# **EMOTION CLASSIFICATION**

A collage of a person with a bun

Description automatically generated

**EUNICE ABIEYUWA IGBINEDION**

**TABLE OF CONTENTS**

1. **Introduction**
2. **Dataset Overview**
3. **Text Normalization**
4. **Data Splitting**
5. **Bag of Words (BoW) Models**
6. **TF-IDF Models**
7. **Hyperparameter Tuning for BoW Models**
8. **Word2Vec Models**
9. **Hyperparameter Tuning for Word2Vec Models**
10. **GloVe and LSTM Models**
11. **Conclusion**

**1. Introduction**

This report provides a comprehensive analysis of emotion classification using various natural language processing (NLP) models. The objective is to identify and predict emotions in text data, with a focus on enhancing accuracy and generalization through advanced modelling techniques.

**2. Dataset Overview**

The dataset comprises multiple emotion categories, predominantly Happiness, followed by Anger and Sadness. The analysis reveals that people tend to send longer and more messages when happy, compared to when they are angry or sad.

A screenshot of a computer code

Description automatically generated

**3. Text Normalization**

Text normalization involves cleaning and standardizing the text data to ensure consistency and improve the performance of NLP models. This process includes removing special characters, converting text to lowercase, and eliminating stop words.

A computer screen shot of a computer code

Description automatically generated

**4. Data Splitting**

The dataset was divided into training, testing, and validation sets to evaluate the model's performance. This step is crucial for assessing how well the model generalizes to unseen data.

A diagram of training and training

Description automatically generated

**5. Bag of Words (BoW) Models**

BoW models were employed to convert text data into numerical representations. Initial results indicated that these models perform well, particularly in identifying the "Happiness" category.

**6. TF-IDF Models**

TF-IDF (Term Frequency-Inverse Document Frequency) models were explored to weigh the importance of words in the dataset. The XGBoost and Random Forest model using TF-IDF showed overfitting on the training data, but a comparative analysis revealed that models using CountVectorization outperformed those using TF-IDF.

A black screen with white text

Description automatically generated

**7. Hyperparameter Tuning for BoW Models**

Hyperparameter tuning was conducted on the best-performing BoW model, SDG & XGBoost with CountVectorizer. The tuned model demonstrated good generalization, performing well on the test set with a reasonable balance between accuracy and overfitting. The Decision tree model using TF-IDF showed overfitting on the training data

**A screenshot of a black and white screen

Description automatically generated**

**8. Word2Vec Models**

Word2Vec models were explored to capture semantic meaning in text data. The XGBoost and Random Forest and Decision Tree shows that there was overfitting in the test dataset with Word2Vec outperformed other models, showing better accuracy in emotion classification.

A screenshot of a black screen

Description automatically generated

**9. Hyperparameter Tuning for Word2Vec Models**

After hyperparameter tuning, the Linear SVM model achieved an accuracy of 87%, with minimal overfitting. This indicates that the model is effective in predicting emotions on new data.

A screenshot of a black and white screen

Description automatically generated

**A graph with different colored lines

Description automatically generated with medium confidence**

**10. GloVe and LSTM Models**

The performance of the GloVe LSTM model was not optimal due to the small dataset size. To enhance emotion prediction and pattern identification, it is recommended to collect more data for better deep learning performance.

**A screen shot of a computer

Description automatically generated**

**11. Conclusion**

The Linear SVM with Word2Vec emerged as the best model, achieving a high validation accuracy of 73%. Despite the challenge of users typing native language slang in English, the model effectively predicted emotions. Future improvements will focus on expanding the dataset to further enhance performance.

A screenshot of a computer

Description automatically generated

**Visualisation and Insights**

**Data Distribution:**

* **Happiness:** Highest number of messages.
* **Anger:** Moderate number of messages.
* **Sadness:** Fewest number of messages.

A graph with a number of bars

Description automatically generated with medium confidence

**Model Performance:**

* **BoW Models:** Good performance, particularly in identifying Happiness.
* **TF-IDF Models:** XGBoost showed overfitting; CountVectorization models performed better.
* **Word2Vec Models:** Linear SVM achieved the highest accuracy.
* **GloVe LSTM:** Requires more data for optimal performance.

A screenshot of a graph

Description automatically generated

**Hyperparameter Tuning:**

* **BoW (XGBoost):** Improved generalization with balanced accuracy.
* **Word2Vec (Linear SVM):** Achieved 74% accuracy with minimal overfitting.

**Future Work:**

* **Data Collection:** Gather more data to improve deep learning models.
* **Advanced Techniques:** Explore other NLP techniques and models.
* **Real-world Application:** Implement the best model in real-world applications to validate its effectiveness.

**Happy Reading**