CE807 – Assignment 2 - Final Practical Text Analytics and Report

Student id: 2201735

Abstract

This study offers predictions regarding the offensiveness of the tweet data points while concentrating on the classification of offensive language in the OLID dataset. The speech classification method is divided into three parts. Two model selection strategies are first picked based on the features of the data set and literature evaluations. Second, a few classifier models are created using the specified training data set, and their performance is evaluated. The four subsets of the original dataset are then used to evaluate the performance of the developed models. Based on a comparison of model accuracy and F1 score, SGD Classifier is selected as the top model for text categorization of hostile languages on the dataset.

1 Materials

The shared link provides the documents and files used for this report,

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

2 Model Selection (Task 1)

The choice of a model for classifying objectionable speech is influenced by several factors, including data type, size, and type of attributes. Additionally, results from a study of the literature on offensive language classification models demonstrate that SGD classifiers and linear SVC classification models outperform other machine learning techniques.

2.1 Summary of 2 selected Models

A well-liked and effective approach for classifying texts is the linear support vector classifier (SVC)(Sulea et al., 2017), especially when dealing with high-dimensional and noisy data. It efficiently develops models from a data source with a wide range of attribute values. L2 regularisation(Cortes et al., 2012) is used by default in linear SVC, which helps to prevent overfitting and enhances generalization performance. Additionally, it works well in high-dimensional settings, trains more quickly, and is noise-resistant. Finding the ideal hyperparameters for a given dataset is simple with linear SVC.

Another well-liked text categorization approach is Stochastic Gradient Descent (SGD)(Yousaf et al., 2020). Because it is based on stochastic gradient descent optimization, which analyses the data one instance at a time, it performs well on large-scale text classification issues. It is one of the adaptable algorithms that can accommodate various penalty and loss functions while using various feature selection methodologies. It can manage millions of instances and features and is appropriate for big data applications. It employs an incremental learning process, so whenever a new data set is introduced, its models are immediately updated. It automatically applies L2 regularisation, which helps prevent overfitting and enhances generalization performance.

2.2 Critical discussion and justification of model selection

Three steps make up the overall process of classifying offensive speech. As input, tweets are provided, and these tweets are preprocessed to clean the data. In the preprocessing stage(Kadhim, 2018), the digits, stopwords, and punctuation are eliminated. The provided tweets are also tokenized and lowered.A Tfid vectorizer(Fautsch and Savoy, 2010)is used for feature extraction and vectorization. Vectorized tweets are then fed into classifier models as input. Models for linear SVC and SGD classifiers have been developed. The classification of offensive

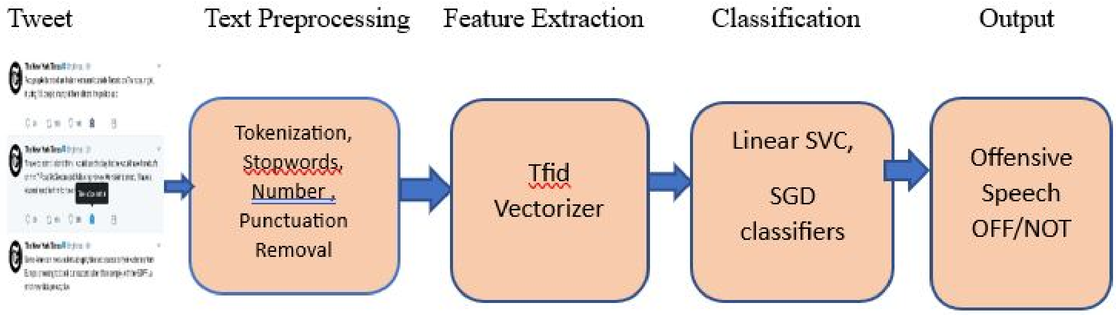

Figure 1: Classification Process Diagram.

speech is done using this model. 1

3 Design and implementation of Classifiers (Task 2)

The train set, valid set, and test set are three datasets sets. Id, tweet, and label are the three columns found in each of these three data sets. The model is trained using the train set to categorize the labels as NOT or OFF. It has a total of 12313 labels, 33.23 percent of which are turned OFF and 66.76 percent of which are NOT. The model's performance is evaluated using the valid set as it is being trained. There are 927 labels altogether, of which 33.22 percent are OFF labels and 66.77 percent are NOT labeled. The test set is then utilized to assess the model's actual performance. There are 860 instances in the test set, with 27.90 percent OFF labels and 80.23 NOT labels.

Dataset	Total	% OFF	% NOT
Train	12313	33.31	66.69
Valid	927	33.23	66.77
Test	860	27.90	72.10

Table 1: Dataset Details

Model1 in this study is based on the Linear SVC Classification methodology. A linear kernel is used in the specific SVM implementation known as linear SVC to conduct classification on a dataset. An SVM model's performance can be considerably impacted by the hyperparameter choice of the kernel function. The dot product of two input vectors is calculated by the linear kernel, a straightforward kernel. Packages required for model implementation are loaded from sci-kit-learn. The creation of a Linear SVC object and hyper parameterization follow. In which the parameter 'c' determines the trade-off between maximizing the margin and minimizing the classification error, and the parameter 'loss' stands for the loss function. 'Penalty' stands for the regularisation term. The model is built on the trained data set and saved. The saved liner SVC model is used to train the test set to find the actual performance of the classifier model.

SGD Classifier is used to generate Model 2. It is a flexible and effective technique for jobs involving text classification. L1 and L2 regularisation are just a couple of the regularisation techniques supported by SGD Classifier. By doing so, overfitting can be avoided, and the model's generalization abilities are enhanced. 'Penalty' is the phrase used to describe it. This model's "hinge" loss function is referred to as the "loss" parameter. When using SGD Classifier, the parameters of the model are iteratively updated to reduce the loss function, which calculates the discrepancy between the predicted and actual labels. The algorithm is quicker than conventional techniques thanks to this stochastic approach. The classifier is implemented using imported Sci-kit-Learn tools. After creating an SGD Classifier object, hyper parametrization is carried out by selecting the loss function as the hinge and penalty as 12. The model is built on the trained data set and saved. This saved model is used to train the test set to evaluate the actual performance of the model.

A 100%dataset is used to compare two models in Table 2. It is evident from the data that Model 2 has the highest F1 Score, at 0.7206. SGD Classifier outperformed the Linear SVC model in terms of performance. Model 1 currently has an F1 score of 0.7018. The performance of the model differs significantly. Model 2 is more accurate than Model 1 in terms of accuracy. Model 1 is 0.7686 percent accurate, whereas Model 2 is 0.7860 percent accurate. Table 5 compares the performance of the two models utilizing 100% data, sample examples, and model output for the two models that were chosen. Except for a few fields, the majority of the forecasts made by both models are identical.

Model	F1 Score
Model 1	0.70182123
Model 2	0.72069247

Table 2: Model Performance

4 Data Size Effect (Task 3)

The given dataset is divided into four different subsets,25%,50%,75%, and 100%. The 25% dataset contains a total of 3079 values, in which 33.91% is categorized as OFF label and 66.09% is categorized as NOT label. Similarly, the 50%,75%, and 100% datasets contain 6157,9235 and 12313 samples respectively. Moreover, in all four datasets, the percentage of OFF and NOT labels are nearly equal and all values are tabulated in Table 3.

The identical preparation and data-cleansing procedures outlined in Task 2 are applied to all four datasets. For Model 1, the F1 score for the 25% dataset was 0.67185, the 50% dataset was 0.68772, and the 75% dataset was 0.68351. The 100% dataset's F1 score is 0.71699. The findings demonstrate that data amount significantly affects model performance. The score increases with the reported data size. As a result, the 25 percent dataset had the lowest F1 score while the 100 percent dataset had the highest F1 score. The 25 percent, 50 percent, 75 percent, and 100 percent datasets for Model 2 received F1 Scores of 0.69791, 0.68315, 0.734983, and 0.72069, respectively. All different size datasets in Model 2—with the exception of 75%of datasets—had their F1 scores increase.The dataset with a 75 percent accuracy rate has the highest accuracy score, 0.73498. It demonstrates that, in contrast to Linear SVC, the SGD classifier is less affected by the size of the data set.

Data %	Total	% OFF	% NOT
25%	3079	33.91	66.09
50%	6157	32.73	67.27
75%	9235	33.11	66.89
100%	12313	33.31	66.69

Table 3: Train Dataset Statistics of Different Size

The following figures plot model performance comparison on the validation set and test set. In Figure 2, Model performance on Test Dataset based on F1 Score is depicted and in Figure 3, Model evaluation is done based on the accuracy score on the validation dataset.

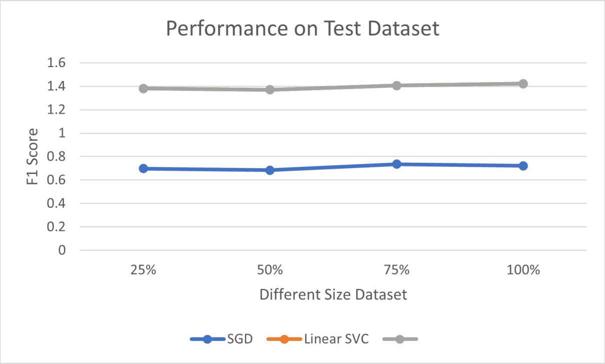

Figure 2: Comparision of Models based on Different data sizes Vs F1 Score on Test Dataset.

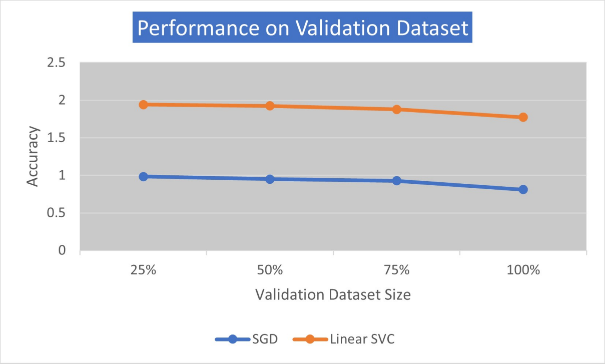

Figure 3: Comparision of Models based on Different data sizes Vs Accuracy on Validation Dataset.

5 Summary (Task 4)

5.1 Discussion of work carried out

In this report offensive text classification is done based on Linear SVC and SGD Classifier models. The given OLID dataset is large and has a higher number of features. The classification of objectionable text in this report is based on the Linear SVC and SGD Classifier models. The provided OLID dataset is sizable and contains more features. Tfid Vectorizer is used to extract the features from the text input file. On the training dataset, linear SVC and SGD classifiers are developed. The performance of this model is then assessed using datasets of varying sizes. The models are assessed using the F1 Score and Accuracy. The SGD classifier received the test set's highest F1 score of 0.7206.

Example %	GT	M1(100%)	M2(100%)
BoycottLeviStrauss for supporting leftist gun control orgs	NOT	NOT	NOT
WCW goes to my beautiful and sexy @USER Without you I would be lost, you have brought me from the dark and into the light	NOT	NOT	NOT
Levi Strauss has “never been afraid to take an unpopular stand to support a greater good.	NOT	OFF	NOT
ConfirmKavanaughKavanaugh SCOTUS POTUS45 PresidentTrump AmericaFirst MAGAKAG Anti Trump liberal activist Accuser	OFF	NOT	NOT
Does anyone want to bet it costs more than that 50k election referendum?	NOT	NOT	NOT

Table 4: Comparing two Model’s using 100% data

Example %	GT	M1(25%)	M1(50%)	M1(75%)	M1(100
DECLASFISA Democrats Support Antifa, Muslim Brotherhood, MS13, ISIS, Pedophilia, Child Trafficking,	OFF	OFF	OFF	OFF	OFF
ConstitutionDay is revered by Conservatives , hated by Progressives/Socialist/ Democrats that want to change it	NOT	OFF	NOT	NOT	NOT
fakereporting THESWAMP dnc	NOT	NOT	NOT	NOT	NOT
NoPasaran: Unity demo to oppose the far-right in London –	NOT	NOT	NOT	NOT	NOT
What the fuck did he do this time?	OFF	NOT	NOT	NOT	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%)	M2(10
Always smack URL	OFF	OFF	OFF	OFF	OFF
RAP is a form of ART! Used to express yourself freely.	NOT	NOT	OFF	NOT	NOT
5 Tips to Enhance Audience Connection on Facebook URL	NOT	NOT	NOT	NOT	NOT
Conservatives - the party of low taxation Tories Tory URL	NOT	NOT	NOT	NOT	NOT
ConsTOO THE PLACE FOR FED UP CONSERVATIVES !!!	OFF	NOT	NOT	NOT	NOT

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

A similar F1 Score of 0.7018 has been assigned to Linear SVC. The findings indicate that both models perform well across all datasets.

5.2 Lessons Learned

The study of the results shows that data amount significantly affects the model’s performance. The classifier models that were chosen had the best training results on the huge dataset. The Stochastic Gradient Descent (SGD) classifier outperforms the Linear SVC Classifier when comparing the two models. However, for smaller datasets Linear SVC performed well, especially in the 75% dataset. In the SGD classifier, the data is analyzed one instance at a time and it is appropriate for big data applications. In terms of accuracy, in both models accuracy is decreased with the increasing size of datasets.

6 Conlusion

To conclude, the distribution and size of the dataset are directly linked to model performance. Feature extraction and data cleaning are other factors that affect model accuracy and F1 Score. The Stochastic Gradient Descent (SGD) classifier performed well in overall datasets.

References

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2012. L2 regularization for learning kernels. *arXiv preprint arXiv:1205.2653*.

Claire Fautsch and Jacques Savoy. 2010. Adapting the tf idf vector-space model to domain specific information retrieval. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1708–1712.

Ammar Ismael Kadhim. 2018. An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(6):22–32.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Anam Yousaf, Muhammad Umer, Saima Sadiq, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi. 2020. Emotion recognition by textual tweets classification using voting classifier (lr-sgd). *IEEE Access*, 9:6286–6295.