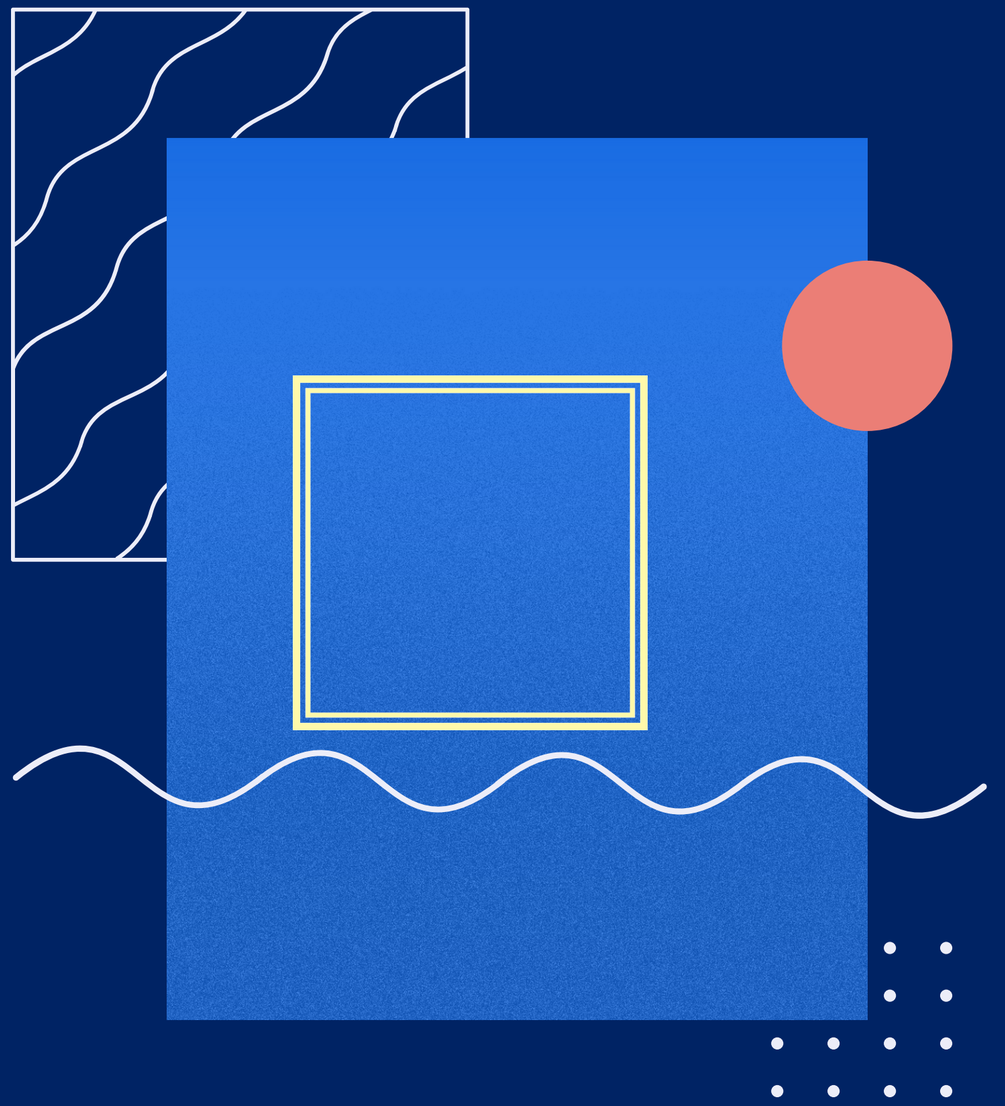
NLP-Classification Assignment

GROUP 2



WORKING FLOW:

In our project we used general form of coding on our chosen data then we started the preprocessing removing the special characters, stop words and the unusual spaces as cleaning for the data

Now we have our data prepared so we grouped our data by the name of the book and we have encoded the labels using LabelEncoder

To be able to deal with our data we transformed it in many forms as a feature engineering like (BOW, TF-IDF, BOW with n-gram and TF-IDF with n-gram)

To reduce the number steps in our code and to have a better shape for the data, we used the pipeline technique

Then we started training our models (SVM, KNN, Naive bayes and Decision tree) with different way of transformation each time

After that we chose the best model of each type of models then we performed evaluation on the selected four models to obtain the champion model

After that we chose the best model of each type of models then we performed evaluation on the selected four models to obtain the champion model

Exploring Data Analysis

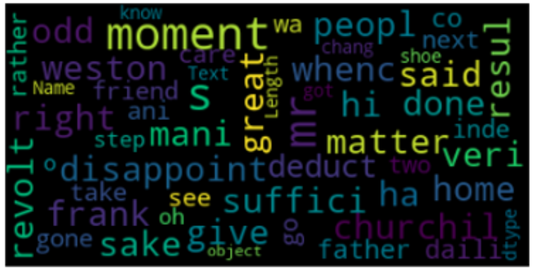
1) Top words for Alice's adventures in wonderland

2) Top words for Dracula

3) Top words for the war of the worlds

4) Top words for uncle tom's cabin or life among the lowly

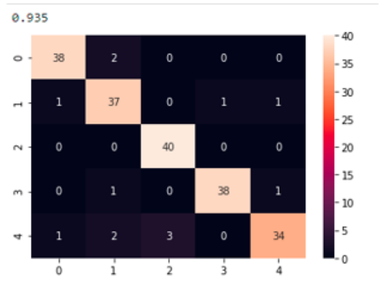
5) Top words for Emma



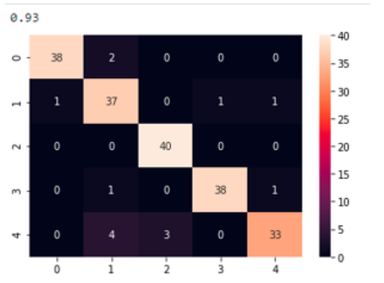
Models Confusion matrices AFTER TRAINING WITH THE DIFFERENT TYPES OF TRANSFORMATION

Naïve Bayes

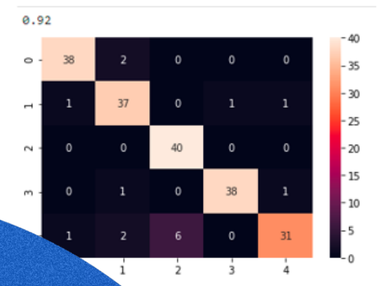
Naïve Bayes with BOW



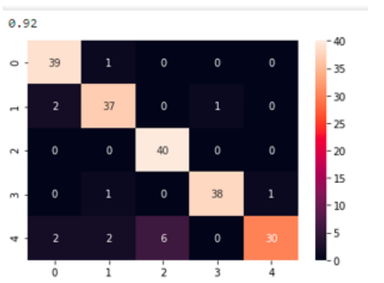
Naïve Bayes with BOW and N-gram



Naïve Bayes with TF-IDF

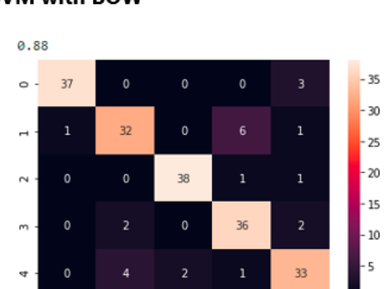


Naïve Bayes with TF-IDF and N-gram



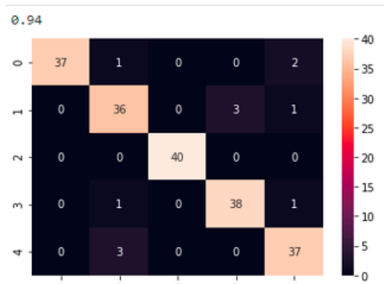
SVM

SVM with BOW



SVM with BOW and N-grams

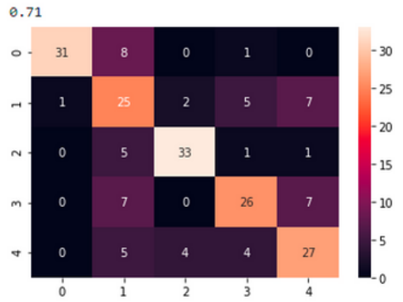
SVM with TF-IDF



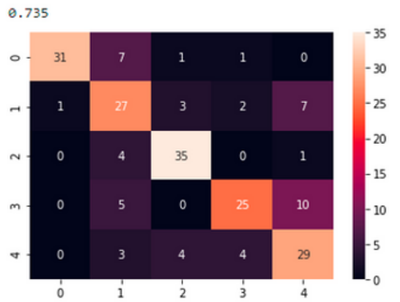
SVM with TF-IDF and N-grams

Decision tree

Decision tree model with BOW



Decision tree model with BOW and N-gram

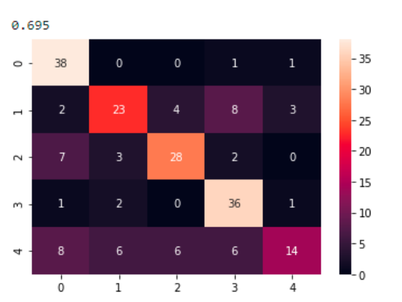


Decision tree with TF-IDF

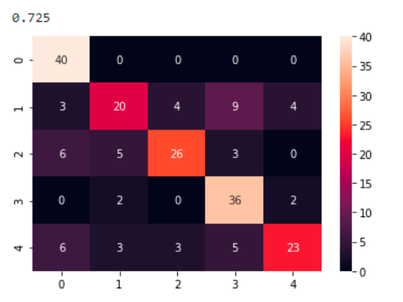
Decision tree with TF-IDF N-gram

KNN

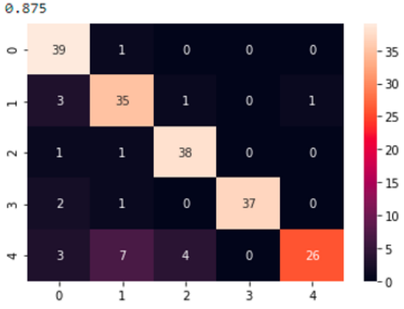
KNN model with BOW



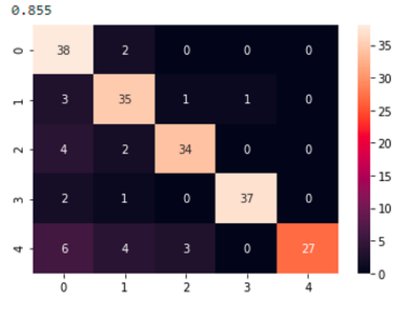
KNN model with BOW and N-grams



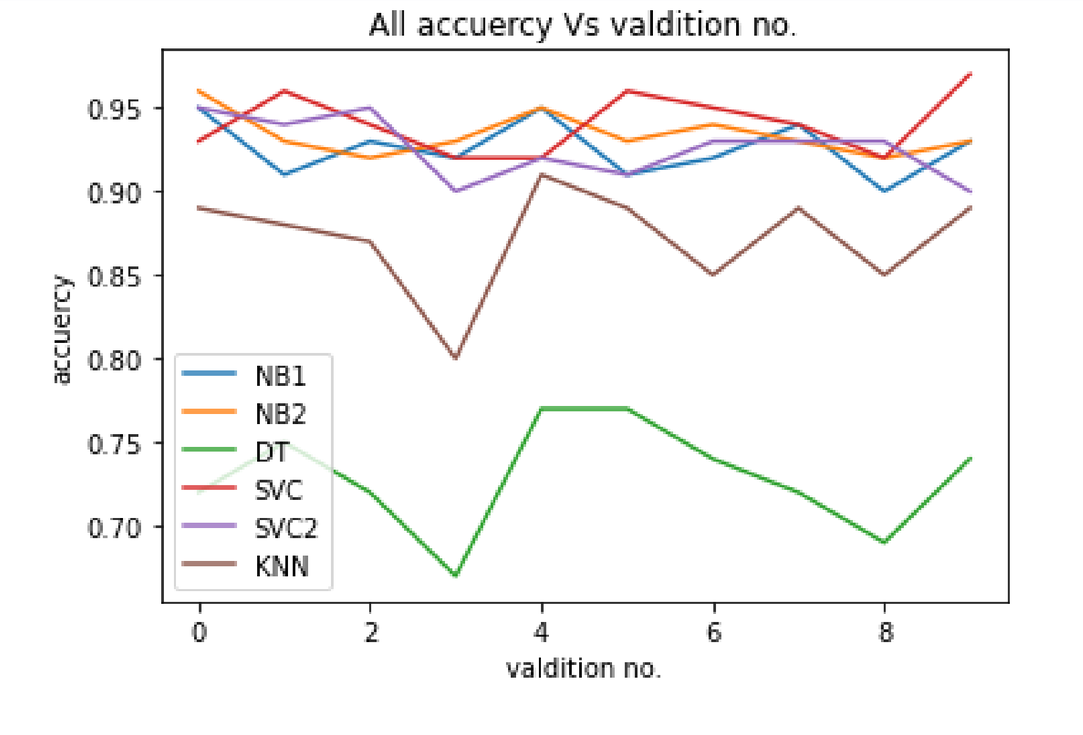
KNN with TF-IDF only



KNN with TF-IDF N-grams



After applying cross validation on the four selected



confusion matrix after error analysis and Reducing champion model accuracy by nearly 30%