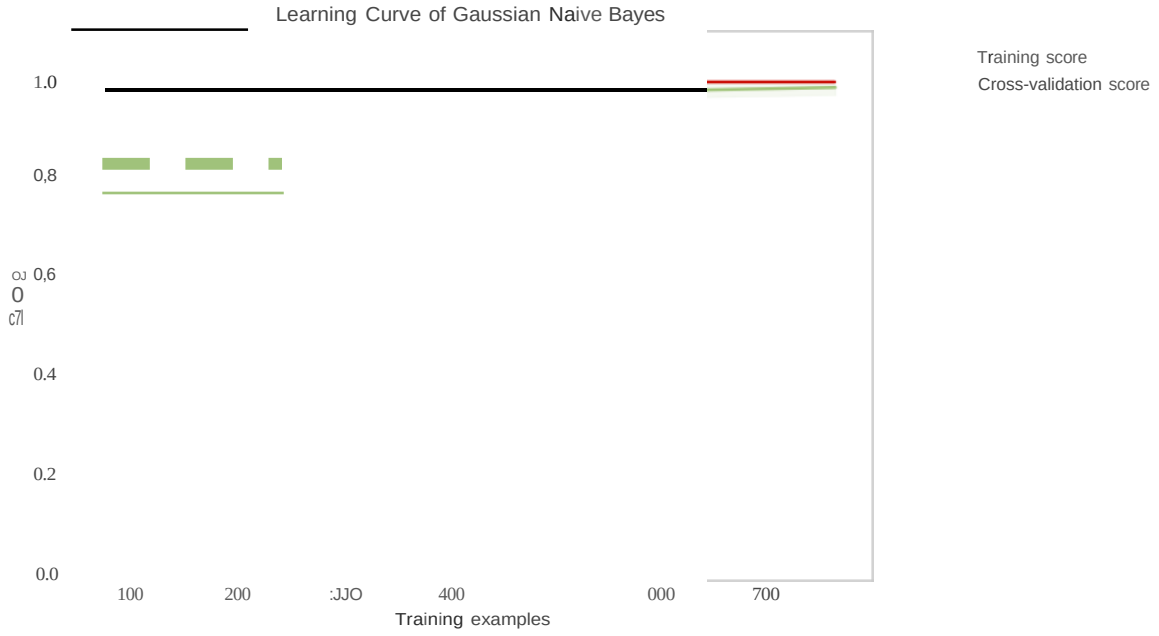


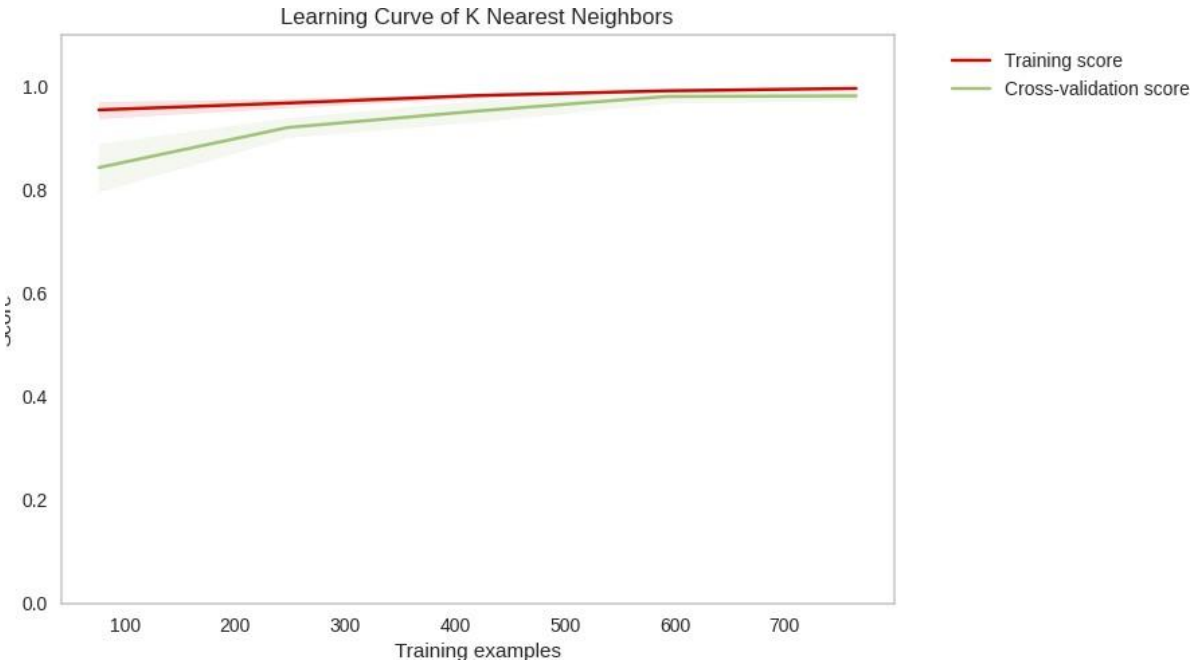
Accuracy Scores for Models

	Model	Accuracy_Score
0	Gaussian Naive Bayes	0.986667
1	Random Forest	0.966667
2	K Nearest Neighbors	0.953333




Learning Curve of Random Forest
Training score
Cross-validation score
1.0
0.8
0.6
0.4
0.2
0,0
100
200
:JJO
400
000
700
Training examples


Learning Curve of Gaussian Naive Bayes
Training score
Cross-validation score
1.0
0,8
0,6
0.4
0.2
0.0
100
200
:JJO
400
000
700
Training examples



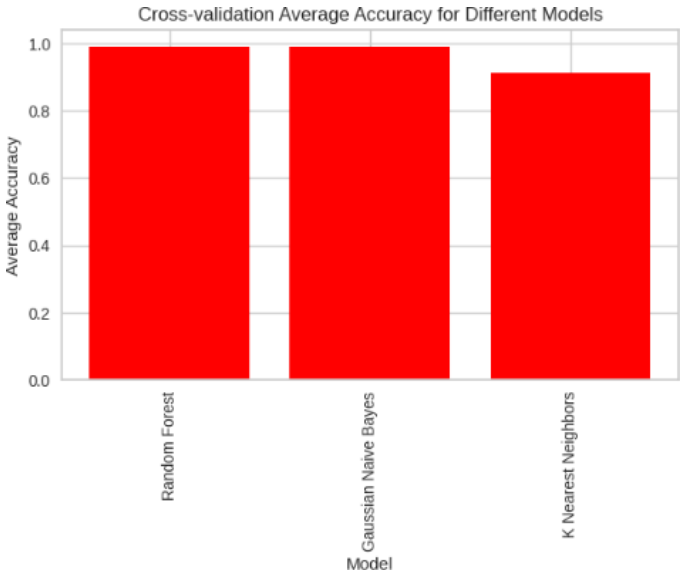
3. N-gram

We use it because it preserves the meaning of words and sentences In our model, we used bigram

	Model	Accuracy_Score
0	Gaussian Naive Bayes	100.00
1	Bernoulli Naive Bayes	98.67
2	Stochastic Gradient Descent SGD	98.67
3	Logistic Regression	97.33
4	XGB Extreme X Gradient Boosting	94.67
5	Support Vector Machine SVM	93.33
6	K Nearest Neighbors	92.00
7	Random Forest	89.33
8	Decision Tree Classifier	88.00
9	Catboost	88.00
10	AdaBoost	60.00

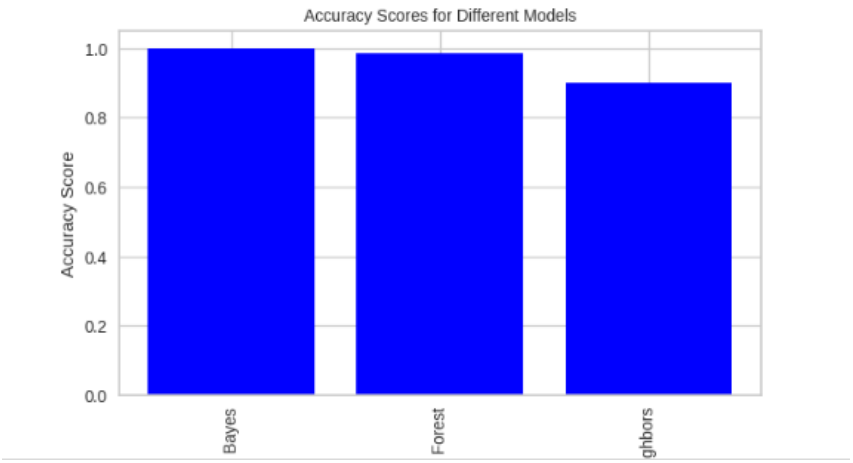
##

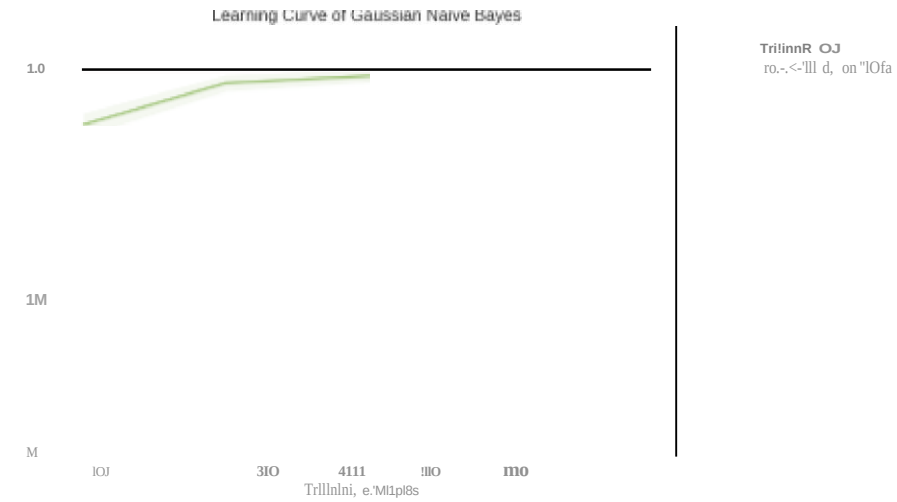
Cross-Validation for the classifications models

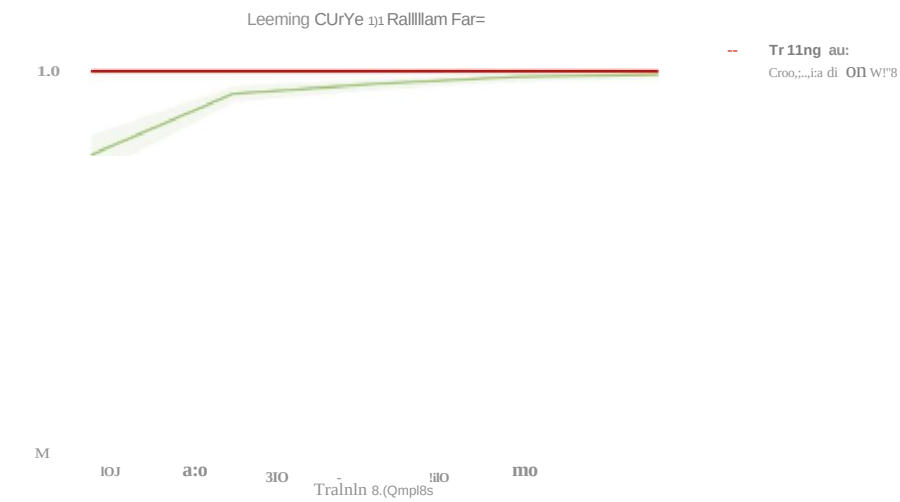


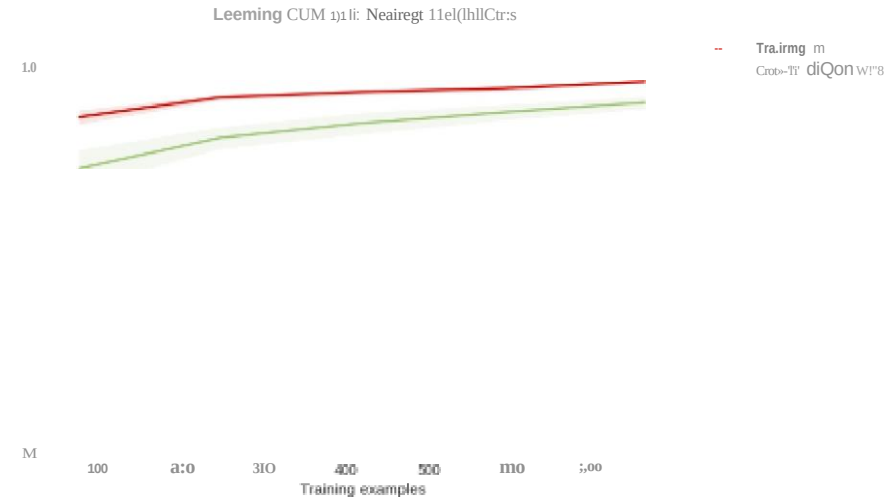
Accuracy Scores for Models

	Model	Accuracy_Score
0	Gaussian Naive Bayes	1.000000
1	Random Forest	0.986667
2	K Nearest Neighbors	0.900000




Learning Curve of Gaussian Naive Bayes
1.0
1M
M
lOJ
3IO
4111
!llO
mo
Trlllnlni, e.'Ml1pl8s
Tri!innR OJ
ro.-.<-'lll d, on "lOfa


Leeming CUrYe 1)1 Rallllam Far=
1.0
M
lOJ
a:o
3IO
!:llO
mo
TraInln 8.(Qmpl8s
-- Tr 11ng au:
Croo,;...i:a di on W!"8

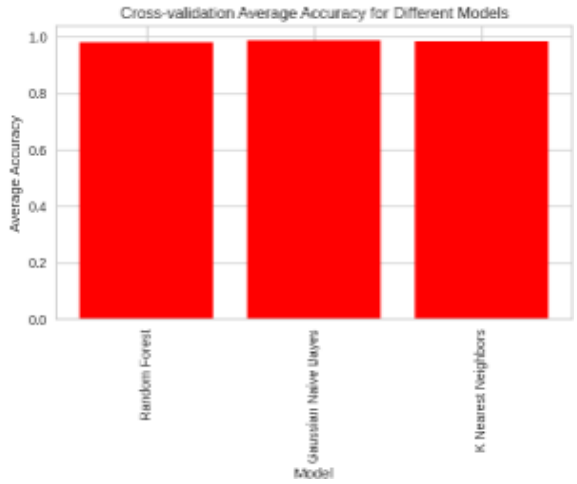

Leeming CUM 1)1 li: Neairegt 11el(lhllCtr:s
1.0
M
100
a:o
3IO
400
500
mo
;,oo
Training examples
-- Tra.irmg m
Crot»-'Tr diQon W!"8

4. Word Embedding

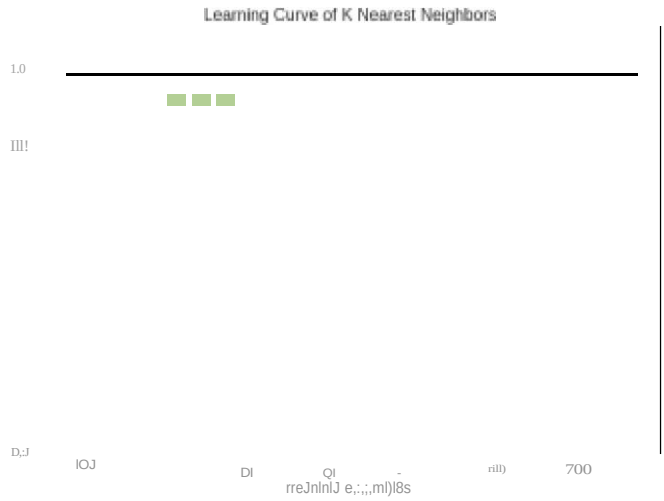
Is a type of word representation that allows words with similar meaning to have a similar representation

	Model	Accuracy_Score
0	XGB Extreme X Gradient Boosting	94.67
1	Catboost	89.33
2	Random Forest	86.67
3	Decision Tree Classifier	84.00
4	K Nearest Neighbors	76.00
5	Gaussian Naive Bayes	58.67
6	Logistic Regression	57.33
7	Stochastic Gradient Descent SGD	53.33
8	AdaBoost	53.33
9	Support Vector Machine SVM	50.67
10	Bernoulli Naive Bayes	34.67

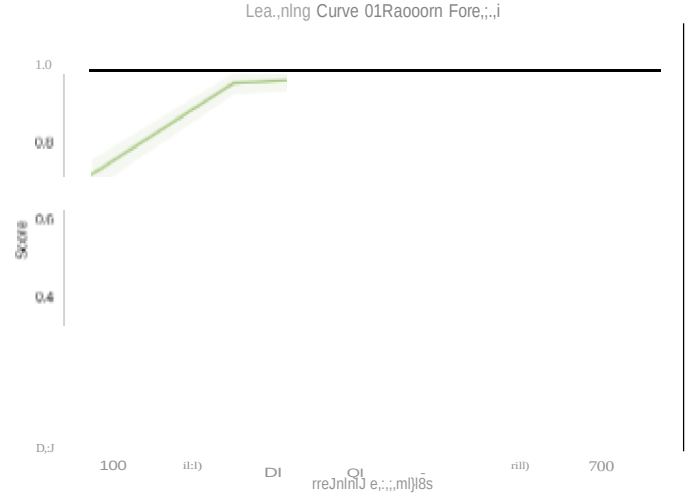
Cross-Validation for the classifications models



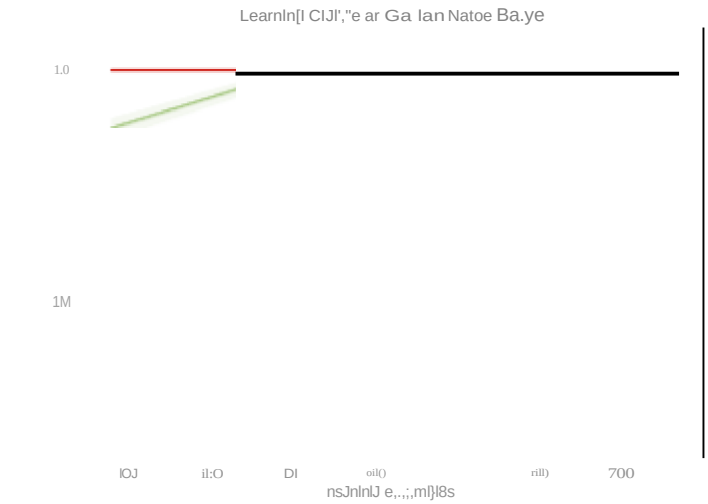
Accuracy Scores for Models



Tri1.u1ng 51:on:
Cr0!!9--• d on=••

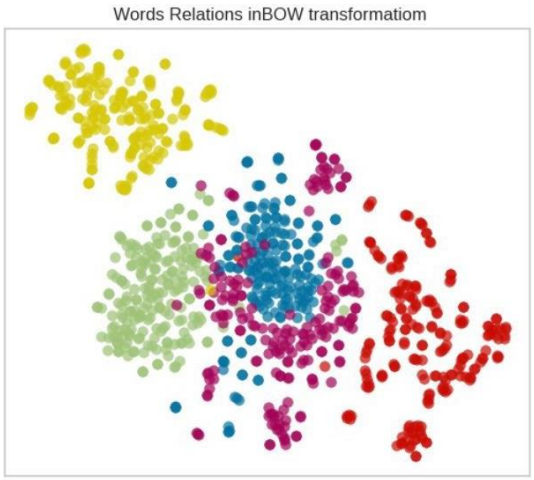


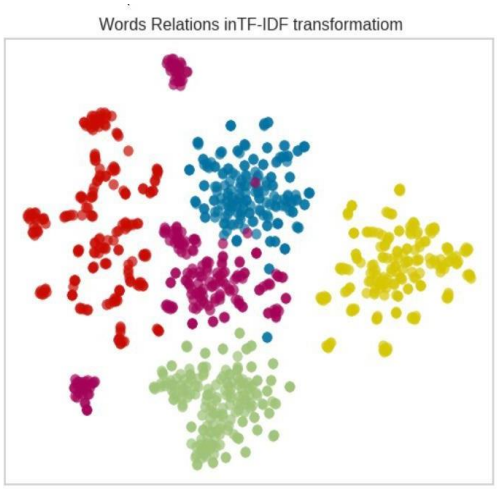
Tri11nng s;c01
CfQtrl.-., dlllon=••

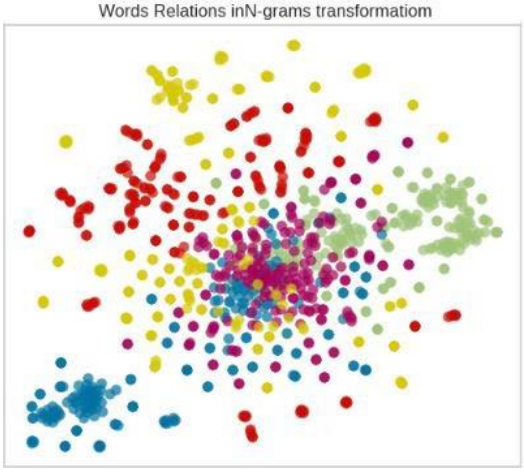


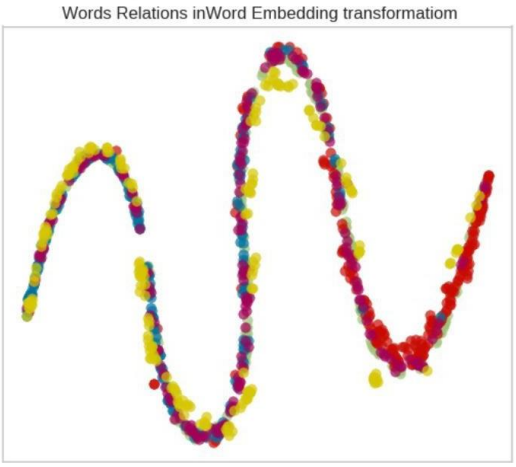
Trail1ng scou:
ro;-.,,;. d, on <oo<11

scatter plots indicates the transforms









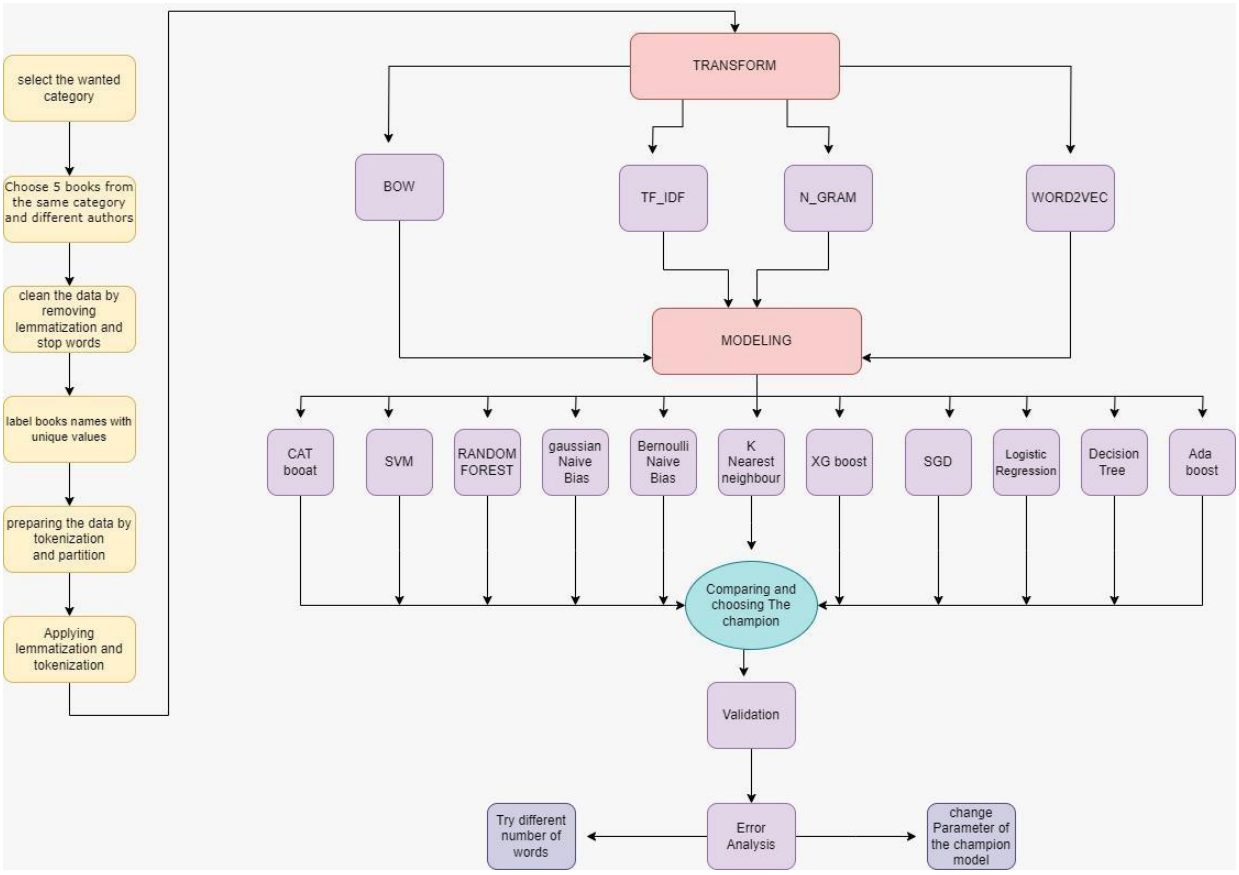
Champion Model:

Based on our models, the best model here is SVM based on TF-IDF with accuracy 100%, so it's our champion model.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17
1	1.00	1.00	1.00	16
2	1.00	1.00	1.00	15
3	1.00	1.00	1.00	9
4	1.00	1.00	1.00	18
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75

The Accuracy of Support Vector Machine SVM is 100.0 %

System workflow



Model Evaluation

Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample to estimate the skill of a machine learning model on unseen data.

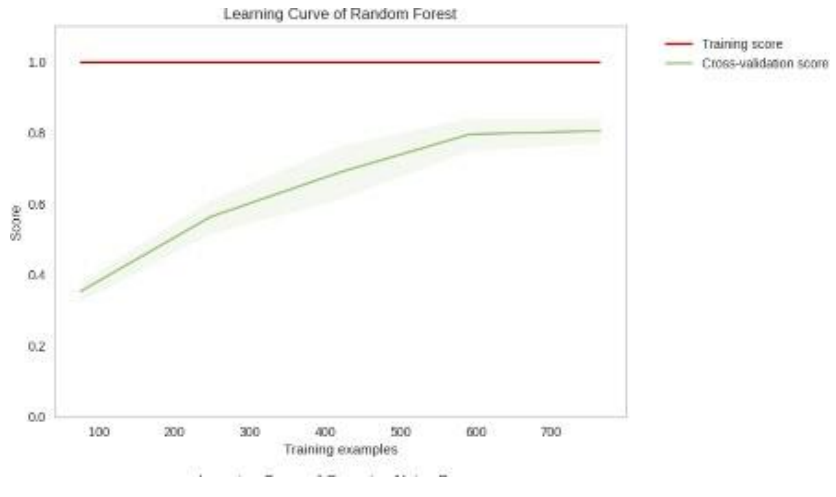
```
Confusion Matrix :
[[26  1  0  2  0]
 [ 2 22  0  3  0]
 [ 0  0 33  1  0]
 [13  4  0  9  0]
 [ 3  0  0  0 31]]

Classification Report :
              precision    recall  f1-score   support

           0       0.59      0.90      0.71        29
           1       0.81      0.81      0.81        27
           2       1.00      0.97      0.99        34
           3       0.60      0.35      0.44        26
           4       1.00      0.91      0.95        34

    accuracy                           0.81       150
   macro avg       0.80      0.79      0.78       150
weighted avg       0.82      0.81      0.80       150


Testing Accuracy of Random Forest is 80.67 %
Average bias: 0.000
Average variance: 1.994
```

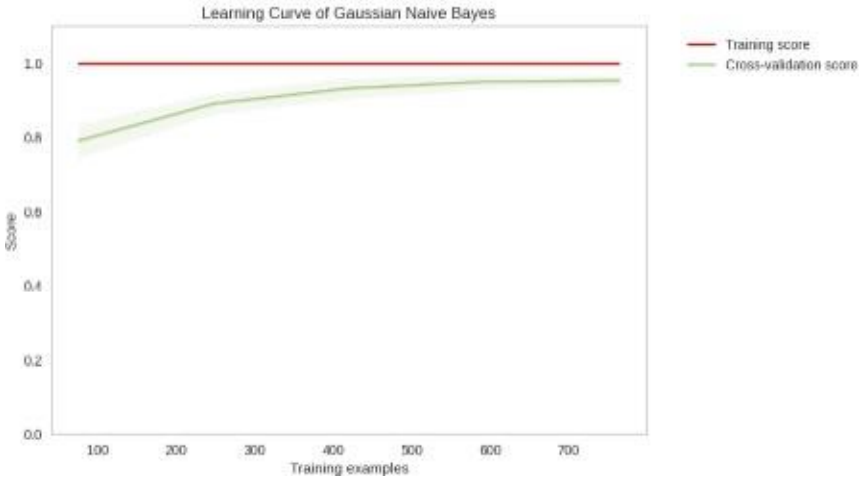
```
Confusion Matrix :
[[24  3  2  0  0]
 [12  7  5  3  0]
 [ 0  3 28  2  1]
 [12  4  3  6  1]
 [22  7  1  0  4]]

Classification Report :
              precision    recall  f1-score   support

           0       0.34      0.83      0.48        29
           1       0.29      0.26      0.27        27
           2       0.72      0.82      0.77        34
           3       0.55      0.23      0.32        26
           4       0.67      0.12      0.20        34

    accuracy                           0.46       150
   macro avg       0.51      0.45      0.41       150
weighted avg       0.53      0.46      0.42       150


Testing Accuracy of Gaussian Naive Bayes is 46.0 %
Average bias: 3.906
Average variance: 1.299
```

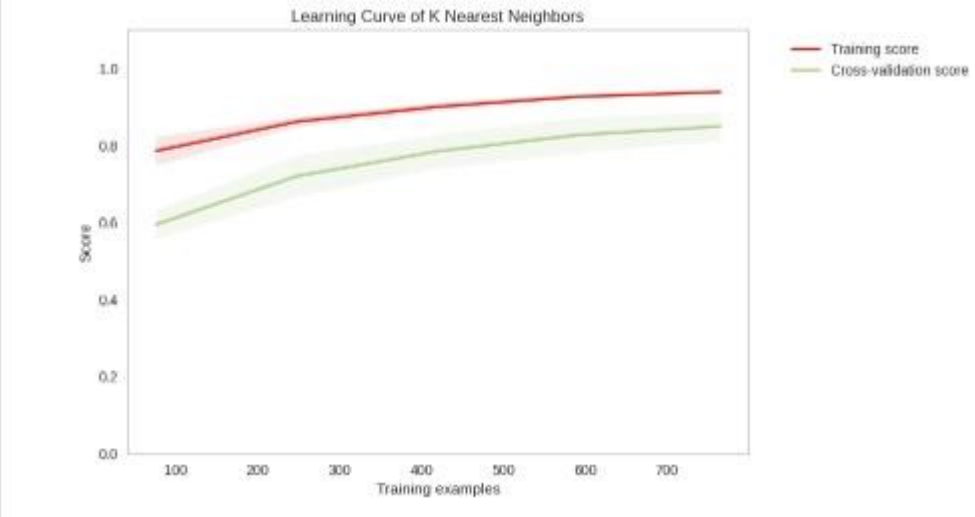
```
Confusion Matrix :
[[28  1  0  0  0]
 [ 4 21  0  2  0]
 [ 2  1 29  2  0]
 [15  5  2  4  0]
 [ 7  1  0  2 24]]

Classification Report :
              precision    recall  f1-score   support

           0       0.50      0.97      0.66        29
           1       0.72      0.78      0.75        27
           2       0.94      0.85      0.89        34
           3       0.40      0.15      0.22        26
           4       1.00      0.71      0.83        34

    accuracy                           0.71       150
   macro avg       0.71      0.69      0.67       150
weighted avg       0.74      0.71      0.69       150


Testing Accuracy of K Nearest Neighbors is 70.67 %
Average bias: 0.962
Average variance: 2.063
```

Learning curve

Another way to get an estimate of model's generalization performance is Learning Curves, it shows the model performance on training and validation sets as a function of training set size (training iterations), Ideal Learning Curve Model generalizes on testing and training data. the smaller the gap between the training and cross-validation scores, the better the model in generalization.







Error Analysis of Champion Model

-By reducing the number of words, it will lead to reduce the accuracy of our champion model as follows:

```
Accuracy with number of words 100 is 98.67 %

Accuracy with number of words 70 is 97.33 %

Accuracy with number of words 50 is 96.0 %

Accuracy with number of words 40 is 93.33 %

Accuracy with number of words 30 is 88.0 %

Accuracy with number of words 20 is 77.33 %
```

-indicate that the n estimators' parameter is not significantly impacting the model's performance on our dataset.

```
Accuracy with number of words 100 is 97.33 %

Accuracy with number of words 70 is 97.33 %

Accuracy with number of words 50 is 97.33 %

Accuracy with number of words 40 is 97.33 %

Accuracy with number of words 30 is 97.33 %

Accuracy with number of words 20 is 97.33 %
```