SENTIMENT CLASSIFICATION

1.Introduction

Sentiment classification is the process to use text and based upon the words used we infer if the text gives a positive or negative feedback. This report is generated based on the task for sentiment classification using traditional machine learning techniques.

2.Dataset used – IMDB, which has 50,000 reviews.

https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data

3.Aims and Objectives

Aim of this project is given a text review, classify if it is a positive or negative review.

Objectives:

1. classification of a text as binary classification task
2. Evaluate various models through comparison of accuracy and other metrics

4.Methodology

With respect to constraints established we have restricted to implement two vectorization techniques – TF-IDF () and Word2vc

TF-IDF – Term Frequency-Inverse Document Frequency , this method tells us the importance of a word relative to its frequency in entire dataset , so that the common words don't get more importance and more important words could get more weightage .

Word2Vec- This methodology captures the semantic menaing based on word context.it helps the model to understand the relationships between different words which can give better view in sentiment classification.

4.2 Model selection

The following models were used for our objective.

1.Logistic Regression

2.SVM

3.XGboost

4.MultinomialNB

5.Random Forest

4.3 Model Training and evaluation

The dataset was split into training and testing subsets to ensure that the models were evaluated on unseen data. The models were trained on the TF-IDF and Word2Vec representations, and their performances were assessed using accuracy scores and classification reports, which included precision, recall, and F1 scores.

5. Results

Model Name	TF-IDF (Accuracy)	Word2Vec(Accuracy)
Logistic Regression	89.36	58.71
SVM	90.9	58.99
XGBoost	85.65	58.79
RandomForest	86.17	58.86
MultinomialNB	86.21	(GaussianNB)56.92

Table 1 : Comparison of all the models and methodologies

Please see that for TF-idf we have used Multinomial and for Word2vec we have used GaussainNb, as when we use word2vec it can generate negative values and our algorithm MultinomialNB doesn’t accepts negative values.

The best model in terms of the metric we have chosen,i.e. accuracy is SVM (Support Vector Machine) with TF-IDF as the vectorizer. Precision and recall stand at 91%.

6. Conclusion

The sentiment classification task on the IMDb movies dataset successfully demonstrated the capabilities of various machine learning models and text representation techniques. The combination of TF-IDF and SVM gave the best results.\