[DTI5125 [EG] Data Science Applications Group 3, Assignment 1, Text Classification]

[Ahmed Abdelaal]

[Motaz Habib]

[Shahd Mohamed]

[Yousef Mohamed]

Introduction:

The project focuses on detecting the writing style of authors with mixed nationalities in the romance genre using natural language processing (NLP) by choosing five different books and experimenting with different techniques in NLP in each step in the methodology and sees how it will affect the results and also if the nationality is a contributing factor in the detecting process.

Dataset:

We choose five different books from the Gutenberg corpus in the romance genre with different authors from different origins.

The chosen books are:

1. Pride and Prejudice by Jane Austen (English)
2. Anna Karenina by Leo Tolstoy (Russian)
3. Little Women by Louisa M. Alcott (American)
4. Wuthering Heights by Emily Brontë (English)
5. This Side of Paradise by F. Scott Fitzgerald (American)

Data preprocessing and cleansing steps:

1. We loaded the data from the Gutenberg corpus using URL.

2. We took from each book 200 random partitions with each partition contains 100 words.

3. We Removed stop words, special characters, numbers, and punctuation and then performed lemmatization on each partition.

	Author	Paragraphs
0	Jane Austen	advantage miss lydia bennet come upon happiest...
1	Jane Austen	unaffectedly civil force younger sister civil ...
2	Jane Austen	give one two early speeches slight come seriou...
3	Jane Austen	report engagement could elizabeth loss till re...
4	Jane Austen	higher others solidity reflections often strik...
...	...	...
995	F. Scott Fitzgerald	tell worry seem progress perfectly amory lose ...
996	F. Scott Fitzgerald	lately reread aeschylus divine irony find answ...
997	F. Scott Fitzgerald	well know little man laugh conscientious stop ...
998	F. Scott Fitzgerald	afterward repeat anecdotes life could make sou...
999	F. Scott Fitzgerald	every one sew bag rest college amory find writ...

1000 rows × 2 columns

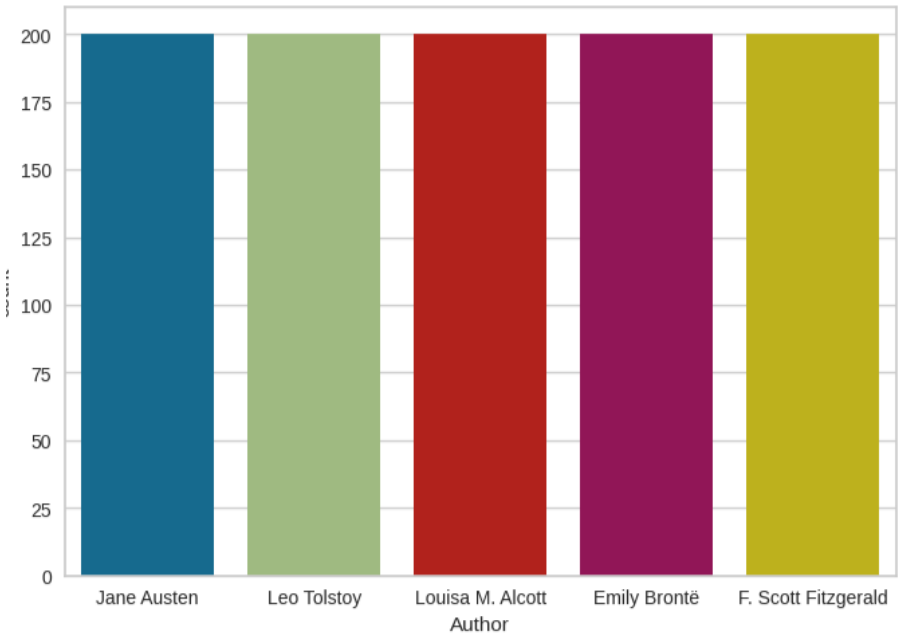
Feature Engineering:

1. We applied sentiment analysis to have insight into the overall mood.

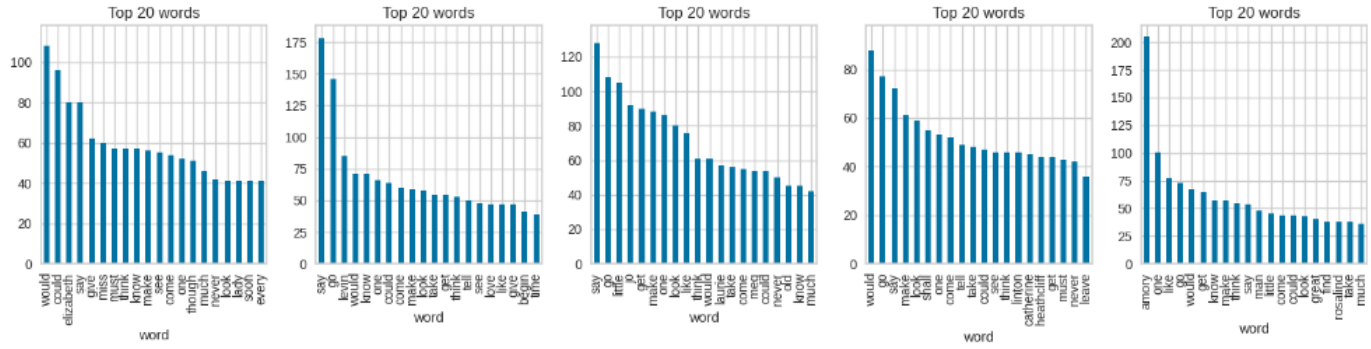
Name	Score	Overall_Emotion
Jane Austen	0.19	Positive
Leo Tolstoy	0.08	Positive
Louisa M. Alcott	0.2	Positive
Emily Brontë	-0.02	Neutral
F. Scott Fitzgerald	0.08	Positive

2. We generated a word cloud for each book to visualize the most frequent words.

3. We checked if there is bias in the data by visualizing it.

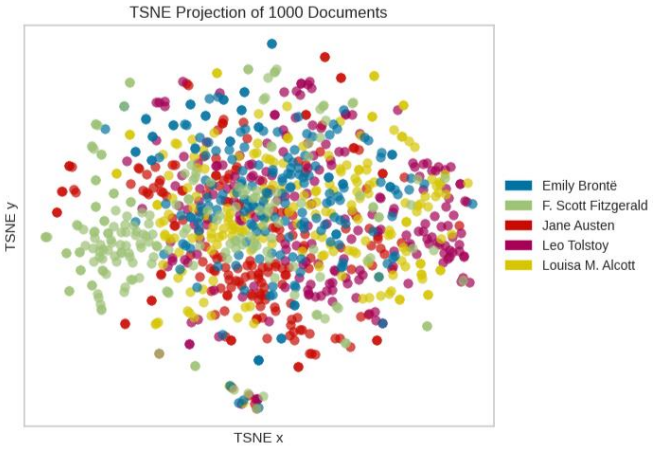


4. We plotted the top 20 words in each book to have a general idea of the content.

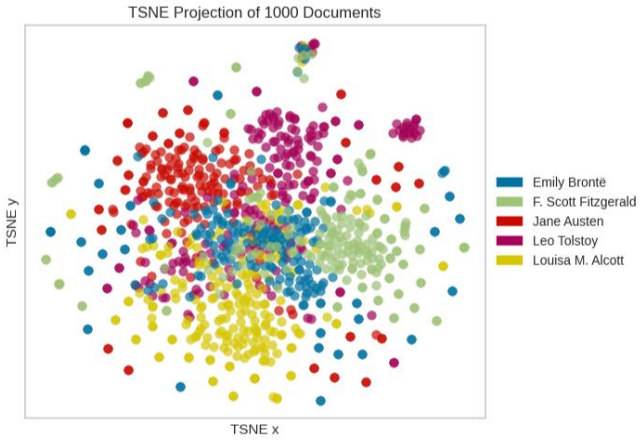


5. We applied various transformation techniques to the data such as BOW, TF-IDF, and N-gram.

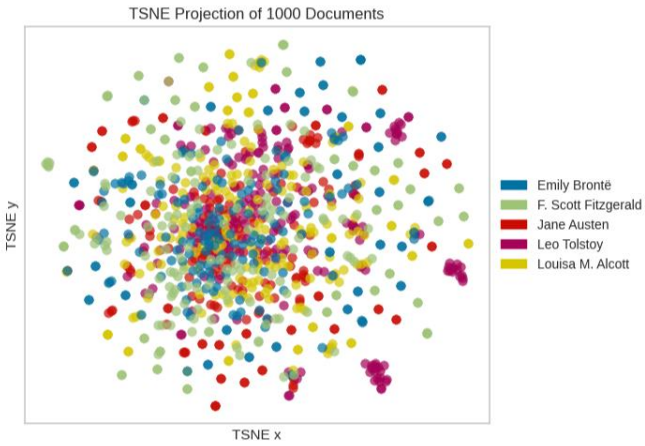
6. We Visualized our transformed data using TSNE and we noticed that TF-IDF has the better separation of all three.



BOW



TF-IDF



N-gram

Modeling:

We applied the transformed data on various models:

1. SVM
2. Random Forest
3. Naïve bayas,
4. KNN
5. XG-boost
6. Logistic regression
7. BERT

with the default parameters with data split of 80 – 20 for training and testing respectively and with each model 10 folds cross validation.

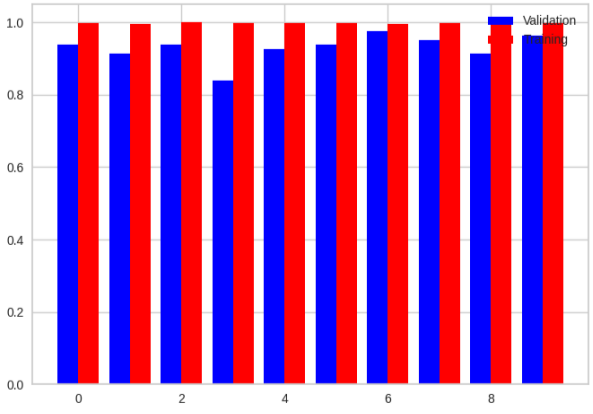
Accuracy of each model						
	SVM	Random Forest	Naïve bayas	KNN	XG-boost	Logistic regression
BOW	0.87	0.855	0.85	0.545	0.845	0.875
TF-IDF	0.9	0.89	0.805	0.795	0.84	0.91
N-gram	0.56	0.43	0.64	0.225	0.45	0.595
Cross Validation						
	SVM	Random Forest	Naïve bayas	KNN	XG-boost	Logistic regression
BOW	0.874	0.880	0.855	0.469	0.845	0.904
TF-IDF	0.932	0.886	0.826	0.780	0.853	0.929
N-gram	0.495	0.417	0.665	0.244	0.357	0.616

Deep Learning Model
BERT
0.804

Choosing the champion model:

- Based on models' performance, the best combination is Logistic regression (Default parameters) with TF-IDF that got an accuracy of 91%
- Based on the accuracies of Logistic regression on TF-IDF in model and cross validation we see that the model fits correctly the data and does not overfitting nor underfitting.

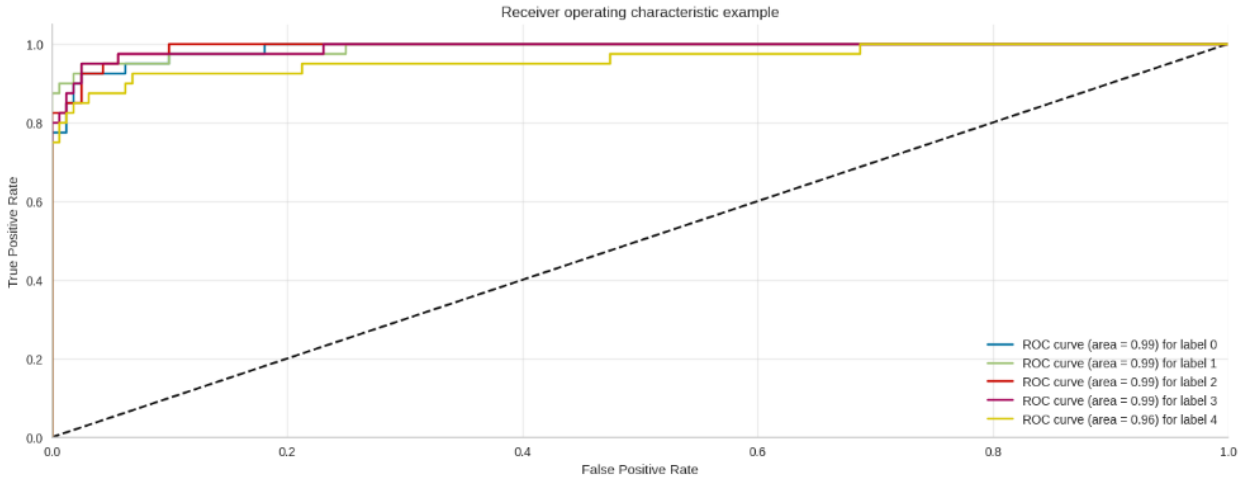
```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.93      0.87        40
           1       0.95      0.93      0.94        40
           2       0.92      0.90      0.91        40
           3       0.95      0.93      0.94        40
           4       0.92      0.88      0.90        40

    accuracy                           0.91       200
   macro avg       0.91      0.91      0.91       200
weighted avg       0.91      0.91      0.91       200


Best Model: Logistic Regression with Accuracy: 0.91
```



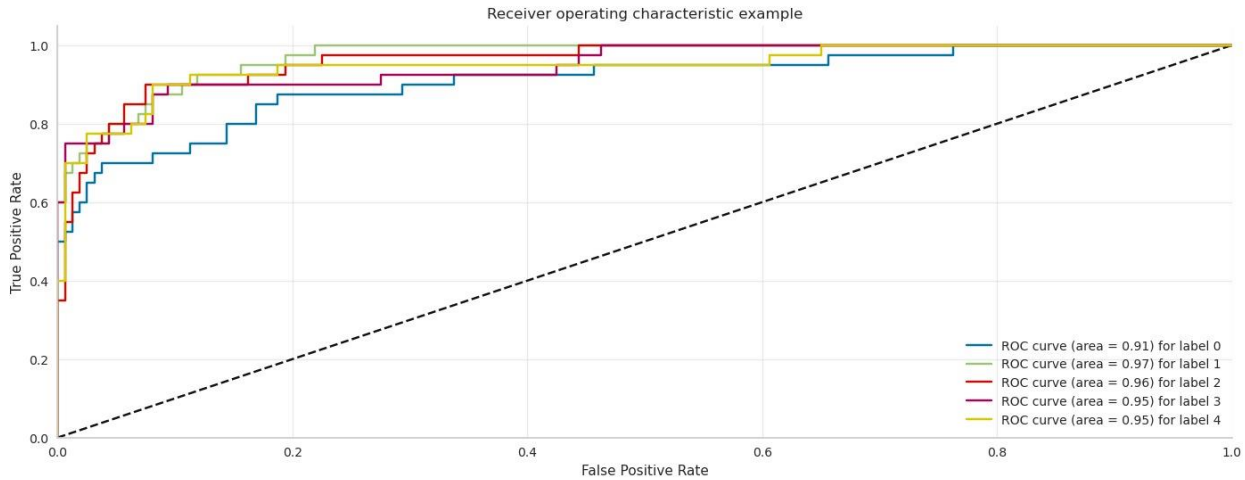

```
Cross Validation accuracy scores: [0.9375 0.9125 0.9375 0.8375 0.925  0.9375 0.975  0.95   0.9125 0.9625]
Cross Validation accuracy: 0.929 +/- 0.036
Logistic Regression Accuracy: 0.91
```

Bias and Variability

- From the performance results that we got; the results indicates that nationality of the authors didn't affect much in the style of writing. As the champion model predicted each class well.
- The AUC of each class is quite similar.
- After we reduced the number of words in each partition to fifty words, the classification of each class became low.



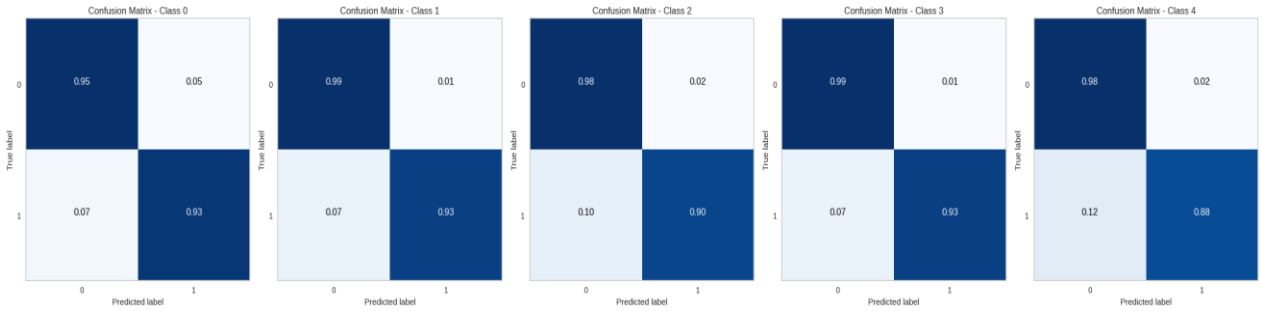
ROC before decreasing the number of words



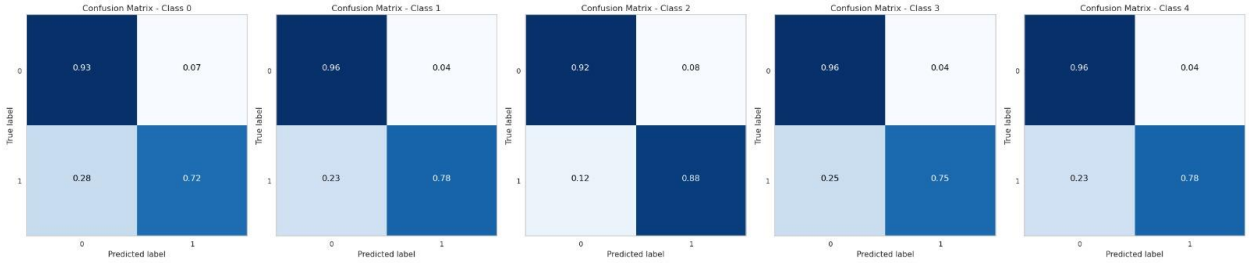
ROC after decreasing the number of words

Error Analysis

- From our confusion matrices results, our champion model got lower accuracy with class four compared to the other classes.
- Class two has lower accuracy but higher than accuracy of class 4.
- The main problem with TF-IDF is that it doesn't understand how words relate to each other or consider the order they appear in a sentence. It treats each word separately, which can make it difficult to understand the overall meaning of the text.
- After we reduced the number of words in each partition to fifty words, the overall accuracy decreased which imply to under fitting.



Confusion matrices before decreasing the number of words



Confusion matrices After decreasing the number of words

Program README:

- The project consist of 3 files:
  1. “Text_Classification_V1.ipynb” which contains the pipeline with original dataset.
  2. “Text_Classification_V2.ipynb” contains the pipeline with alternative dataset.
  3. “Text_Classification_Using_BERT.ipynb” contains pretrained deep learning model.