# Sentiment Analysis Using Machine Learning Models: Mini Project 2


## Abstract

Sentiment analysis has become an indispensable tool in understanding public opinion in various domains such as market research and social media monitoring. This study presents an empirical analysis of sentiment classification using the Sentiment140 dataset, comparing the performance of two distinct machine learning models.


## 1. Introduction

User-generated textual data has proliferated since the emergence of social media platforms. Sentiment analysis attempts to quantify the public's perceptions and opinions by automating the sentiment classification of these texts. Our work uses a dataset of about 1.6 million tweets to tackle the problem of sentiment classification..

## 2. Dataset

The Sentiment140 dataset, created by Stanford University researchers, serves as the basis for our analysis. It consists of tweets automatically annotated for sentiment. This dataset is a benchmark in sentiment analysis research due to its volume and real-world applicability.

## 3. Data Processing

Data preprocessing is crucial in text analytics to convert raw text into a machine-readable format. Our preprocessing pipeline involved tokenization, stopword removal, and vectorization using Term Frequency-Inverse Document Frequency (TF-IDF) transformation.

## 4. Modelling Approaches

We employed two distinct models for our sentiment analysis task:

### 4.1. Support Vector Machine (SVM)

The SVM model, known for its effectiveness in high-dimensional spaces, was implemented with a linear kernel and optimized using a grid search for hyperparameter tuning.

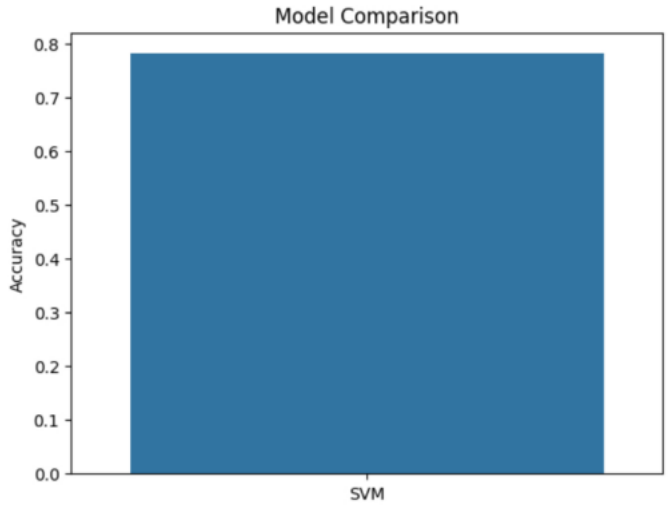
### 4.2. Multilayer Perceptron (MLP)

As a representative of deep learning models, the MLP was configured with a single hidden layer. The model was trained using backpropagation with an emphasis on generalization to avoid overfitting.

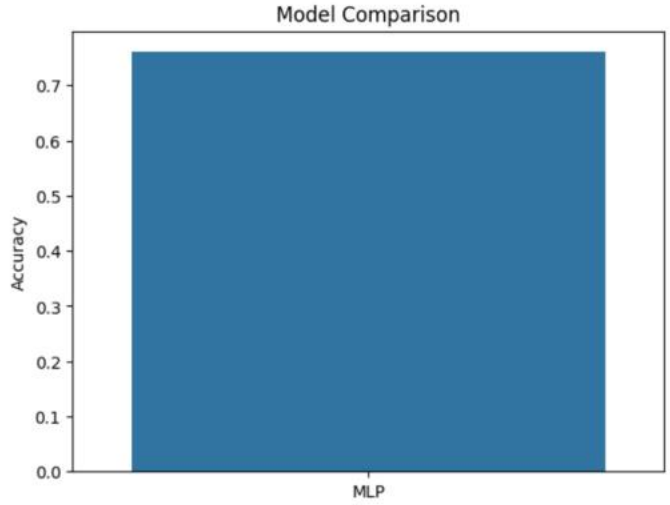
## 5. Performance Comparison

The SVM model demonstrated a notable precision and recall, achieving an accuracy of approximately 78%. The MLP showed comparable performance with an accuracy of 76%. The comparison was visualized using accuracy bar charts to illustrate the relative performance clearly.

```
SVM Accuracy: 0.78059375
              precision    recall  f1-score   support

           0       0.79      0.77      0.78     16002
           4       0.77      0.79      0.78     15998

    accuracy                           0.78     32000
   macro avg       0.78      0.78      0.78     32000
weighted avg       0.78      0.78      0.78     32000
```




```
MLP Accuracy: 0.76046875
              precision    recall  f1-score   support

           0       0.76      0.77      0.76     16002
           4       0.76      0.76      0.76     15998

    accuracy                           0.76     32000
   macro avg       0.76      0.76      0.76     32000
weighted avg       0.76      0.76      0.76     32000
```



## 6. Discussion

In accuracy, the SVM model fared marginally better than the MLP. This could be explained by the SVM's resilience when dealing with sparse data, which is typical of tasks involving text classification. Still, there is hope for future research on MLPs' ability to capture complicated patterns through deeper architecture.

## 7. Conclusions

Our study encountered challenges in balancing model complexity with computational efficiency. The SVM's performance suggests that traditional machine learning models are still competitive in sentiment analysis tasks. Meanwhile, MLPs indicate the growing relevance of neural networks, although their performance may be improved with further tuning and more complex architectures.

## 8. Future Work

Future research could explore the integration of pre-trained word embeddings and more advanced neural network architectures, such as recurrent neural networks (RNNs) or BERT and transformers, to potentially enhance classification accuracy.