SYNOPSIS

Report on

Sentiment Analysis on Social Media Data

by

Govind Gupta [2300290140065]
Harsh Sharma [2300290140069]
Kanchan Sagar [2300290140084]
Kartikey Raghuvanshi [2300290140086]

Session:2024-2025 (IV Semester)

Under the supervision of

Mr. Rabi Narayan Panda (Associate Professor & Addl. HOD**)**

KIET Group of Institutions, Delhi-NCR, Ghaziabad

DEPARTMENT OF COMPUTER APPLICATIONS
KIET GROUP OF INSTITUTIONS, DELHI – NCR ,
GHAZIABAD-201206
(2023-2025)

ABSTRACT

In today's digital landscape, social media and online platforms serve as primary channels for expressing opinions and emotions. Understanding these sentiments is crucial for businesses, policymakers, and researchers to analyze public perception and make data-driven decisions. This project focuses on developing an advanced **Sentiment Analysis** system using **Natural Language Processing (NLP)** and **machine learning techniques** to classify sentiments in textual data.

The system leverages **both English and Hindi datasets**, incorporating diverse sources such as **tweets, product reviews, and social media comments**. By applying **preprocessing techniques**—including tokenization, stopword removal, stemming, and lemmatization—the model ensures high-quality text input. Feature extraction methods like **TF-IDF (Term Frequency-Inverse Document Frequency)** and **word embeddings (Word2Vec, GloVe)** further enhance data representation.

To achieve accurate sentiment classification, various **machine learning algorithms** such as **Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM)** are implemented and evaluated. The performance of these models is measured using **accuracy, precision, recall, and F1-score** to ensure reliable predictions. Additionally, deep learning techniques using **Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)** networks are explored for enhanced sentiment understanding.

The project also introduces sentiment analysis for **sarcasm detection in Hindi tweets**, addressing challenges in multilingual NLP. The use of **real-time data visualization** through word clouds and interactive dashboards provides meaningful insights into sentiment trends.

By integrating **machine learning and NLP techniques**, this sentiment analysis system offers a powerful tool for businesses and researchers to monitor public sentiment, improve customer engagement, and drive informed decision-making. The solution provides a scalable, efficient, and multilingual approach to sentiment analysis, advancing the capabilities of natural language understanding in the digital era.

CONTENT Page No.

CONTENT Page No.

CHAPTER 1: INTRODUCTION

The rapid growth of social media has transformed the way people communicate, with millions of users expressing their thoughts, opinions, and emotions online every day. This unstructured textual data contains valuable insights that can be leveraged for sentiment analysis—helping businesses, governments, and organizations make informed decisions.

This project focuses on building a **Sentiment Analysis system** that classifies text into positive, negative, or neutral categories. Using **Natural Language Processing (NLP) and Machine Learning**, the system processes textual data from multiple sources, including **tweets, product reviews, and social media comments**.

By applying state-of-the-art **feature extraction techniques** such as **TF-IDF and word embeddings**, the project ensures a high-quality understanding of text data. The integration of **deep learning models**, including **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks**, enhances the accuracy and reliability of sentiment classification.

With applications in areas such as **brand monitoring, customer feedback analysis, political opinion tracking, and market research**, sentiment analysis plays a crucial role in data-driven decision-making.

CHAPTER 2: LITERATURE REVIEW

A comprehensive study of existing **sentiment analysis techniques** was conducted to understand the advancements and challenges in the field.

Table: 2.1 Literature Review

S.No.	Year	Author	Contribution
1	2014	Bing Liu	Discussed the importance of sentiment lexicons and feature-based sentiment classification.
2	2017	Saif Mohammad	Explored the use of deep learning in emotion detection and sentiment analysis.
3	2019	Bo Pang	Introduced a hybrid approach combining rule-based and machine learning methods for sentiment classification.
4	2021	Google Research	Proposed a Transformer-based model (BERT) for sentiment analysis, achieving state-of-the-art accuracy.
5	2023	OpenAI	Developed a large-scale sentiment analysis model using generative AI and reinforcement learning.

CHAPTER 3: PROJECT OBJECTIVES

3.1 Importance of Sentiment Analysis

Sentiment analysis is a crucial tool in modern data science that helps businesses and organizations understand customer opinions, market trends, and public sentiment. By analyzing textual data from sources like social media, product reviews, and surveys, organizations can make informed decisions, improve customer experience, and develop targeted marketing strategies.

3.2 Applications in Business and Social Media

Sentiment analysis has wide-ranging applications:

- **Customer Feedback Analysis**: Companies analyze product and service reviews to enhance user satisfaction.
- **Brand Monitoring**: Businesses track brand mentions on social media to assess public perception.
- **Stock Market Predictions**: Investors analyze financial news and social media sentiment to predict stock trends.
- **Political Analysis**: Governments and political parties assess public sentiment during election campaigns.

3.3 Challenges in Sentiment Analysis

Despite its significance, sentiment analysis faces multiple challenges:

- **Sarcasm and Irony**: Detecting sarcasm, especially in textual form, remains difficult.
- **Contextual Ambiguity**: Words may have different meanings based on context (e.g., “sick” could mean unwell or amazing).
- **Multilingual and Code-Mixed Text**: Handling sentiment in multiple languages, including mixed-language texts like “Hinglish” (Hindi + English).
- **Data Imbalance**: Some datasets have significantly more positive reviews than negative ones, making it harder to train balanced models.

3.4 Role of NLP and Machine Learning

NLP and machine learning techniques enhance sentiment analysis accuracy. Some key techniques include:

- **Tokenization & Lemmatization**: Breaking down text into words and converting them into their base form.
- **TF-IDF & Word Embeddings**: Converting text into numerical features for machine learning models.
- **Naïve Bayes & SVM**: Traditional machine learning models used for text classification.

- **Deep Learning Models (LSTM, BERT)**: Advanced models that capture complex language structures and sentiment patterns.

3.5 Sentiment Analysis for Multilingual Data

As sentiment analysis expands globally, multilingual processing is essential. This project includes sentiment analysis for **Hindi** and **English** datasets. Specialized NLP libraries like **IndicNLP** and **polyglot** are used for Hindi text processing.

3.6 Advancements in Deep Learning for NLP

Recent developments in **deep learning** have significantly improved sentiment analysis accuracy. Models like **BERT (Bidirectional Encoder Representations from Transformers)** and **GPT (Generative Pre-trained Transformer)** provide state-of-the-art performance by understanding sentence context and word relationships more effectively than traditional models.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Problem Statement

The objective of this project is to develop an **efficient and scalable sentiment analysis system** that classifies textual data into **positive, negative, and neutral categories** using machine learning and NLP techniques. The system must handle **sarcasm detection, multilingual text processing, and real-time analysis**.

4.2 Data Collection and Preprocessing

To build a high-quality sentiment analysis model, data must be collected and preprocessed effectively:

- **Data Sources**: Publicly available datasets from **Twitter, IMDb reviews, Kaggle, and sentiment lexicons**.
- **Preprocessing Steps**:
    - Removing **stopwords, special characters, emojis, and URLs**.
    - **Tokenization**: Splitting text into words or phrases.
    - **Lemmatization & Stemming**: Converting words to their root form to maintain consistency.
    - **Handling Missing Data**: Removing or filling missing values using NLP techniques.

4.3 Feature Extraction (TF-IDF, Word Embeddings)

Text needs to be converted into numerical representations for machine learning models:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Assigns importance to words based on their occurrence in a document.
- **Word Embeddings (Word2Vec, GloVe, FastText)**: Captures contextual meaning and relationships between words.

4.4 Machine Learning and Deep Learning Models

Several models are tested to determine the best approach:

- **Baseline Models**: Naïve Bayes, Logistic Regression, and SVM.
- **Deep Learning Models**:
    - **LSTM (Long Short-Term Memory)**: A recurrent neural network that remembers long-term dependencies in text.
    - **BERT**: A transformer-based model that captures bidirectional context, improving sentiment classification.

    - **CNNs for Text Classification**: Convolutional neural networks applied to NLP tasks.

- **Evaluation Metrics**: Accuracy, Precision, Recall, and F1-score to measure model performance.

CHAPTER 5: PROJECT OUTCOME

5.1 Improved Sentiment Classification Accuracy

The implementation of advanced **deep learning models (BERT and LSTM)** significantly improves accuracy compared to traditional machine learning approaches. By training on **real-world datasets**, the model achieves a robust understanding of sentiment patterns.

5.2 Real-Time Analysis for Social Media Insights

A key feature of this project is the ability to analyze social media sentiment in **real-time**. This allows businesses, researchers, and policymakers to:

- Monitor **brand perception** across different regions.
- Detect **emerging trends** and shifts in public opinion.
- Identify **crisis situations** based on negative sentiment spikes.

5.3 Multilingual Sentiment Analysis

Unlike many existing sentiment analysis models that focus only on **English**, this project extends support for **Hindi sentiment analysis**. It includes specialized **tokenization, stopword removal, and word embeddings** for Hindi text processing.

5.4 Application in Business Decision-Making

Businesses can leverage sentiment analysis for:

- **Product Development**: Analyzing customer reviews to improve products.
- **Marketing Strategies**: Identifying customer preferences for targeted advertising.
- **Customer Support Automation**: Using chatbots with built-in sentiment detection for better engagement.

5.5 Advanced Sarcasm Detection

Sarcasm is a major challenge in sentiment analysis. This project incorporates:

- **Custom Hindi Sarcasm Detection Dataset**: Labeled sarcastic and non-sarcastic tweets.
- **Hybrid NLP Model**: Combining **TF-IDF, Word2Vec, and deep learning** to improve sarcasm detection.

5.6 Visualization of Sentiment Trends

To enhance **interpretability**, the project includes:

- **Word Clouds**: Visualizing most frequently used words in positive, negative, and neutral sentiments.
- **Sentiment Distribution Graphs**: Analyzing sentiment trends over time.

- **Topic Modeling**: Identifying key discussion themes in textual data.

5.7 Scalability and Future Enhancements

- **Expansion to Additional Languages**: Adding support for **French, Spanish, and regional Indian languages**.
- **Integration with Chatbots**: Enhancing customer support by integrating sentiment-aware AI chatbots.
- **Cross-Platform Deployment**: Making the system accessible via **mobile apps and web-based dashboards**.

CHAPTER 6: PROPOSED TIME DURATION

Task	Start Date	End Date	Duration (Days)
Data Collection	1-Feb-25	5-Feb-25	5
Preprocessing	6-Feb-25	10-Feb-25	5
Model Training	11-Feb-25	18-Feb-25	8
Evaluation	19-Feb-25	22-Feb-25	4
Implementation	23-Feb-25	28-Feb-25	6

CHAPTER 7: REFERENCES/ BIBLIOGRAPHY

- Bing Liu, **"Sentiment Analysis: Mining Opinions, Sentiments, and Emotions"**, 2015.
- Google AI, **"Transformers for NLP"**, 2021.