CE807 – Assignment 2 - Final Practical Text Analytics and Report

Student id: 2201010

Abstract

The report focuses on classifying the text in the OLID dataset using a range of classification techniques. The objective was to choose the two most suitable models for application. After a detailed analysis of various text categorization techniques, the study applied the OLID dataset to assess each technique's performance. Two models have been determined to have the best performance, and their use was noted. This study provides significant knowledge on text categorization techniques and shows how well they handle the OLID dataset.

1 Materials

The Google Colab code and Recorded presentation links are given below.

- Code.
- Google Drive Folder containing models and saved outputs
- Presentation

2 Model Selection (Task 1)

2.1 Summary of 2 selected Models

SVM is a powerful algorithm used for both linear and nonlinear classification. SVM is effective for handling high-dimensional data and can handle large datasets. Its main advantage is its ability to handle non-linear data by mapping it to a higher-dimensional space. However, SVM can be computationally expensive and can require more time and resources than other algorithms.
The F1 value of the SVM classifier reaches 86.26%, according to experimental data reported in the study (Liu et al., 2010), demonstrating the effectiveness of SVM as a machine learning method for text classification. SVM's usefulness is further illustrated by the findings, which indicate that it performs better than alternative classification techniques. Overall, the work emphasizes SVM's potential for text categorization and offers details on how it might be used in NLP.

Logistic Regression uses previous observations from a data set to predict a binary outcome, such as yes or no. A logistic regression model forecasts a dependent data variable by examining the correlation between one or more already present independent variables. It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions. Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets. Regularization (L1 and L2) techniques can be considered to avoid over-fitting in these scenarios.
The research paper I have studied revealed that Logistic Regression can generate better results in identifying and classifying text than any other methods like Naïve Bayes, Decision Tree, and others.

2.2 Critical discussion and justification of model selection

- According to the research paper (Liu et al., 2010), I have studied, The F1 value of the SVM classifier reaches 86.26%, demonstrating the effectiveness of SVM as a machine-learning method for text classification. SVM's usefulness is further illustrated by the findings, which indicate that it performs better than alternative classification techniques. The work emphasizes SVM's potential for text categorization and offers details on how it might be used in NLP. The SVM classifier pipeline is shown in figure 1
- The research paper (Indra et al., 2016), I have studied revealed that Logistic Regression can generate better results in identifying and classifying text than any other methods like Naïve Bayes, Decision Tree, and others. In a study (Indra et al., 2016), a set of features vectors was used for the task of classifying tweets into

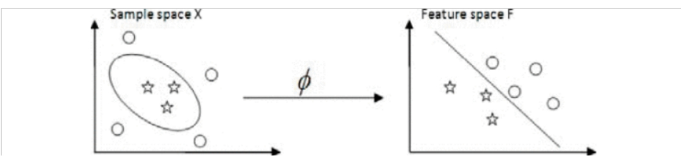

Figure 1: Non-linear SVM

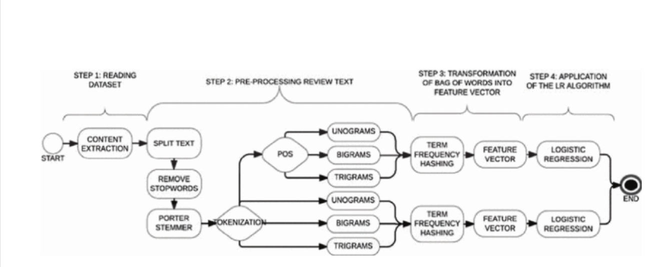

Figure 2: Logistic Regression

four topics (health, music, sport, and technology) using the Logistic Regression algorithm. The trained classifier was tested with 1800 tweets, with 450 tweets for each topic. The accuracy of the classification was evaluated using the Confusion Matrix technique. The results indicated that the classifier was able to classify tweets into the selected topics with an accuracy of 92%, which is considered to be very high and the pipeline of logistic regression is shown in figure 2

3 Design and implementation of Classifiers (Task 2)

In this section, you should add

- The given data set has train, test, and valid files which have a total of 12313 values having 33.23% of offensive tweets, 927 valid values having 33.22% of offensive tweets, and 860 test values out of which 27.9% of offensive tweets. More detail is shown in the table 1.

Dataset	Total	% OFF	% NOT
Train	12313	33.23	66.76
Valid	927	33.22	66.77
Test	860	27.9	72.0

Table 1: Dataset Details

- Implementation of SVM Classifier:
  After Mounting the google drive and initializing Gdrive, data, and file paths, we are all set to work on our models by importing the necessary files. First, the training file is loaded as a train data frame. From sklearn.model_selection, train_test_split and pandas is imported to split our given train data set. The training dataset is split into 25,50 and 75 respectively using stratification.
  The split files are then saved into Gdrive as csv files in the respective path.

- The dataset is then prepared by calling the function 'prepare_dataset', having three arguments: data, count_vectorizer, and split. This pre-processing step removes Twitter-specific elements from the text and the CountVectorizer transforms the text into tokens.
  We call the function to train a linear SVM(support vector machine) classifier on a set of text data. The classifier will separate the data points with a linear decision boundary using a linear kernel. The trained classifier then can be used to predict the labels of new text data.
  The pickle module is called to use as serialize the objects and write them in the disk as binary files as the trained model and vectorizer objects are saved to disk to be loaded and used later.

- The model taught by (Shekhar, 2022) taken as reference is loaded to train as the train and valid data are called by the function 'pd.read.csv()', and the function trains a linear SVM on text data using a train-validation split. After saving it to the disk, the performance is evaluated on both train and valid data using F1-score using the compute_performance function. The F1-score and accuracy are calculated for each set of 25, 50, 75, and 100 % split of data.
  The test_method is used to test method 1 by reading the data and loading the model from the disk, then preparing the data, and the computation is done for the F1 score and then the performance metrics are saved to the disk and printed to the console.
  The performance of the model(F1-Score) is computed on the entire dataset, followed by 75, 50, and 25. The results(F1 Score and accu-

racy) are printed after loading the model, and the final outputs of each split are saved to the disk respectively.

- The second model taught in lab by (Shekhar, 2022) is taken as a reference for logistic regression model, after studying various research papers, this model is selected so as to provide justice to the given OLID dataset.
  Since the dataset is loaded, preprocessed, and splitted already for the first model, the logistic model has to be trained in order to work on the data.
  Logistic Regression is imported from the sklearn.linear_model. The model is saved and loaded in order to perform the classification on the given dataset.
  The model is loaded to train as the train and valid data are called by the function 'pd.read.csv()'. and the function trains a linear SVM on text data using a train-validation split. After saving it to the disk, the performance is evaluated on both train and valid data using F1-score using the compute_performance function. The F1-score and accuracy are calculated for each set of 25, 50, 75, and 100 % split of data.
  The test_method is used to test method 1 by reading the data and loading the model from the disk, then preparing the data, and the computation is done for the F1 score and then the performance metrics are saved to the disk and printed to the console.
  The performance of the logistic regression model(F1-Score) is computed on the entire train dataset, followed by 75, 50, and 25. The results(F1 Score and accuracy) are printed after loading the model, and the final outputs of each split are saved to the disk respectively.

- Finally, the F1-Scores of SVM and a logistic regression model are shown in the table 2.

Model	F1 Score
Model 1	0.7179029061165239
Model 2	0.7204537302725968

Table 2: Model Performance

4 Data Size Effect (Task 3)

In this section, you should add

  - The dataset of different sizes are taken into consideration. 25% of the dataset has 1154 tweets, out of which 32.35% are labeled as offensive and 66.81% tweets are considered as not offensive. 50% of the dataset has 4617 tweets, out of which 33.26% tweets are offensive and 66.77% tweets are not offensive. 75% of the dataset has 9234 tweets, in which 33.24% of tweets are offensive and 66.76% are not offensive. 100% dataset has 12313 tweets in total, out of which 33.23% of tweets are offensive and 66.76% of tweets are labeled as not offensive. The overall data is represented in the below table 3

Data %	Total	% OFF	% NOT
25%	1154	32.35	66.81
50%	4617	33.26	66.77
75%	9234	33.24	66.76
100%	12313	33.23	66.76

Table 3: Train Dataset Statistics of Different Size

  - The performance of the classifiers on different-sized data sets is analyzed here. The F1 Score when applied on 100% of data gave 0.88196 for train and 0.71757 for the valid split. F1 score on 75% of the dataset is 0.9026585 and 0.67262 for train and valid split. 50% of the dataset's F1 score for train split is 0.950211 and the valid is 0.67955, for the 25% of data, the train split got 0.99413 and the valid split got 0.631629 respectively. The testing performance has given the following F1 Scores for 100% 0.71790 with an accuracy of 0.781395, for 75%, 0.687016 with an accuracy of 0.74186, for 50%, 0.687156 with an accuracy of 0.753488.
  - The plot is drawn in figure 3, to show the comparison of the performance of the method 1 f1 score on the test and valid dataset of different splits.
  - Similarly, for model 2 the F1 Score for the different splits of the dataset is seen. The F1 Score for the testing dataset of 100% yielded 0.7204 with an accuracy of 0.79767, 75% has a score of 0.72429, 50% of the dataset has got 0.70487 and 25% of the dataset has 0.638703.

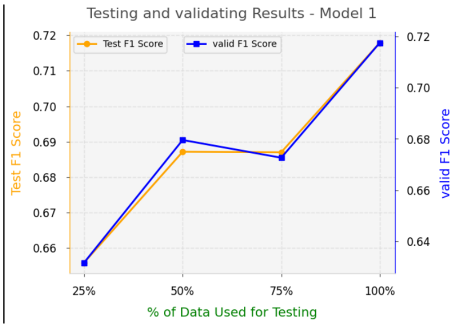

Figure 3: Comparision of Model 1 validation and testing based on Different data sizes.

- The whole F1 score comparison of model 2 is plotted in the below figure 4.

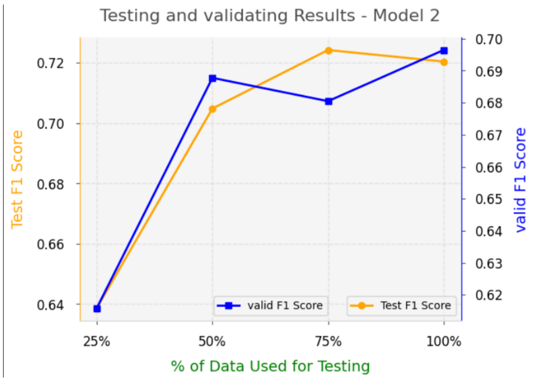

Figure 4: Comparision of Model 2 validation and testing based on Different data sizes.

Table 4 demonstrates the offensive and not offensive tweets in the given dataset based on the GT(Ground Truth) comparing with model 1 and model2 with some of the examples. It is noticed that some examples having not offensive labels in GT are showing as offensive in model 1 & model 2 and vice-versa. While a few examples show exactly the same as the GT.

Table 5 projects the comparison of model 1 among output files of different sizes of data sets. It compares with GT and 25%, 50%, 75%, and 100% of the data. It is seen that from the table, one example having Offensive in GT is predicted as not offensive in 25% of data and 75% of data and predicted exactly the same in the 50% and 100% of the dataset.

Table 6 shows the comparison of model 2 with output files of model 2 with different sizes of the dataset. As it is seen in model 1, two examples which are labeled as offensive are offensive in two sets of data, while shows not offensive in other sets of output files of the data.

5 Summary (Task 4)

The two models SVM and Logistic Regression are carried out to classify the given OLID dataset, after studying various research papers. The classifiers are implemented after cleaning and preprocessing the data, and vectorization is done. The classification is done and by seeing the performance score of both the models, Logistic Regression outperforms the SVM, as it has the F1 score of 0.720453 with an 0.79767 accuracy, whereas SVM has the F1 score of 0.71790 with an accuracy of 0.78139. Although there is a slight difference between two the classification models, the logistic model having the more value is considered the best classification model for the OLID dataset.

5.1 Discussion of work carried out

After preprocessing of dataset and splitting, the model is trained, validated and tested, and stored on disk to classify the given OLID dataset, after studying various research papers. After the classification is done and by seeing the performance score of both the models, Logistic Regression outperforms the SVM, as it has the F1 score of 0.720453 with an 0.79767 accuracy, whereas SVM has the F1 score of 0.71790 with an accuracy of 0.78139. There is a slight difference between two the classification models, the logistic model having the more higher value is considered the best classification model for the OLID dataset.

5.2 Lessons Learned

The classification models have to be chosen in such a way that they should classify the data maintaining the highest accuracy possible. It can only be possible by studying various classification models and analyzing which model will be applied to the particular dataset to get the best results.

There might be a few drawbacks in a particular model, but the model can perform better based on the dataset given.

Text classification is a vast topic, which needs to be studied more in order to understand the different models and their advantages as well as advantages based on the size of the data sets.

6 Conlusion

By conducting the classification methods, it is evident that logistic regression outperforms SVM in classifying the OLID dataset providing the better F1 Score than SVM. With this, Logistic Regression

Example %	**GT**	**M1(100%)**	**M2(100%)**
#HurricaneFlorence not good	NOT	NOT	NOT
#CheeringTheChutiyapa just because she is a woman.....LOL URL	NOT	NOT	NOT
NoPasaran: Unity demo to oppose the far-right in	OFF	NOT	NOT
Are you fucking serious?	NOT	OFF	OFF
EmmyAwards2018 - Ratings tank as expected.	OFF	NOT	NOT

Table 4: Comparing two Model's using 100% data: Sample Examples and model output using Model 1 & 2. GT (Ground Truth) is provided in the test.csv file.

Example %	**GT**	**M1(25%)**	**M1(50%)**	**M1(75%)**	**M1(100%)**
Liberalismisamentaldisorder,	OFF	NOT	OFF	NOT	OFF
StopKavanaugh he is liar like the rest of the GOP URL	OFF	OFF	OFF	OFF	OFF
ArianaAsesina? Is that serious?!	OFF	OFF	OFF	OFF	OFF
Are you fucking serious? URL	NOT	OFF	OFF	OFF	OFF
@USER Do you get the feeling he is kissing	OFF	NOT	NOT	OFF	NOT

Table 5: Comparing Model Size: Sample Examples and model output using Model 1 with different Data Size

Example %	**GT**	**M2(25%)**	**M2(50%)**	**M2(75%**
ConstitutionDay is revered by Conservatives, hated by Progressives	NOT	NOT	NOT	NOT
Are you fucking serious?	NOT	NOT	OFF	OFF
FortniteBattleRoyale XboxShare @USER Please ban this cheating scum.	OFF	NOT	NOT	OFF
Conservatives Govt ' @USER made my life hell':	OFF	NOT	NOT	NOT
Georgetown Classmate Says Left Accuser is Absolutely NUTS"	NOT	OFF	OFF	OFF

Table 6: Comparing Model Size: Sample Examples and model output using Model 2 with different Data Size

is concluded to be the better one among the both that are carried out.

References

ST Indra, Liza Wikarsa, and Rinaldo Turang. 2016. Using logistic regression method to classify tweets into the selected topics. In *2016 international conference on advanced computer science and information systems (icacsis)*, pages 385–390. IEEE.

Zhijie Liu, Xueqiang Lv, Kun Liu, and Shuicai Shi. 2010. Study on svm compared with the other text classification methods. In *2010 Second international workshop on education technology and computer science*, volume 1, pages 219–222. IEEE.

Dr Ravi Shekhar. 2022. Ce07 text analytics lab codes, referred for model implementation. MSc Data Science, University of Essex.