International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Opinion Analysis using Recurrent Neural Networks and Apache Spark

Mrs. A.V.L. Prasuna[1], A. Sai Aakarsh[2], A. Hemanth Reddy[3]
[1]*Assistant Professor, Mahatma Gandhi Institute of Technology*
[2,3] *UG Student, Mahatma Gandhi Institute of Technology*

***Abstract*: The internet is full of opinions, reviews, articles, and discussions that shape decisions in business, politics, and society. But finding meaningful insights in this massive and diverse data is a big challenge. Current methods often focus on specific sources and struggle with the large scale and complexity of internet data. This project aims to solve this problem using advanced tools like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models, combined with Apache Spark for fast data processing. It can analyze different types of online content in real-time and better understand the true meaning behind words. The results show that this approach is accurate and efficient, making it easier to uncover useful insights from online data. This system helps businesses, governments, and organizations make smarter decisions and respond quickly to what people think and feel.**

Keywords: **Opinion Analysis, deciphering, distributed Processing framework, Hadoop, Apache Spark, Deep Learning, Recurrent Neural Networks (RNN), sentiment score.**

I. INTRODUCTION

A key job in natural language processing (NLP) is opinion analysis, which is essential for comprehending and gaining emotions and views from textual data. The need for real-time opinion analysis tools has increased due to the exponential rise of online material on social media, e-commerce websites, and consumer evaluations. Real-time opinion analysis gives companies a competitive edge in the ever-changing business world of today by allowing them to monitor brand perception, react quickly to client input, and identify market trends.

Because of their computing complexity and scalability constraints, traditional opinion analysis techniques frequently encounter difficulties when processing huge amounts of real-time data. Emerging technologies like Hadoop and Apache Spark have drawn attention for their distributed computing capabilities, which allow for the processing of large datasets in parallel, in order to address these issues. Furthermore, by automatically learning hierarchical representations of text, deep learning algorithms—in particular, Recurrent Neural Networks (RNN)—have shown impressive success in opinion analysis tasks.

In this work, we integrate Apache Spark, and the deep learning algorithm RNN to conduct a thorough research of real-time opinion analysis systems. Our goal is to assess these frameworks' effectiveness, scalability, and performance in handling real-time text data streams for opinion analysis. We want to offer important insights into the relative advantages and disadvantages of each strategy by carrying out tests on huge datasets and comparing different metrics including processing speed, accuracy, and resource use.

A. Problem Statement.

In today's world, emotions are increasingly conveyed through text on platforms like social media, emails, customer feedback, and reviews. These texts hold valuable insights into public opinion, customer satisfaction, and mental health, making emotional analysis essential for applications like marketing, crisis intervention, social media monitoring, and customer service.

However, analyzing emotions in text is challenging due to the complexity of human language. Emotions are often subtly expressed through word choices, sentence structures, and context, and many systems rely on simple techniques like keyword matching, which miss deeper nuances and struggle with slang, sarcasm, or ambiguity.

Additionally, the rapid growth of real-time communication, such as on Twitter or live chat, has led to massive volumes of unstructured text data. Analyzing

this data in real-time requires powerful infrastructure and advanced machine learning techniques. Traditional models often face delays and lack adaptability, leading to suboptimal predictions. The dynamic nature of language, influenced by cultural trends and new terms, further complicates matters, as systems must continuously update and retrain to remain effective.

B. Existing Systems.

Lexicon-based opinion analysis uses predefined sentiment dictionaries, where words are assigned sentiment scores. While simple and not requiring labeled data, these systems struggle with understanding context, sarcasm, irony, and domain-specific variations. They also fail to adapt to evolving language, such as slang or new terms, limiting their effectiveness for more complex or dynamic datasets.

Machine learning-based opinion analysis, on the other hand, learns patterns from labeled data using algorithms like Naïve Bayes, SVM, or Decision Trees. These systems perform better at identifying opinions in varied contexts but require large amounts of labeled data, which can be expensive and time-consuming to acquire. Machine learning methods also face challenges with sarcasm and irony, as well as the computational resources required for processing large datasets. Additionally, feature engineering is needed to identify text characteristics, which can be labor-intensive and prone to missing subtle patterns. While machine learning provides advancements over lexicon-based systems, it still faces issues related to data requirements, context understanding, and scalability.

II. PROPOSED SYSTEM

A. Architecture of Proposed System.

The proposed opinion analysis system utilizes deep learning and Apache Spark to enhance emotion detection in text. It combines Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) layers for improved contextual understanding, including complex emotions like sarcasm and irony. Apache Spark is used for efficient, real-time processing of large datasets. The system integrates advanced opinion layers for deeper emotional insights and offers a user-friendly interface for easy interaction. Data preprocessing techniques, such as tokenization and cleaning, are implemented to prepare input for analysis, with real-time opinion classification from sources like social media and reviews. Performance is evaluated using accuracy, sensitivity, and specificity, and scalability testing ensures the system can handle increasing data loads.

B. Advantages of Proposed System.

- Improved Accuracy
- Versatility
- Robustness
- Real Time Processing

III. LITERATURE SURVEY

This project builds upon various studies that explore content analysis and opinion classification techniques, providing key insights into the design and functionality.

In the paper Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis by authors Yule Kim (2022-23), the authors argue that combining lexicon-based inputs with Convolutional Neural Networks (CNNs) can significantly enhance sentiment classification accuracy in online reviews. Despite this, the authors emphasize the importance of integrating lexicon-based data with deep learning models to improve sentiment analysis in diverse online review contexts [1].

In Using Online Reviews for Customer Sentiment Analysis (2021), Rae Yule Kim explores the use of a lexicon-based approach for sentiment analysis, incorporating alternative metrics such as sentiment scores and review lengths. Their paper discusses the biases inherent in online reviews, particularly extreme positive or negative opinions, which may distort sentiment interpretation. The study suggests that these biases should be addressed to provide a more accurate view of customer sentiment, offering new ways to interpret consumer feedback beyond traditional methods [2].

The paper Opinion Mining: A Survey of Techniques and Applications by authors in 2019 provides a detailed review of various opinion mining techniques, including

sentiment analysis, feature extraction, and classification methods. Their study presents the challenges faced by older techniques, especially in current applications where some may be less effective. Their study suggests that identifying gaps in current research could pave the way for better future methods in opinion mining [3].

In Twitter Sentiment Analysis Based on Ordinal Regression (2019), Shihab Elbagir and Jing Yang investigate the use of ordinal regression with machine learning algorithms such as Decision Trees, SoftMax, Support Vector Regression (SVR), and Random Forest to improve sentiment analysis accuracy on Twitter data. This study emphasizes the need for better models to handle the unique structure and informal nature of social media text, which often complicates sentiment analysis [4].

In Sentiment Analysis in social media: A Review (2021), the authors provide a comprehensive review of sentiment analysis methodologies. They discuss the difficulties of working with social media data, including the reliance on existing datasets that might skew the analysis results. The paper suggests potential improvements for future studies, proposing more adaptable and robust methods to address the unique challenges posed by social media sentiment analysis, such as varying language and context [5].

IV. SCOPE

The project focuses on developing an advanced opinion analysis system that leverages deep learning techniques, specifically Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, combined with Apache Spark for scalable data processing. The system is designed to analyze large volumes of text data, such as social media posts, product reviews, or customer feedback, to extract and classify opinions as positive, negative, or neutral. The key components of this system include:

- Feature Extraction:
  Word Embeddings: Converting text data into numerical representations using techniques like Word2Vec, GloVe, or FastText. These embeddings capture the semantic meaning of words and are essential for deep learning models.
  Sequence Padding: Ensuring that all input sequences (sentences or phrases) are of the same length by padding or truncating them, which is necessary for feeding data into RNNs and LSTMs.

- Deep Learning Model Development:
  RNN Architecture: Implementing a Recurrent Neural Network (RNN) to model sequential data. RNNs are well-suited for text data as they can capture dependencies between words in a sentence.
  LSTM Enhancement: Enhancing the RNN with Long Short-Term Memory (LSTM) units to address the vanishing gradient problem and improve the model's ability to capture long-range dependencies in text.
  Model Training: Training the RNN + LSTM model on labeled datasets (e.g., sentiment-labeled reviews) to classify opinions as positive, negative, or neutral. The training process will be optimized using techniques like dropout and batch normalization to prevent overfitting.

- Distributed Training with Apache Spark:
  Data Parallelism: Utilizing Apache Spark's distributed computing capabilities to parallelize the training process across multiple nodes, enabling the model to handle large datasets efficiently.

  Model Parallelism: Implementing model parallelism to distribute the deep learning model across multiple GPUs or nodes, further speeding up the training process.

V. CONCLUSION

The opinion analysis system proposed in this project tackles the complex task of extracting and interpreting emotions from large volumes of text data. Unlike traditional opinion analysis tools that rely on basic keyword matching or sentiment polarity, this system leverages advanced deep learning techniques and distributed computing to enhance emotion detection.

With its flexible design, the system processes data from multiple sources such as social media, customer feedback, and live streaming platforms and offers real-time predictions of emotions like happiness, sadness,

anger, and sarcasm. The integration of Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) layers allows for improved understanding of contextual relationships, making the system more accurate in detecting complex emotions and nuances in text.

Overall, this project demonstrates the power of combining deep learning with real-time data processing to deliver a robust, efficient, and accurate opinion analysis solution. It not only improves opinion detection but also provides deeper insights into emotions, enabling better decision-making and personalized responses.

REFERENCES

[1] Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis - IEEE Access (2022)
[2] Using Online Reviews for Customer Sentiment Analysis - Rae Yule Kim, IEEE Engineering Management Review (2021)
[3] Opinion Mining: A Survey of Techniques and Applications - ACM Computing Surveys (2019)
[4] Twitter Sentiment Analysis Based on Ordinal Regression - Shihab Elbagir, Jing Yang, IEEE Access (2019)
[5] Sentiment Analysis in Social Media: A Review - Journal of Information Science (2021)
[6] Real-Time Lexicon-Based Sentiment Analysis Experiments on Twitter - Yusuf Arslan, Aysenur Birturk, Bekjan Djumabaev, Dilek Kucuk, IEEE (2017)
[7] Sentiment Analysis of Students' Comment Using Lexicon Based Approach - Khin Zezawar Aung, Nyein Nyein Myo, ICIS (2017)
[8] A Survey on Sentiment Analysis Techniques - International Journal of Computer Applications (2020)
[9] Machine Learning Techniques for Sentiment Analysis: A Survey - IEEE Transactions on Neural Networks and Learning Systems (2020)