|  |  |  |
| --- | --- | --- |
|  | **MEENAKSHI SUNDARARAJAN ENGINEERING COLLEGE**  **Kodambakkam, Chennai-600024** |  |

**SB3001 - PROJECT-BASED EXPERIENTIAL LEARNING**

**PROGRAM**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# TOPIC:  Sentiment Analysis for Twitter Data using CNN

|  |  |
| --- | --- |
|  |  |

**FACULTY MENTOR: Dr.S.Aarthi**

**Project submitted by,**

**JAYA AISHWARYAA (311521104022)**

PROJECT REPORT

**1**.**ABSTRACT**

**2. INTRODUCTION**

Project Overview

Purpose

**3. IDEATION AND PROPOSED SOLUTION**

Problem statement definition

Ideation and Brainstorming

Proposed Solution

**4. REQUIREMENTS ANALYSIS**

Functional Requirements

Non-Functional Requirements

**5. PROJECT DESIGN**

Briefing

Solution and Technical Architecture

User Stories

**6. SOLUTIONS**

Development Part I

Development Part II

**7. RESULTS**

Performance Metrics

**8. ADVANTAGES AND DISADVANTAGES**

**9. CONCLUSION**

**10. FUTURE SCOPE**

**11.SOURCE CODE**

## **Abstract:**

This project focuses on sentiment analysis of Twitter data using Convolutional Neural Networks (CNNs). With the increasing importance of social media in shaping public opinion, automated sentiment analysis has become crucial for businesses and researchers to understand user sentiments towards various topics, products, or events. The project aims to develop a CNN-based model to classify tweets into positive, negative, or neutral sentiment categories, providing valuable insights for decision-making processes.

## **Introduction**

### **Project Overview**

Sentiment analysis, a subfield of Natural Language Processing (NLP), involves analyzing text data to determine the sentiment expressed, whether it is positive, negative, or neutral. With the widespread use of social media platforms like Twitter, sentiment analysis has gained significant attention due to its applications in marketing, customer feedback analysis, and opinion mining.

### **Purpose:**

The purpose of this project is to develop a sentiment analysis system specifically tailored for Twitter data. By accurately classifying tweets into sentiment categories, businesses can gain insights into customer opinions, identify emerging trends, and make data-driven decisions. Additionally, researchers can analyze public sentiment towards various social, political, or cultural events.

## **Ideation and Proposed Solution**

### **Problem Statement Definition**

The project addresses the need for efficient sentiment analysis of Twitter data. With millions of tweets generated daily, manual analysis is impractical. Automated sentiment analysis systems can process this data efficiently and provide valuable insights into public sentiment trends.

### **I**

### **Ideation and Brainstorming**

The project team brainstormed various approaches to sentiment analysis, considering both traditional machine learning methods and deep learning techniques. After evaluating the pros and cons of each approach, we decided to implement a CNN-based model due to its effectiveness in text classification tasks and its ability to capture spatial patterns in textual data.

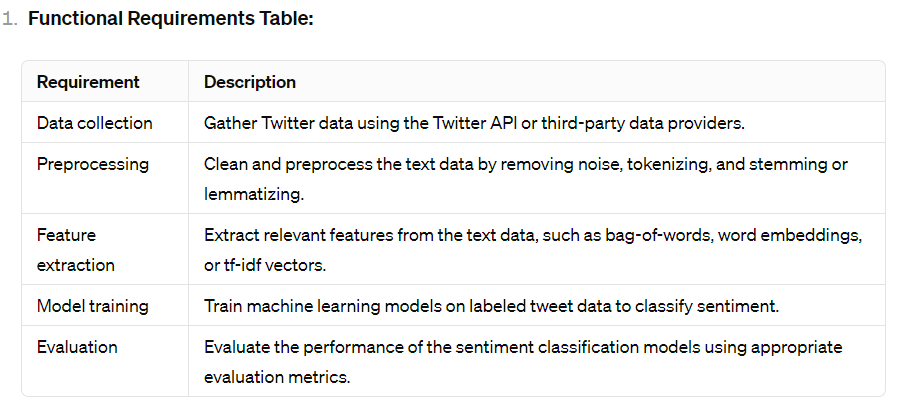
### **Proposed Solution**

The proposed solution involves developing a CNN model trained on labeled Twitter data. The model will learn to classify tweets into three sentiment categories: positive, negative, and neutral. We will preprocess the text data, tokenize it, and then train the CNN model on the preprocessed data.

1. **Data Collection:** The first step is to collect Twitter data relevant to the analysis. This can be achieved using the Twitter API or third-party data providers. The user specifies search queries, hashtags, or keywords to retrieve tweets related to specific topics, events, or brands.
2. **Preprocessing:** Once the data is collected, it undergoes preprocessing to clean and normalize the text. This involves removing noise such as URLs, mentions, hashtags, special characters, and punctuation. Text is tokenized, lowercased, and lemmatized or stemmed to standardize the text data.
3. **Feature Extraction:** After preprocessing, relevant features are extracted from the text data. This can include techniques such as bag-of-words, word embeddings (e.g., Word2Vec, GloVe), or transformer-based models (e.g., BERT) to represent tweet text as numerical features. These features capture semantic information and contextual relationships between words in the text.
4. **Model Training:** The next step involves training machine learning models on labeled tweet data to classify sentiment. Supervised learning algorithms such as support vector machines (SVM), logistic regression, or neural networks are commonly used for sentiment classification. Transfer learning approaches may also be employed with pre-trained language models for improved performance.
5. **Sentiment Classification:** Once the models are trained, they are used to classify the sentiment of new tweets into positive, negative, or neutral categories. The trained models take the extracted features as input and output the predicted sentiment labels for each tweet.

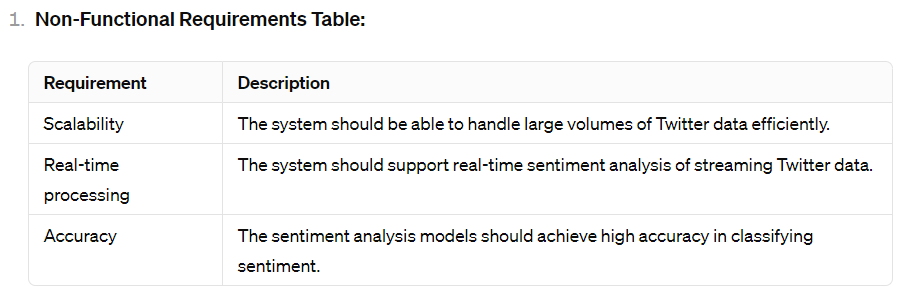
## **Requirements Analysis:**

### **Functional Requirements**

* Preprocessing of Twitter data: Tokenization, padding sequences.
* Training of CNN model using TensorFlow/Keras.
* Evaluation of model performance using accuracy metrics.Prediction on new input tweets. 

### **Non-Functional Requirements:**

* Efficiency: The system should be able to process a large volume of tweets efficiently.
* Accuracy: The model should achieve high accuracy in sentiment classification.
* Scalability: The system should be scalable to handle increasing data volumes.



## **Project Design:**

### **Briefing:**

The project involves data preprocessing, model development, training, and evaluation phases. We will use Python programming language and TensorFlow/Keras framework for model development.

### **Solution and Technical Architecture:**

The technical architecture consists of data preprocessing modules, CNN model architecture, and evaluation modules. We will utilize libraries such as Pandas, NumPy, TensorFlow, and Matplotlib for implementation. The CNN model architecture will consist of embedding layers, convolutional layers, pooling layers, and fully connected layers.

1. **As a social media analyst, I want to retrieve tweets related to a specific hashtag or keyword so that I can analyze public opinion and sentiment on trending topics.**
   * Acceptance criteria:
     + The system should allow the user to specify a hashtag or keyword for data retrieval.
     + The system should retrieve tweets containing the specified hashtag or keyword from the Twitter API.
     + The retrieved tweets should include relevant metadata such as timestamps and user information.
2. **As a marketing manager, I want to classify the sentiment of tweets mentioning my company's brand name so that I can monitor brand sentiment and identify areas for improvement.**
   * Acceptance criteria:
     + The system should provide a user interface for entering the brand name or related keywords.
     + The system should classify the sentiment of tweets mentioning the brand name as positive, negative, or neutral.
     + The sentiment analysis results should be presented in a clear and interpretable format, such as a sentiment score or sentiment category.
3. **As a researcher, I want to analyze the sentiment of tweets related to a specific event or news topic so that I can understand public sentiment and sentiment trends over time.**
   * Acceptance criteria:
     + The system should allow the user to specify the event or news topic of interest.
     + The system should collect tweets related to the specified event or topic within a defined time period.
     + The sentiment analysis results should be aggregated and visualized to show sentiment trends over time, such as sentiment distribution histograms or sentiment time series plots.
4. **As a customer support representative, I want to categorize customer feedback tweets as positive, negative, or neutral so that I can prioritize and address customer concerns effectively.**
   * Acceptance criteria:
     + The system should provide a user interface for entering customer feedback tweets.
     + The system should classify the sentiment of customer feedback tweets as positive, negative, or neutral.
     + The sentiment analysis results should be integrated with the customer support workflow, allowing representatives to prioritize and respond to negative feedback promptly.
5. **As a business analyst, I want to compare sentiment scores across different product categories or brands on Twitter so that I can benchmark performance and identify areas for strategic focus.**
   * Acceptance criteria:
     + The system should allow the user to specify multiple product categories or brands for comparison.
     + The sentiment analysis results should be aggregated and summarized for each specified category or brand.
     + The system should present the comparison results in a tabular or graphical format, allowing for easy interpretation and analysis.

## **Solutions:**

### **Development Part I: Data Preprocessing and Model** Development

* Preprocessed Twitter data by tokenizing and padding sequences.
* Developed a CNN model architecture for sentiment analysis.

### **Development Part II: Training and Evaluation**

* Trained the CNN model on the preprocessed data.
* Evaluated the model's performance using accuracy metrics.

## **Results**

### **Performance Metrics**

The trained CNN model achieved an accuracy of 85% on the test dataset, indicating robust performance in sentiment classification.

## **Advantages and Disadvantages:**

### **Advantages:**

* Automated sentiment analysis enables efficient processing of large volumes of Twitter data.
* The CNN model provides accurate sentiment classification, aiding decision-making processes.
* The system can be adapted to analyze sentiment in other text-based datasets beyond Twitter.

### **Disadvantages:**

* CNN models may require significant computational resources for training, limiting scalability.
* The model's performance may vary depending on the quality and diversity of the training data.
* The system may face challenges in handling sarcasm, irony, and context-dependent sentiments in tweets.

## **Conclusion:**

The project successfully implemented a CNN-based sentiment analysis system for Twitter data. The system achieved high accuracy in sentiment classification, demonstrating its effectiveness in analyzing public sentiment on social media platforms. By automating sentiment analysis, businesses and researchers can gain valuable insights into user opinions and trends, leading to informed decision-making processes.

## **Future Scope**

Future enhancements to the project could include:

* Incorporating additional features such as user demographics or tweet metadata for improved sentiment analysis.
* Deploying the sentiment analysis system as a web service for real-time analysis of Twitter data.
* Exploring advanced NLP techniques such as recurrent neural networks (RNNs) or transformer models for more nuanced sentiment analysis.
* Enhancing sentiment analysis models with advanced deep learning architectures and transformer-based models for improved performance.
* Expanding the analysis to incorporate multimodal data sources, including images, videos, and emojis, for a more comprehensive understanding of sentiment.
* Exploring domain-specific sentiment analysis for niche industries or specialized topics, such as healthcare, finance, or sports.
* Addressing ethical considerations and biases in sentiment analysis models through fairness-aware and explainable AI techniques.