CE807-TEXT ANALYTICS

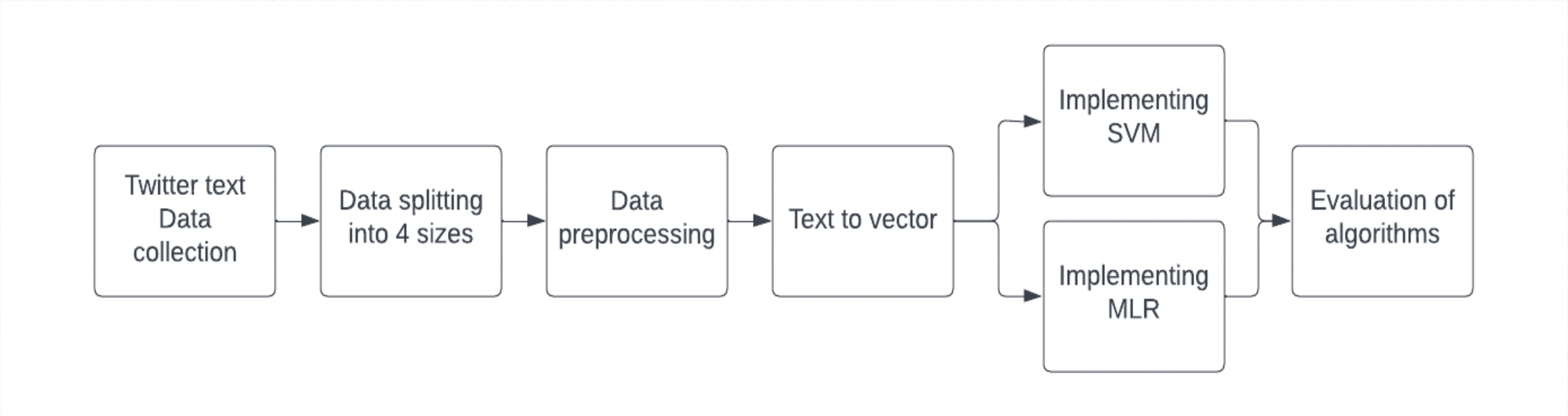
FINAL PRATICAL TEXT ANALYTICS REPORT

Student id: 2202392

TEXT ANALYSIS

- Text analysis is the process of analyzing text to extract useful insights and information.There are several techniques for text analysis, including natural language processing (NLP), machine learning, and data visualization.
- This assignment focuses on two main categories: evaluating the efficacy of SVM and Multinomial Logistic Regression algorithms, and exploring alternative methods.
- Results suggest that the size of the dataset is critical in text analysis, and that SVM and Multinomial Logistic Regression are both acceptable solutions.

Flow of pipeline used

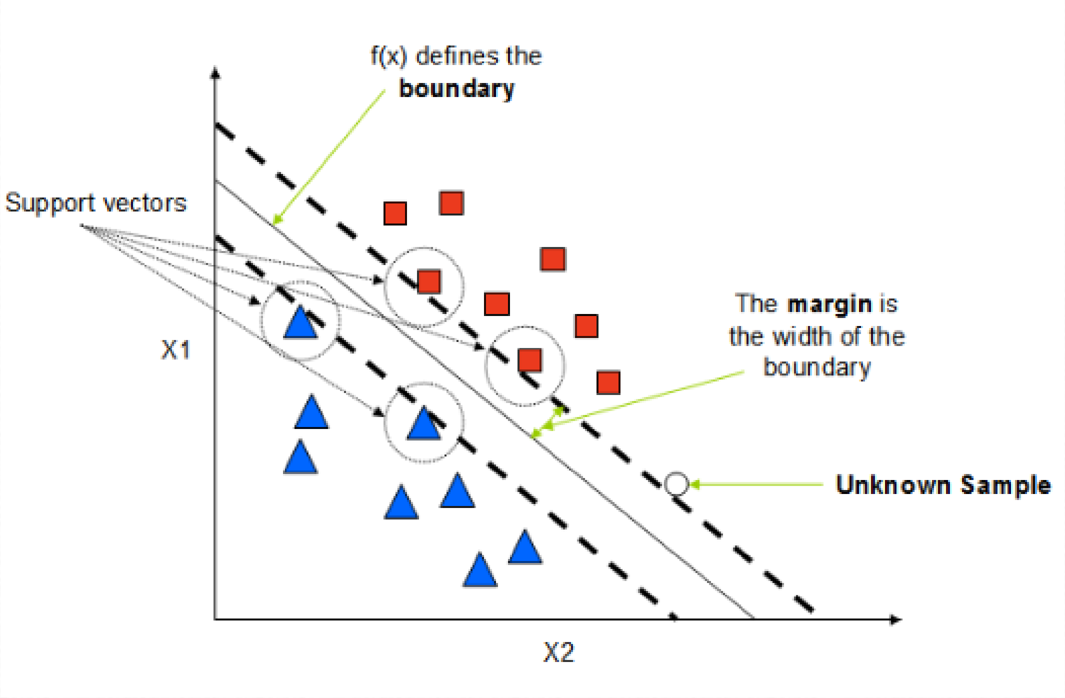


Twitter text Data collection → Data splitting into 4 sizes → Data preprocessing → Text to vector → Implementing SVM / Implementing MLR → Evaluation of algorithms

Support Vector Machine

- SVMs are well-suited for classification tasks with complex data sets due to their ability to create a decision boundary that is both linear and nonlinear.
- They are based on the concept of maximal margin classifiers, which seeks to find a hyper plane that maximizes the margin between the two classes.
- SVMs are robust to outliers and require relatively little data pre-processing, and are computationally efficient and able to handle large amounts of data.
- They are used in applications such as facial recognition, medical diagnosis, and text classification.

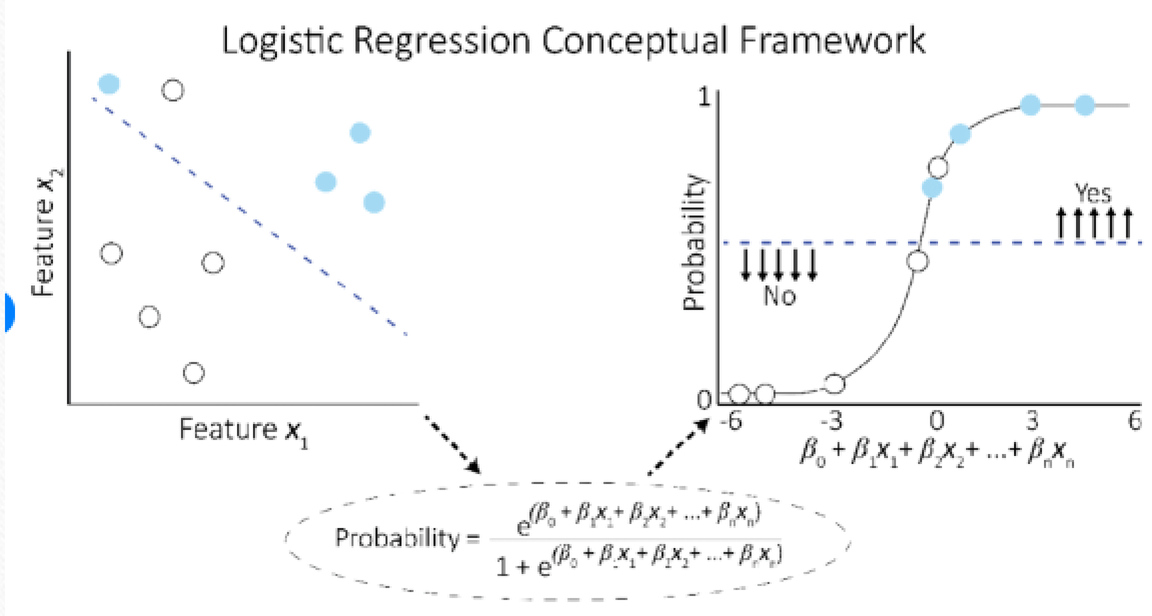
SVM Algorithm



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 uses a logistic function to determine the probability of an event occurring from one of the two categories.
- Multinomial logistic regression does not assume that the independent variables are independent of each other, making it useful for modeling complex relationships between independent variables.
- It can also be used to determine the probability of an event occurring from one of several categories, such as a customer purchasing a product or a patient being diagnosed with a certain disease based on a set of symptoms.

Multinomial Logistic Regression



Logistic Regression Conceptual Framework

Feature x_2
Feature x_1
Probability
1
0
-6
-3
0
3
6
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$
Yes
No
Probability = $\frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$

Justification of Models - 1

- Support Vector Machines (SVMs) are one of the most popular and powerful methods for text classification.
- SVMs are supervised machine learning algorithms that can classify data by finding the best hyper plane that divides the data into two sets.
- They are particularly effective for text classification because they can learn complex non-linear boundaries, handle high dimensional data, and are computationally efficient.

Justification of Models - 2

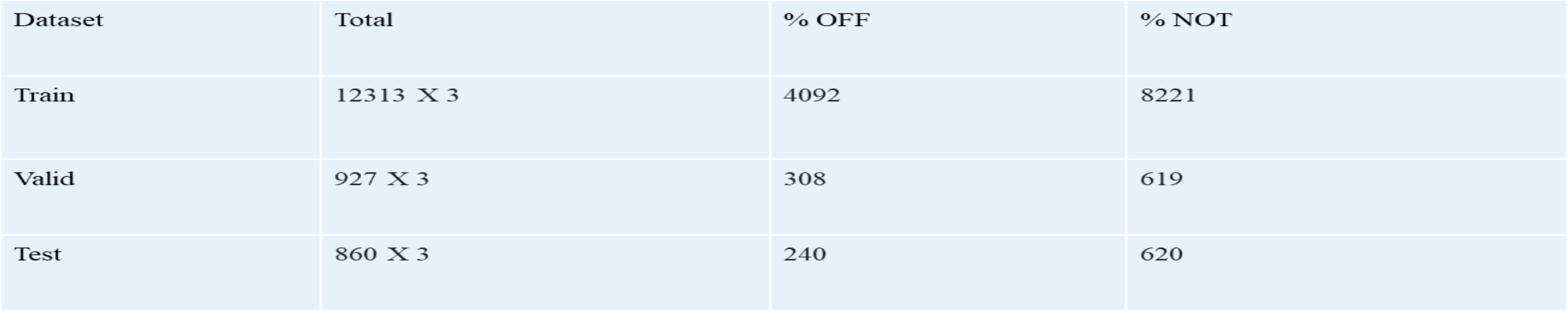
- Multinomial logistic regression is able to capture the complex, non-linear relationships between text features and text classes that are often present in text classification tasks.
- It models the probability of each class as a function of the independent variables, allowing for a more accurate estimation of the probability of a given text belonging to a particular class.
- Additionally, it 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.

Design and Implementation

- Dataset details
- Model implementation details of Hyper parameters
  - SVM hyper parameter
  - Multinomial Logistic regression hyper parameter
- Performance of the models

Dataset Details

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



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

Hyper paramaters used for Models

SVM hyper parameter

- The degree hyper parameter of a SVM is an integer that controls the complexity of the model. It is used to calculate the polynomial kernel, which separates data points in the feature space. If the degree is too high, the model may over fit the data, leading to poor generalization performance. To adjust the degree for optimal results, k-fold cross-validation is used. This involves splitting the data into k different subsets and training and testing the model on each of the k subsets.

Multinomial Logistic regression hyper parameter

- The two most commonly used solvers for Multinomial Logistic Regression are 'sag' and 'saga'.'sag' is a modification of Stochastic Average Gradient Descent and is 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’ . They are used to find the optimal solution to a convex optimization problem.

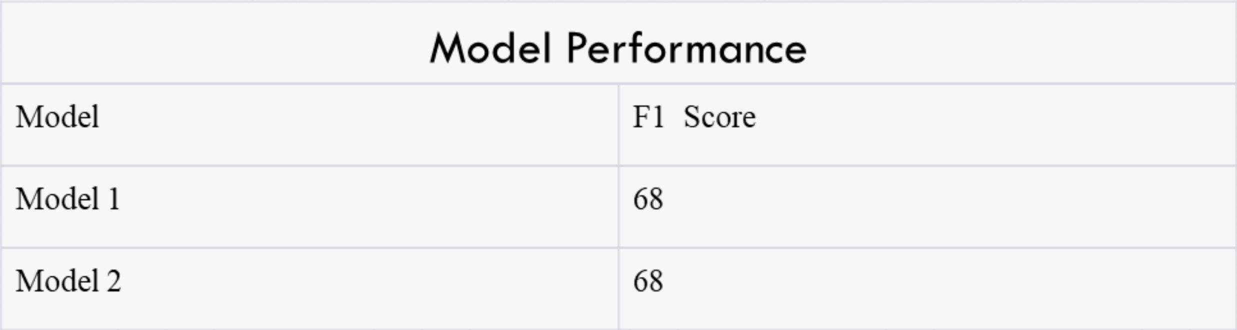
Data size effect and comparision

This graph shows the accuracy two models SVM and MLR for 4 data split quantities.(25,50,75,100)

Data %	Total	% OFF	% NOT
25%	3078 × 3	1014	2064
50%	6156 × 3	2010	4146
75%	9234 × 3	3008	6226
100%	12312 × 3	4092	8220

Performance of the Models

- The models of SVM and Multinomial Logistic Regression are compared with the metric of F1 score and the results.



Model Performance

Model	F1 Score
Model 1	68
Model 2	68

Testing Dataset accuracy score for Model 1

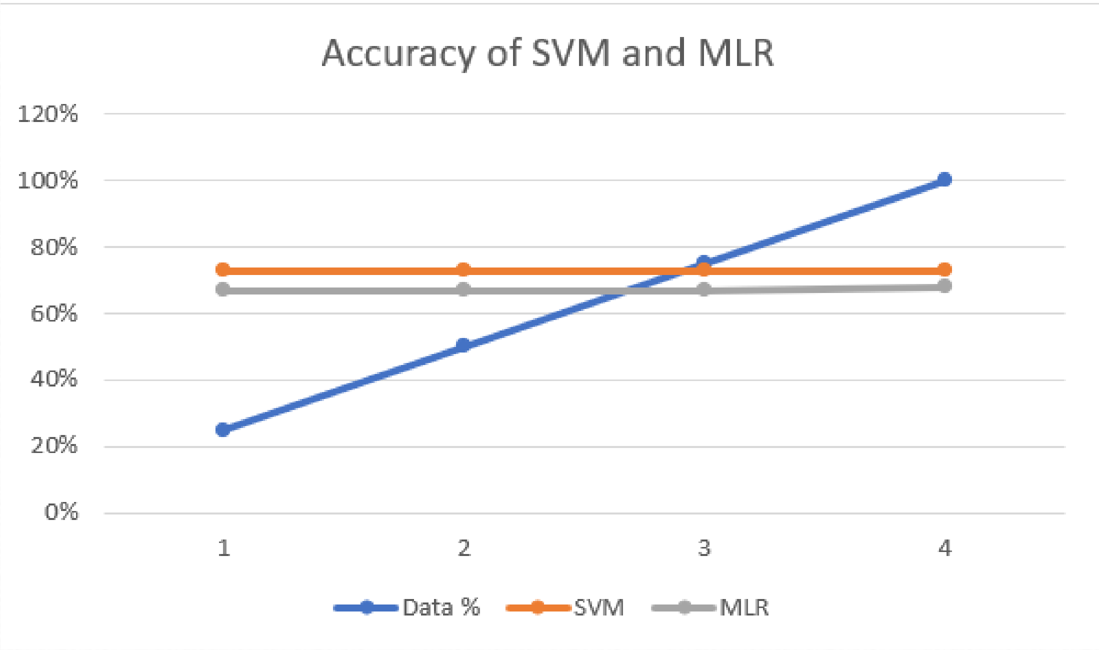
Data	Precision	Recall	F1-score	Accuracy
25%	0.51	0.73	0.46	0.73
50%	0.53	0.72	0.48	0.73
75%	0.52	0.73	0.47	0.73
100%	0.52	0.72	0.47	0.73

Testing Dataset accuracy score for Model 2

Data	Precision	Recall	F1- score	Accuracy
25%	0.56	0.57	0.56	0.67
50%	0.54	0.56	0.54	0.67
75%	0.54	0.56	0.54	0.67
100%	0.54	0.56	0.53	0.68

Accuracy of both models

This shows the accuracy of both models used SVM and MLR with respect to size of data that is splitted and trained



Multilingual Languages

- Text analysis of the Twitter dataset is performed using two models, SVM and Multinomial   Logistic Regression.
- A comprehensive table is provided that details the accuracy of the models. The performance results of five different datasets for Hate Speech Detection are shown in the report, with the highest F1 score at 100%.
- The accuracy scores of Multinomial Logistic Regression trained on five different datasets for offensive language detection with different languages like Portuguese, German, Chinese, Dutch, English.
- The accuracy scores are highest for the model trained on 100% of the data, indicating that the model performs best when trained on the full dataset.

Summary

This project describes a text analysis procedure using Support Vector Machines (SVM) and Multinomial Logistic Regression (MLR). The dataset chosen for this assignment was used to evaluate model performance and investigate the impact of dataset size on accuracy. The classification results showed that the SVM method was more accurate than the MLR approach. The optimal precision, recall, F1-score, and accuracy were achieved when the SVM model was trained on 100% of the data set, while the MLR model was trained on 75% of the dataset. Multilingual datasets produced the same results.

Conclusion

- The Support Vector Machines (SVM) and Multinomial Logistic regression (MLR) are useful for text analysis. It has been shown that the appropriate strategy for achieving the best results may change depending on the size of the dataset. SVM outperformed MLR on this dataset, according to the results. This task has increased the understanding of the importance of dataset size and the usefulness of SVM and MLR for text analysis. It also illustrates how text analysis can be used to multilingual datasets with varying amounts.