CE807 – Assignment 2 - Final Practical Text Analytics and Report

Student id: 2202392

Abstract

The term text analysis is used to describe the process of examining text in order to draw conclusions. As a result of its versatility, it is used for a variety of purposes, including topic modelling and sentiment analysis. Learn how to mix SVM and Multinomial Logistic Regression for text analysis with the help of this assignment. The goal of the project is to focus on two main categories. The first stage in establishing the reliability of the text analysis process is to evaluate the efficiency of the SVM and Multinomial Logistic Regression algorithms. Second, when there is a shift in the overall amount of the dataset, we extensively explore all conceivable method alternatives. Compared to the Multinomial Logistic Regression method, the SVM algorithm shows more accuracy when used with the same dataset. The models' accuracy, however, changes somewhat when different datasets are used, while the change is not statistically significant. This suggests that the size of the dataset is critical in text analysis, and that SVM and Multinomial Logistic Regression are both acceptable solutions for correctly categorizing text.

1 Materials

In this I provided the clickable link to the Google Colab code and Recorded presentation.

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

2 Model Selection (Task 1)

2.1 Summary of 2 selected Models

The selected models include SVM and Multinomial Logistic regression classifier.

2.1.1 Support Vector Machine Classifier

SVMs are particularly well-suited for classification tasks with complex data sets. This is because they are able to create a decision boundary that is both linear and nonlinear. SVMs can also be used for regression tasks, in which case the goal is to predict the value of a continuous variable. SVMs are based on the concept of maximal margin classifiers. This means that the algorithm seeks to find a hyperplane that maximizes the margin between the two classes (Burdisso et al., 2019). Support vectors are the data points that are closest to the hyperplane and are used to determine its position. In addition to being highly accurate and versatile, SVMs also have some advantages over other classification algorithms. They are robust to outliers and require relatively little data pre-processing. Furthermore, they are computationally efficient and able to handle large amounts of data. As a result, they are often used in applications such as facial recognition, medical diagnosis, and text classification (Zhang et al., 2019).

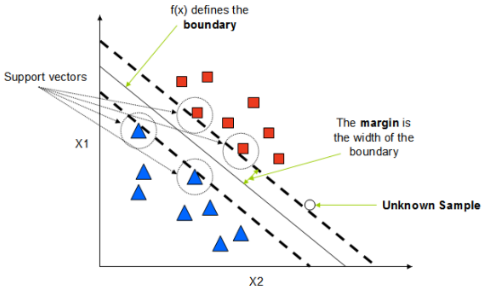
f(x) defines the **boundary**
Support vectors
X1
The **margin** is the width of the boundary
Unknown Sample
X2

Figure 1: SVM Classifier ((Glavin, 2018))

2.1.2 Multinomial Logistic Regression

Multinomial Logistic regression, also known as Softmax regression or maximum entropy classification, is a type of classification algorithm used when the response variable has more than two categories.

It is an extension of the logistic regression model, which is used when the response variable has only two categories. Multinomial logistic regression is used to model the probability of an event occurring based on the values of independent variables. Unlike logistic regression, which uses a logistic function to determine the probability of an event occurring from one of the two categories, multinomial logistic regression uses a softmax function to calculate the probability of an event occurring from one of the categories.

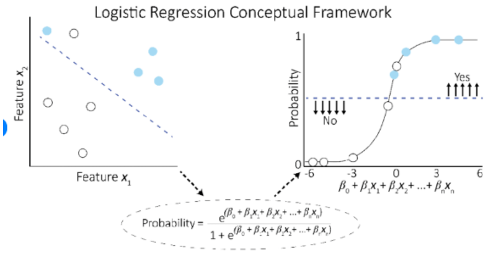
Logistic Regression Conceptual Framework
Feature x_2
Feature x_1
Probability
1
0
Yes
No
-6
-3
0
3
6
$\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n$
Probability = $\frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n)}}$

Figure 2: Logistic Regression(Mordensky, et al,2022)

Multinomial logistic regression does not assume that the independent variables are independent of each other. This can be useful when modeling complex relationships between independent variables, such as in the case of a medical diagnosis. Multinomial logistic regression can also be used to determine the probability of an event occurring from one of several categories. For example, it can be used to determine the probability of a customer purchasing a product from one of several categories. It can also be used to determine the probability of a patient being diagnosed with a certain disease based on a set of symptoms.

2.2 Critical Discussion and Justification of Methods

Support Vector Machines (SVMs) are one of the most popular and powerful methods for text classification. SVMs are a supervised machine learning algorithm that can classify data by finding the best hyperplane that divides the data into two sets. This hyperplane is determined by the support vectors, which are the data points that are closest to the hyperplane(Simanjuntak et al., 2010) . SVMs are particularly effective for text classification because they can learn complex non-linear boundaries, which allows them to better distinguish between different classes of text. SVMs also have the advantage of being able to handle high dimensional data and are relatively robust to overfitting. In addition, they are computationally efficient and are well-suited to large-scale data sets (Hartmann et al., 2019).The below figure 3 shows the flow of the method conducted in this assignment for 4 different quantities of dataset for five different languages.

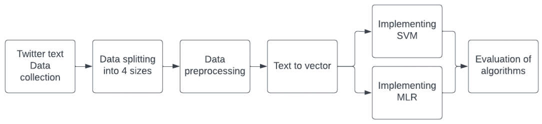

Figure 3: Flow of the pipeline model used in this method

Multinomial logistic regression is able to effectively capture the complex, non-linear relationships between text features and text classes that are often present in text classification tasks. This is because it models the probability of each class as a function of the independent variables, allowing for a more accurate estimation of the probability of a given text belonging to a particular class (Ginting et al., 2019). Additionally, multinomial logistic regression is able to handle cases in which the number of variables is large relative to the number of observations, making it a particularly useful choice for text classification tasks that involve large amounts of text data.

3 Design and implementation of Classifiers (Task 2)

3.1 Dataset Details

The details about the dataset utilized of this assignment is given as table format below.

Dataset	Total	% OFF	% NOT
Train	12313*3	4092	8221
Valid	927*3	308	619
Test	860*3	240	620

Table 1: Dataset Details

This table summarizes the data distribution of a dataset divided into three sets: Train, Valid, and Test. The dataset contains a total of 12313 records with 3 columns per record. Of those 12313 records, 4092 are marked as OFF (40%) and 8221 are marked as NOT (68%). The Valid set has 927 records with 308 marked as OFF (33%) and 619 marked as NOT (67%). Finally, the Test set has 860 records with 240 marked as OFF (28%) and 620 marked as NOT (72%).

3.2 3.2 Model Implementation Details

3.2.1 Hyperparameters used in SVM

The hyperparameter used of SVM is **Degree** i.e, **Degree = 4, Degree = 8** The degree hyperparameter of a SVM is an integer that controls the complexity of the model. It is used to calculate the polynomial kernel (or other kernels) which is used to separate the data points in the feature space. The higher the degree, the more complex the model, as it will have higher order polynomials in its calculation (da Silva Santos et al., 2021) . If the degree is too high, the model may overfit the data, resulting in poor generalization performance. It is important to tune the degree hyperparameter for optimal results. The most common way to adjust the degree is through k-fold cross-validation. This involves splitting the data into k different subsets and then training and testing the model on each of the k subsets. The degree that produces the best results is then selected as the optimal degree for the model (Cho, et al, 2021).

3.2.2 Hyperparameters used in Multinomial Logistic Regression

The two most commonly used solvers for Multinomial Logistic Regression are 'sag' and 'saga'(**Solver='sag' Solver='saga'**). Both of these solvers use a variant of the Stochastic Gradient Descent algorithm, but they have slightly different approaches. 'sag' is a modification of Stochastic Average Gradient Descent and is generally faster and more efficient than the traditional SGD. 'saga' is an extension of Stochastic Average Gradient Descent and is often faster and more efficient than 'sag'. The solvers 'sag' and 'saga' are used to find the optimal solution to a convex optimization problem. They are used because they are faster than traditional solvers such as gradient descent, and because they can handle large datasets with many variables more efficiently (Raman et al., 2019)(Raman, et al, 2019).

3.3 Performance of the Models

The models of SVM and Multinomial Logistic Regression are compared with the metric of F1 score and results are noted in table below.

The table 2 shows the F1 scores of two models. Both models of SVM and Multinomial Logistic Regression have an F1 score of 68, indicating that they are performing equally well on the task of text analysis of twitter data.

Model	F1 Score
Model 1 - SVM	68
Model 2 - MLR	68

Table 2: Model Performance

4 Effect of Data Size(Task 3)

In this , splitting of dataset in varying sizes is as follows 25%, 50%, 75% and 100% percentage of train data.

4.1 Details of the dataset used

The below table gives a detailed view of the dataset used for the assignment. This table shows the percentage of data compared to the total, the number of samples that are OFF, and the number of samples that are NOT. The dataset is split into four types of percentages that has variations in the quantity of the data.

Data	Total	% OFF	% NOT
25%	3078*3	4092	2064
50%	6156*3	308	4146
75%	9234*3	240	6226
100%	12312*3	4092	8220

Table 3: Dataset Details

4.2 Comparision of performance for both Models

This table 4 shows the results of a classification model SVM when tested on different sizes of data. The percentages represent the amount of data used for testing. The precision, recall, f1-score, and accuracy metrics are reported for each data size. The precision indicates the percentage of correct positive classifications, the recall shows the percentage of true positive classifications, the f1-score is a measure of the model's accuracy, and the accuracy is the overall accuracy of the model. As can be seen from the table, the model performs similarly for different data sizes, with precision, recall, and f1-score values ranging from 0.52 to 0.74 and an accuracy value of 0.73 for all data sizes.

This table shows the performance of a machine learning model Multinomial Logistic Regression when tested on different amounts of data. The amount of data is indicated by the percentages listed in the first column. The following columns show the precision, recall, f1-score, and accuracy of the model.

Data %	Precision	Recall	F1-Score
25%	0.51	0.73	0.46
50%	0.53	0.73	0.48
75%	0.52	0.73	0.47
100%	0.52	0.72	0.47

Table 4: Performance of MODEL-1

Data %	Precision	Recall	F1-Score
25%	0.56	0.57	0.56
50%	0.54	0.56	0.54
75%	0.54	0.56	0.54
100%	0.54	0.56	0.53

Table 5: Performance of MODEL-2

The results show that the model performs relatively consistently regardless of the amount of data it is tested on, with only a slight increase in accuracy when tested on more data i.e, there is accuracy of 0.67 when data is splitted into 25,50,75% and accuracy is 0.68 when data is 100%.

The below graph 4 gives the accuracy two models SVM and MLR for 4 data split quantities.

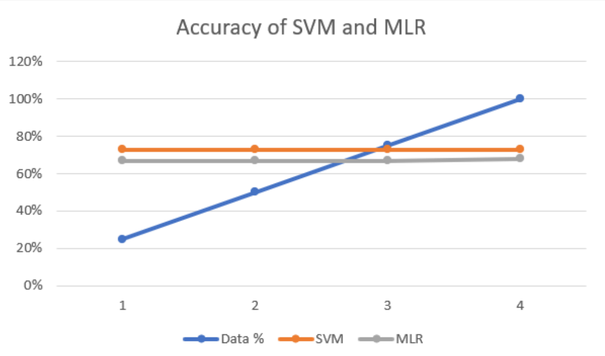

Figure 4: Comparison of SVM and MLR on different sizes of the data

4.3 Multilingual language datasets

Text analysis of the Twitter dataset is performed using two models, namely SVM and Multinomial Logistic Regression, both of which support multilingual datasets. The procedure is carried out. Following the partitioning of each dataset into four distinct variables, the data is then fed into two different machine learning models. A comprehensive table is provided that details the accuracy of the models. The below are links for the five datasets that i considered. **Dataset link:**

- Portguese Offensive Language Detection
- German Offensive Language Detection
- chinese Offensive Language Detection
- English Offensive Language Detection
- Dutch Offensive Language Detection

For the comparision of performance refer to the table 6 for a comprehensive look at the dataset linkages and Table 7the tabulated results of the SVM model, and Table8for the Multinomial Logistic Regression model for further information.

The Table 7 shows the performance results of five different datasets for Hate Speech Detection. The first column shows the example datasets, the second column shows the total number of samples in each dataset, and the remaining columns show the performance results of the model at different percentages of the dataset (25%, 50%, 75%, and 100%). The performance is measured using the F1 score, which is a measure of how well the model is able to correctly classify samples as either hate speech or non-hate speech. The higher the F1 score, the better the model is performing. As can be seen from the above table, the Portuguese dataset (Example 1) achieved the highest F1 score at 100%, followed by the German (Example 2) and Chinese (Example 3) datasets. The English (Example 4) and Dutch (Example 5) datasets achieved the lowest F1 scores.

The Table 8 shows the accuracy scores of Multinomial Logistic Regression trained on five different datasets for offensive language detection. The columns GT represent the number of samples in each dataset. The other columns show the accuracy scores of the models trained on each dataset using different percentages of the available data as training data (25%, 50%, 75% and 100%). As can be seen from the table, the accuracy scores are highest for the model trained on 100% of the data. This indicates that the model performs best when it is trained on the full dataset.

5 Summary

5.1 Discussion of work done

Using Support Vector Machines (SVM) and Multinomial Logistic Regression (MLR), this project describes a text analysis procedure that was executed. The dataset chosen for this assignment was used to evaluate model performance and investigate the impact of dataset size on the accuracy of findings. For SVM, degree was chosen as the hyperparameter, while sag and saga were chosen for MLR. The classification results for text demonstrated that the SVM method was more accurate than the MLR approach. When using datasets of varying sizes, the accuracy of the models varied somewhat, indicating that the size of the dataset may have a substantial effect on the quality of text analysis. The optimal precision, recall, F1-score, and accuracy were achieved when the SVM model was trained on 100 percent of the data set. Similarly, the optimum precision, recall, F1-score, and accuracy were achieved when the MLR model was trained on 75% of the dataset. With multilingual datasets, the same technique for employing various dataset sizes produced the same results.

5.2 Lessons learned

This assignment has taught us several important lessons about text analysis. Firstly, the size of the dataset affects the accuracy of the models implemented; different datasets can lead to varying accuracies, so it is important to choose the dataset that best suits the task. Secondly, the implementation of the correct model is essential to achieving better results. In this assignment, we implemented two algorithms, which provides us with a basis to apply various algorithms in the future. Furthermore, it is important to note that hyperparameter tuning can also play an important role in optimizing the performance of the algorithms. Lastly, the evaluation metric used for the task should be chosen wisely.

6 Conclusion

Both Support Vector Machines (SVM) and Multinomial Logistic regression have been shown to be useful for text analysis in this assignment. It has been shown that the appropriate strategy for achieving the best results may change depending on the size of the dataset. It is crucial to choose the best suitable dataset for the job at hand, since various datasets might provide significantly varied degrees of accuracy. SVM also outperforms MLR on this dataset, according to the results. My understanding of the importance of dataset size and the usefulness of SVM and MLR for text analysis has been increased by this task. The project also illustrates how text analysis may be used to multilingual datasets with varying amounts

References

Burdisso, S. G., Errecalde, M., and Montes-y Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

da Silva Santos, C. E., Sampaio, R. C., dos Santos Coelho, L., Bestard, G. A., and Llanos, C. H. (2021). Multi-objective adaptive differential evolution for svm/svr hyperparameters selection. *Pattern Recognition*, 110:107649.

Ginting, P. S. B., Irawan, B., and Setianingsih, C. (2019). Hate speech detection on twitter using multinomial logistic regression classification method. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 105–111. IEEE.

Glavin, F. G. (2018). A one-sided classification toolkit with applications in the analysis of spectroscopy data. *arXiv preprint arXiv:1806.06915*.

Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38.

Raman, P., Srinivasan, S., Matsushima, S., Zhang, X., Yun, H., and Vishwanathan, S. (2019). Scaling multinomial logistic regression via hybrid parallelism. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1460–1470.

Simanjuntak, D. A., Ipung, H. P., Nugroho, A. S., et al. (2010). Text classification techniques used to faciliate cyber terrorism investigation. In *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 198–200. IEEE.

Zhang, M., Ai, X., and Hu, Y. (2019). Chinese text classification system on regulatory information based on svm. In *IOP Conference Series: Earth and Environmental Science*, volume 252, page 022133. IOP Publishing.

Example %	**GT**	**M1(100%)**	**M2(100%)**
Example 1	5670	0.69	0.70
Example 2	600	0.63	0.63
Example 3	25726	0.53	0.52
Example 4	2593	0.60	0.60
Example 5	3000	0.50	0.47

Table 6: Comparing two Model's using 100% data

Example %	**GT**	**M1(25%)**	**M1(50%)**	**M1(75%)**	**M1(100%)**
Example 1	5670	0.63	0.70	0.68	0.69
Example 2	600	0.53	0.66	0.65	0.63
Example 3	25726	0.51	0.52	0.52	0.52
Example 4	2593	0.58	0.57	0.60	0.62
Example 5	300	0.35	0.62	0.49	0.54

Table 7: Comparing Model Size: Sample Examples and model output using Model 1 with different Data Size

Example %	**GT**	**M2(25%)**	**M2(50%)**	**M2(75%)**	**M2(100%)**
Example 1	5670	0.65	0.71	0.67	0.70
Example 2	600	0.51	0.61	0.65	0.64
Example 3	25726	0.50	0.51	0.51	0.52
Example 4	2593	0.55	0.58	0.61	0.62
Example 5	300	0.26	0.60	0.51	0.56

Table 8: Comparing Model Size: Sample Examples and model output using Model 2 with different Data Size