uOttawa

Report

Assignment

Data Science Applications

Group Number: G4

Team Members:

Abdallah Ragab, Hossam Mahmoud

Ahmed Ali Ziada, Shaimaa Mamdouh Ahmed

Table of contents

1. Overview

The main objective in this assignment is how to classify large text data. We need to build machine learning models can perform this task on it. Our Goal is achieving this model with best performance by using different algorithms, packages, and tools of python.

2. Methodology

2.1 Import useful libraries

for reading and plotting such as pandas, NumPy, Sklearn, seaborn, matplotlib, etc.

2.2 Load 5 different books of science fictions category for different authors.

loading the Gutenberg dataset from NLTK package.

Selected books:

- Emma, by Jane Austen
- The Parent's Assistant, by Maria Edgeworth
- Moby-Dick, Herman Melville
- The Man Who Was Thursday, by G. K. Chesterton
- Alice's Adventures in Wonderland, by Lewis Carroll

2.3 Pre-processing of the Data

2.3.1 Cleaning

Using regular expression, the book title, chapter title, Volume title, empty lines, punctuation marks, and stopwords have been removed, then convert the book to small text.

2.3.2 Partitioning

creating random samples of 200 documents of each book and put them into a data frame, each record in this data frame contains 100 words for each document labeled with authors' names.

2.3.3 Label Encoding

Converting the names of the authors to numerical type such as 0, 1, 2, … etc.

2.4 Feature engineering

In this step several methods were preformed such as, The Bag of words (BOW) technique, Term Frequency-Inverse Document Frequency (TF-IDF), and Term Frequency-Inverse Ngram to Convert text into vector.

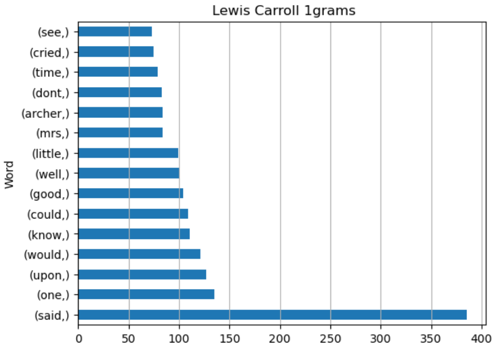
2.5 Subset the data

Define features which are words and target which is the book's title.

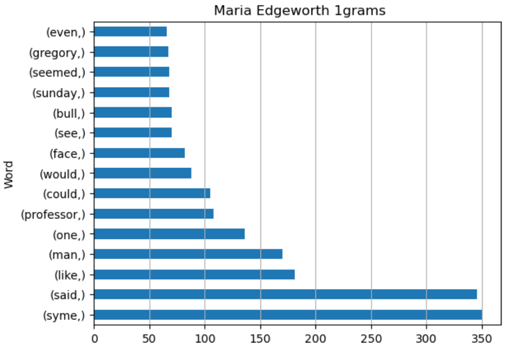
2.6 Split the data

Splitting the data into training data, testing data and validation data

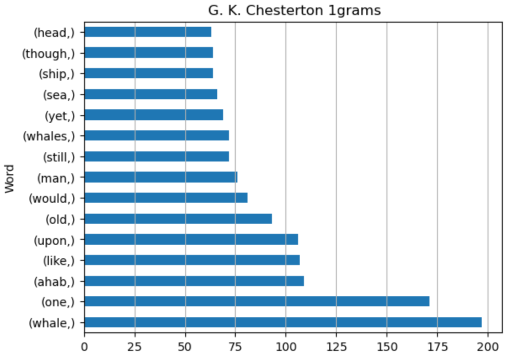
2.6.1 plotting the most repeating fifteen words in each book



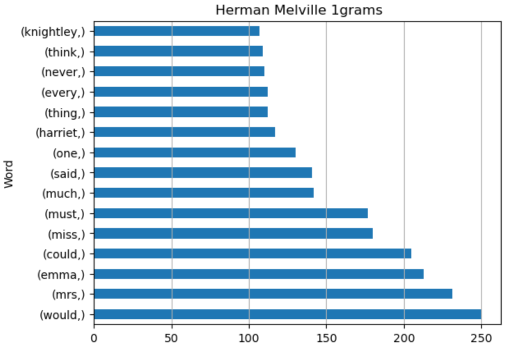
Alice's Adventures in Wonderland



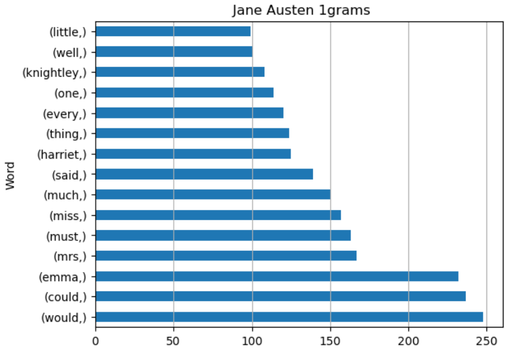
The Parent's Assistant



The Man Who Was Thursday



Moby-Dick



Emma

2.7 Models

Multiple models were preformed:

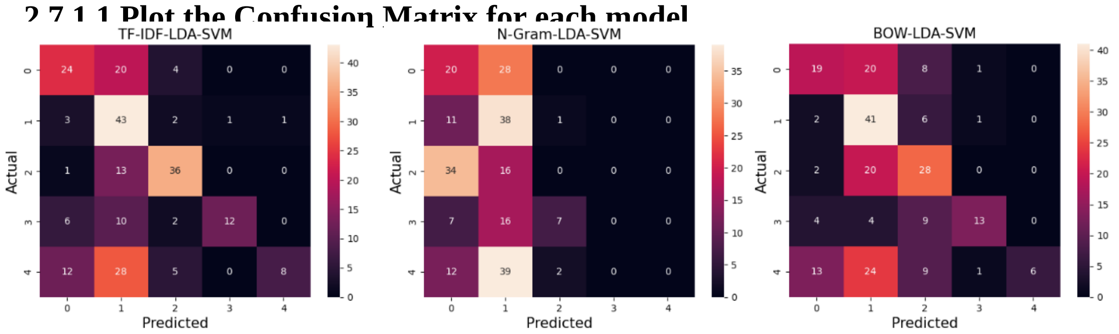
- Support Vector Machine
- Decision Tree
- k-nearest Neighbors algorithm
- XG Boosting
- XG_PCA

These models trained by four different methods of Text transformation

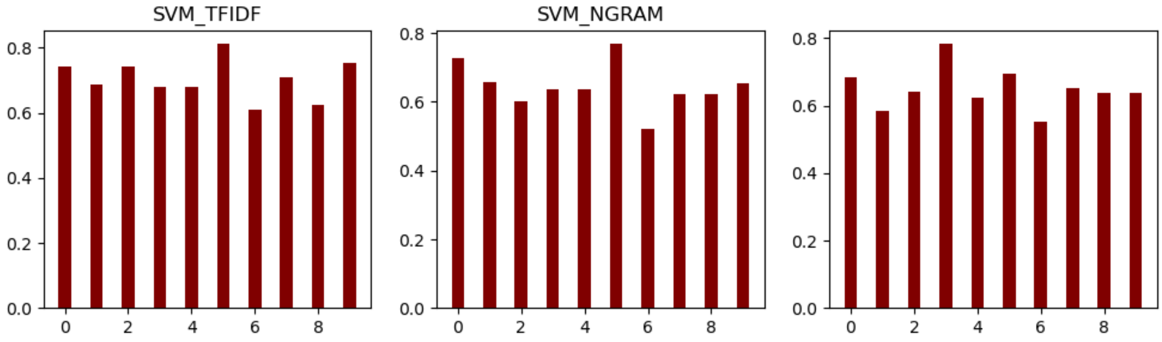
- BOW-LDA
- TF-LDA
- N-Gram
- PCA

2.7.1 Support Vector Machine

2.7.1.1 Plot the Confusion Matrix for each model

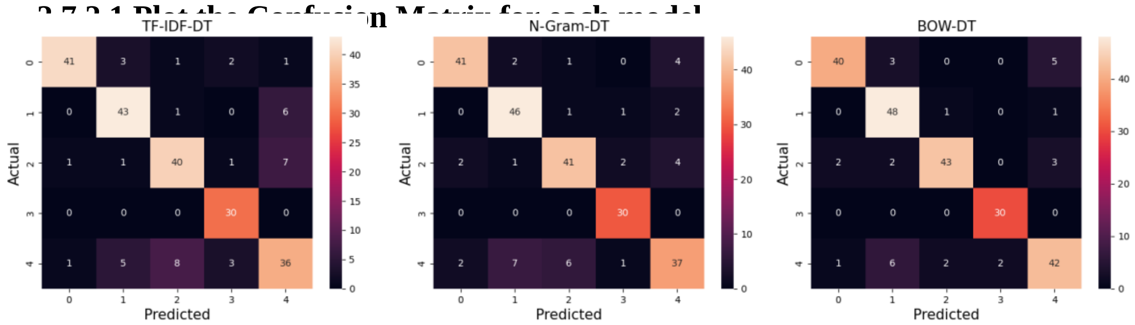


2.7.1.2 Cross Validation and accuracy for each model

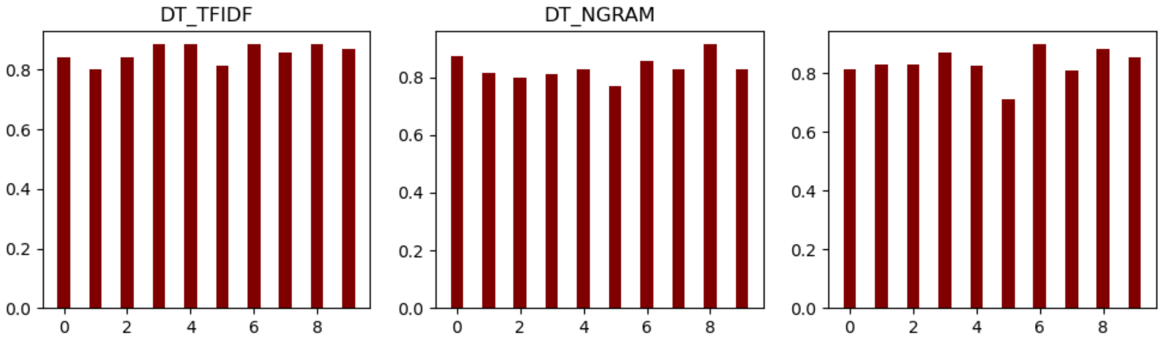


2.7.2 Decision Tree

2.7.2.1 Plot the Confusion Matrix for each model

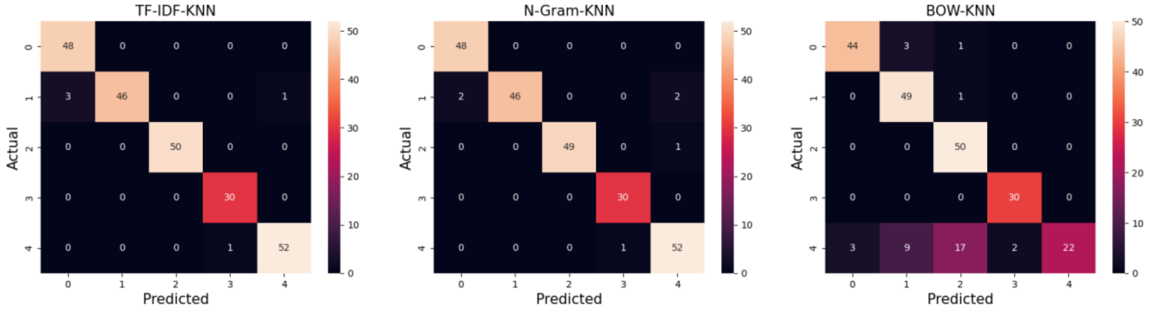


2.7.2.2 Cross Validation and accuracy for each model

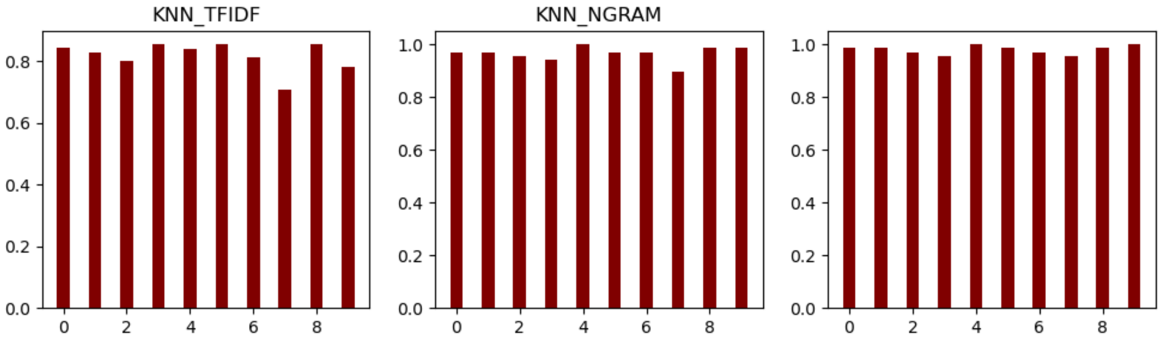


2.7.3 k-nearest Neighbors algorithm

2.7.3.1 Plot the Confusion Matrix for each model.

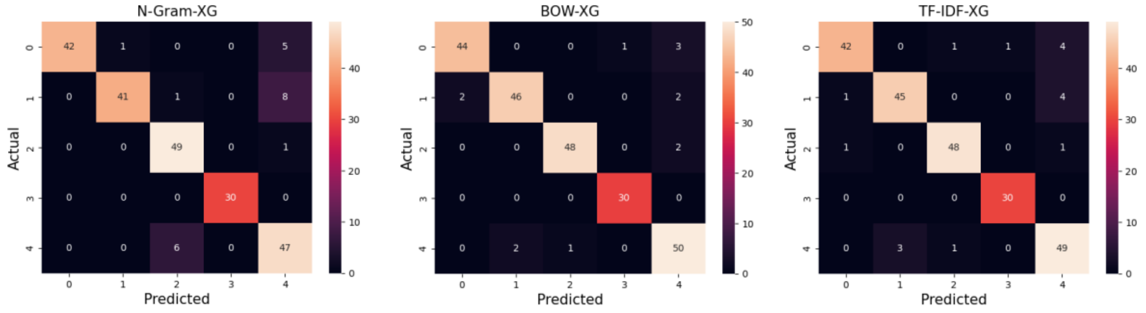


2.7.3.2 Cross Validation and accuracy for each model

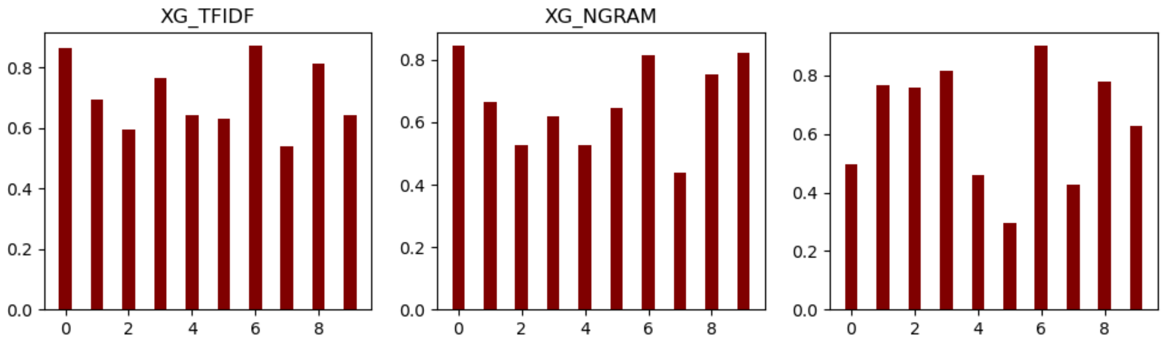


2.7.4 XG Boosting

2.7.4.1 Plot the Confusion Matrix for each model.

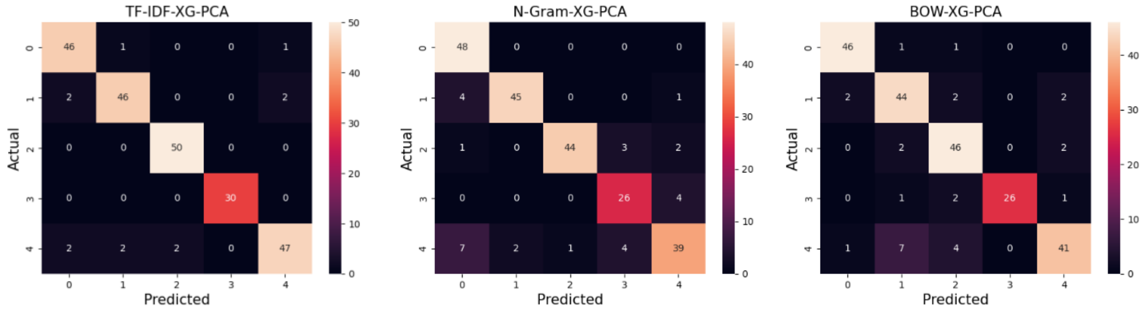


2.7.4.2 Cross Validation and accuracy for each model

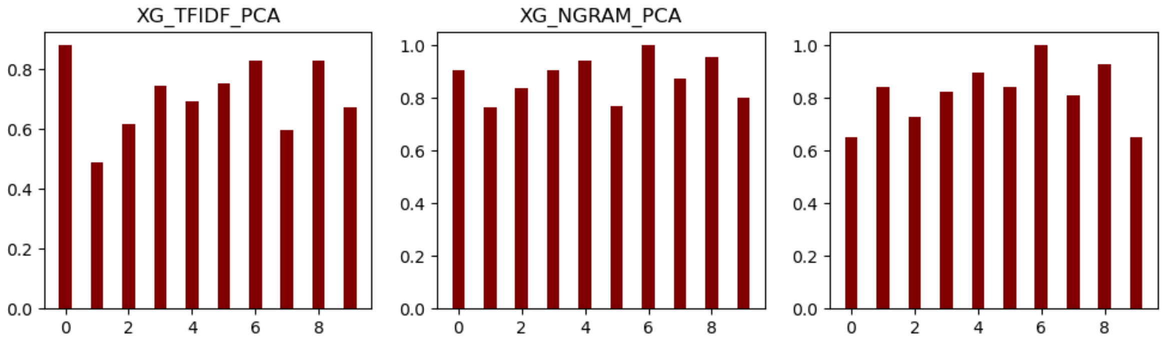


2.7.5 XG_PCA

2.7.5.1 Plot the Confusion Matrix for each model.



2.7.5.2 Cross Validation and accuracy for each model



2.8 Compare models

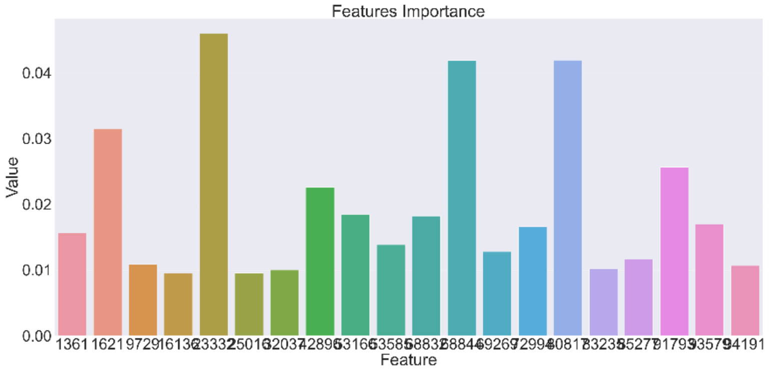
Models	Accuracy	Precision	Recall	F1 Score
SVM_BOW	**0.85**	**0.186**	**0.212**	**0.171**
SVM_TFIDF	**0.84**	**.212**	**0.221**	**0.188**
SVM_NGRAM	**0.69**	**0.042**	**0.2**	**0.06**
DT_BOW	**0.31**	**0.157**	**0.161**	**0.157**
DT_TFIDF	**0.314**	**0 .156**	**0.159**	**0.159**
DT_NGRAM	**0.92**	**0.84**	**0.827**	**0.832**
KNN_BOW	**0.318**	**0.198**	**0.195**	**0.188**
KNN_TFIDF	**0.246**	**0.21**	**0.216**	**0.209**
KNN_NGRAM	**0.988**	**0.983**	**0.987**	**0.984**
XG_BOW	**0.30**	**0.156**	**0.188**	**0.154**
XG_TFIDF	**0.24**	**0.170**	**0.19**	**0.165**
XG_NGRAM	**0.94**	**0.881**	**0.862**	**0.868**
XG_BOW__PCA	**0.28**	**0.208**	**0.218**	**0.168**
XG_TFIDF_PCA	**0.27**	**0.197**	**0.211**	**0.199**
XG_NGRAM_PCA	**0.98**	**0.887**	**0.87**	**0.87**

2.9Champion Model

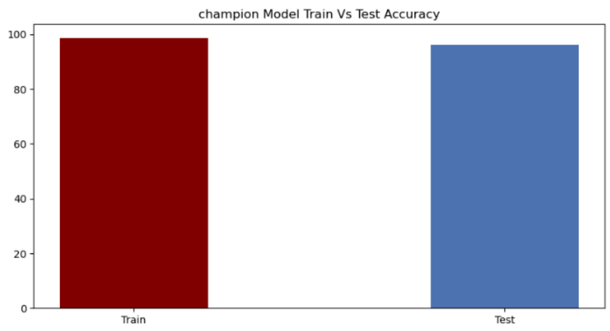
By comparing between all accuracies for each model **KNN Classifier** that trained on Ngram features had the highest accuracy, this model is our champion model.
After the champion model is chosen, it has been tested on test set and the left-hand figure shows the different between train and test accuracies.

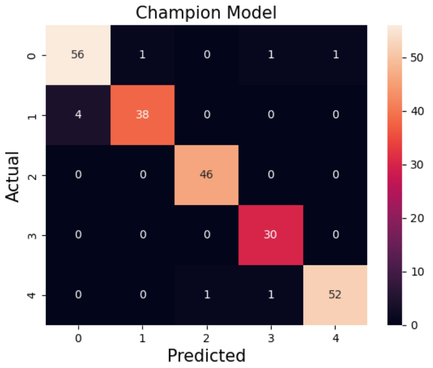
2.10 Make it harder for our model

To make the classification process more difficult, most twenty important features were plotted to recognize them as shown below, then they were dropped from train and test data sets. Finally, champion model was retrained on the new sets shown different accuracy.

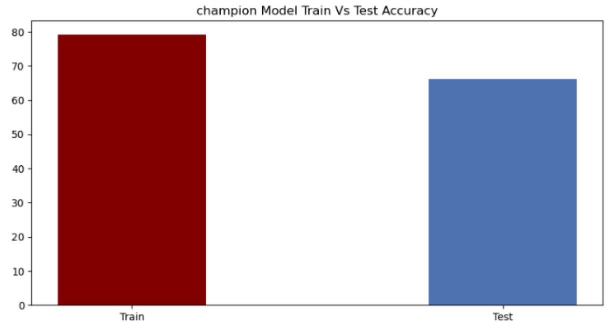


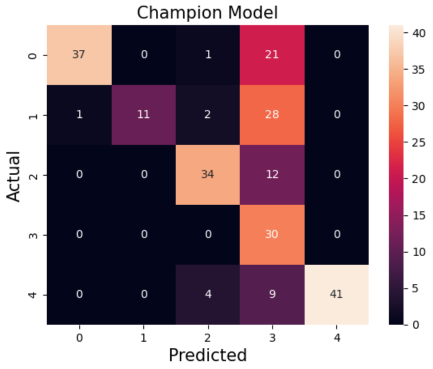
- **Before removing important features**





- **After removing important features**





3. Conclusion

We have learned many new things during this assignment, and we have discovered some useful techniques like SVM, KNN, Decision trees, XG Boost. we have gotten familiar with new Dimensionality reduction techniques like PCA and LDA.
And we learned how to clean, tokenize and vectorize our text data, and we have learned different visualization methods to visualize our results.
And finally, we learned how to improve the accuracies of all the different models that we have crated, by applying cross validation technique and based on that we have chosen the champion model.