CHAPTER 4

RESEARCH DESIGN AND IMPLEMENTATION

4.1 Introduction

This chapter presents the methodology and processes undertaken to analyze social media posts using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). It covers the steps involved in exploratory data analysis (EDA), data preparation, feature extraction, model development, and evaluation. The aim is to demonstrate how text data is processed, trained, and evaluated to extract topics and trends effectively.

4.2 Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution and structure of the dataset. Key features such as post length, frequency of posts, and engagement metrics were analyzed.

The analysis revealed that a majority of posts fell within the 50 to 200-word range, suggesting a preference for concise and impactful messaging, a common trend observed in social media platforms.

Sentiment analysis was performed to assess the overall emotional tone of the dataset. The analysis indicated a relatively balanced distribution across positive, negative, and neutral sentiments.

Preliminary keyword analysis identified several prominent themes within the dataset, including technology, health, and politics.

Histograms and scatter plots were utilized to visualize the relationship between post engagement metrics and sentiment polarity. Word clouds were generated to provide a visual representation of the most frequent keywords, offering insights into the dominant topics and discourse within the dataset.

4.3 Data Preparation

Data preparation is critical for ensuring quality input for the models. Steps included text cleaning which removal of URLs, special characters, and stop words. The tokenization and lemmatization that each post was tokenized, and words were lemmatized to their base forms to ensure uniformity.

The handling missing data means rows with significant missing values were excluded, while others were filled using imputation techniques. The normalization means all text data were converted to lowercase to prevent duplication caused by case sensitivity.

4.4 Feature Extraction

The primary features for topic analysis were derived from the cleaned dataset to enhance the quality of the analysis. First, Term Frequency-Inverse Document Frequency (TF-IDF) was applied to identify and quantify important words within each post, allowing for the distinction of terms that are both frequent within a specific document and rare across the entire corpus. Additionally, pre-trained word embeddings, such as GloVe or Word2Vec, were utilized to capture the semantic relationships between words, enabling the model to understand the contextual meanings beyond simple word frequencies.

Lastly, sentiment scores were computed for each post, utilizing sentiment analysis libraries to assess both the polarity and subjectivity of the

text. The polarity score indicates the emotional tone of the text, while the subjectivity score reflects the degree of personal opinion or factual content. This comprehensive feature extraction process ensured that the dataset was enriched with relevant attributes, laying the groundwork for effective model training.

4.5 Model Development

Two primary models, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), were implemented for the task of topic classification, each contributing distinct advantages for handling sequential data. The RNN model served as a baseline for sequence data processing, capturing the contextual relationships between words in posts by utilizing its inherent capacity to process sequences step-by-step.

In contrast, the LSTM model, an extension of the RNN, incorporated forget gates designed to mitigate the issue of vanishing gradients, which commonly affects the training of deep neural networks, especially on longer sequences. This modification ensured that the LSTM model performed more effectively on longer posts, maintaining more relevant information across extended contexts.

The architecture of the LSTM model was structured as follows: the input layer consisted of word embeddings derived from the dataset, providing dense vector representations of the words. The hidden layers were composed of two LSTM layers, which performed feature extraction by learning the temporal dependencies and context of the input sequences. Finally, the output layer employed a softmax classifier to predict the topic associated with each post. To optimize model performance, hyperparameters such as learning rate, batch size, and the number of epochs were tuned through a grid search process.

4.6 Model Evaluation

To evaluate the performance of the models, several key metrics were employed to provide a comprehensive assessment of their effectiveness. Accuracy and F1-Score were used to measure the overall classification performance, with accuracy reflecting the proportion of correct predictions and the F1-Score providing a balanced measure of precision and recall, especially in the case of imbalanced class distributions.

In addition, a Confusion Matrix was utilized to highlight the strengths and weaknesses in topic classification, offering a detailed breakdown of true positive, false positive, true negative, and false negative predictions for each topic. To further monitor the models' training progress and ensure robust generalization, Loss Curves were analyzed, tracking both training and validation losses to ensure that the models did not overfit to the training data.

The results revealed that the LSTM model outperformed the RNN, demonstrating superior accuracy and better topic identification capabilities, particularly when handling long and complex posts, which benefited from the LSTM's ability to capture long-term dependencies more effectively than the standard RNN.

4.7 Summary

This chapter detailed the methodology for analyzing social media posts using RNN and LSTM. EDA and data preparation laid the groundwork for effective feature extraction. Model development highlighted the application of LSTM in handling sequential data, while evaluation metrics confirmed the robustness of the approach. The next chapter will delve into the results and discussion of the findings.