# **Sentiment Analysis of Twitter Feed Data Using CNN BERT and BiLSTM**

A PROJECT REPORT

#### Submitted by

**S S R SUBRAMANYA HEMANT KONDURI [RA2111050010010]**

**NAGARJUNA LEDALLA [RA2111050010019]**

#### Under the Guidance of

## **DR. M RAMPRASATH**

Assistant Professor, Department of Data Science and Business Systems

### *in partial fulfillment of the requirements* *for the degree of*

## BACHELOR OF TECHNOLOGY

## in

## COMPUTER SCIENCE ENGINEERING

## with specialization in BLOCKCHAIN TECHNOLOGY



## DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

## COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

## KATTANKULATHUR- 603 203

### NOVEMBER 2024

Department of Computational Intelligence

##### SRM Institute of Science & Technology

##### Own Work\* Declaration Form

##### Degree/ Course :B.Tech in Computer Science and Engineering

**Student Name :S S R Subramanya Hemant Konduri, Nagarjuna Ledalla**

##### Registration Number :RA2111050010010, RA2111050010019

**Title of Work : Sentiment Analysis of Twitter Feed Data using CNN, BERT and BiLSTM**

I / We hereby certify that this assessment compiles with the University’s Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

* Clearly referenced / listed all sources as appropriate
* Referenced and put in inverted commas all quoted text (from books, web, etc)
* Given the sources of all pictures, data etc. that are not my own
* Not made any use of the report(s) or essay(s) of any other student(s) either past or present
* Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
* Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

|  |
| --- |
| **DECLARATION:** |
| I am aware of and understand the University’s policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.  S S R Subramanya Hemant Konduri[RA2111050010010]  Nagarjuna Ledalla[RA2111050010019]  Date: |
| S S R Subramanya Hemant Konduri[RA2111050010010]  Nagarjuna Ledalla[RA2111050010019] |



# SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR – 603 203

## BONAFIDE CERTIFICATE

Certified that 18CSP107L - Minor Project [18CSP108L- Internship] report titled “**Sentiment Analysis of Twitter Feed Data using CNN, BERT and BiLSTM**” is the bonafide work of **S S R Subramanya Hemant Konduri[RA2111050010010], Nagarjuna Ledalla[RA2111050010019]** who carried out the project work[internship] under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

**Dr.M.Ramprasath Dr. Kavitha V**

|  |  |  |
| --- | --- | --- |
| **SUPERVISOR**  **Assistant Professor**  DEPARTMENT OF  DATA SCIENCE AND BUSINESS SYSTEMS |  | **PROFESSOR &HEAD**  DEPARTMENT OF  DATA SCIENCE AND BUSINESS SYSTEMS |

**ACKNOWLEDGEMENTS**

We express our humble gratitude to **Dr. C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr. T. V. Gopal,** Dean-CET, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. Revathi Venkataraman**, Professor and Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We encompass our sincere thanks to, **Dr. M. Pushpalatha**, Professor and Associate Chairperson, School of Computing and **Dr. C.Lakshmi,** Professor and Associate Chairperson, School of Computing, SRM Institute of Science and Technology, for their invaluable support.

We are incredibly grateful to our Head of the Department, **Dr. V. Kavitha**, Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Project Coordinators, Panel Head, Dr.M.Ramprasath and Panel Members, Dr.V.Prasanna, Dr.S.Nandhini, Dr.M.Rajakani, Dr.K.S.Arikumar, Mrs.B.Yasotha, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, Dr. Nadana Ravishankar, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, Dr. M. Ramprasath, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for providing us with an opportunity to pursue our project under his mentorship. He provided us with the freedom and support to explore the research topics of our interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank all the staff and students of Department of Data Science and Business Systems, School of Computing, S.R.M Institute of Science and Technology, for their help during our project. Finally, we would like to thank our parents, family members, and friends for their unconditional love, constant support, and encouragement

Authors

S S R Subramanya Hemant Konduri

Nagarjuna Ledalla

**ABSTRACT**

This research investigates the application of deep learning techniques to sentiment analysis of Twitter feed data, utilizing a hybrid model that combines Convolutional Neural Networks (CNN) with Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory (BiLSTM) networks. With the exponential rise in social media interactions, platforms like Twitter have emerged as influential channels for public sentiment, capturing real-time feedback on current events, products, policies, and social issues. The unstructured nature of Twitter data, characterized by abbreviations, slang, and informal language, poses significant challenges for traditional sentiment analysis models. Conventional approaches based on shallow machine learning algorithms and basic word embeddings (such as Word2Vec or GloVe) often lack the depth to fully capture complex linguistic features and contextual subtleties in social media text. This study addresses these limitations by adopting a multi-model architecture that capitalizes on CNN, BERT, and BiLSTM.

The model architecture begins with BERT for word embedding and context-aware feature extraction. BERT, a transformer-based model pre-trained on large corpora, offers contextually rich embeddings that are highly effective in understanding sentiment within sentences, as it considers word order and surrounding context, unlike traditional embeddings. The CNN layer is then applied to extract localized features, especially valuable in recognizing patterns in short texts like Twitter posts. CNN’s strength lies in identifying key phrases or words that may indicate strong sentiments, thereby complementing BERT’s capabilities. Following CNN, BiLSTM processes the text sequence in both forward and backward directions, capturing dependencies across time, and improving sentiment detection by recognizing the relationship between past and future words in a sentence. This sequential dependency, essential for understanding the structure of informal language, aids in fine-tuning sentiment polarity even in complex or nuanced statements.Experiments were conducted on a substantial dataset of Twitter feed data, preprocessed for optimal input by normalizing text, removing noise such as hashtags, and handling emoticons and other unique Twitter characteristics. The hybrid CNN-BERT-BiLSTM model was compared with several baseline architectures, including single BERT, CNN-LSTM, and traditional machine learning approaches, using metrics such as accuracy, F1-score, precision, and recall to evaluate performance. The results demonstrated that the proposed architecture achieved superior accuracy, with enhanced F1-scores indicative of better balance between precision and recall in both positive and negative sentiment classes. The model’s robustness in handling complex expressions, sarcasm, and idiomatic phrases—challenges commonly seen in Twitter language—was also noted.The study underscores the importance of combining transformer-based embeddings with CNN and BiLSTM models to achieve a more nuanced understanding of sentiment in short texts, particularly within the context of social media. This approach proves effective in capturing and classifying sentiment, facilitating improved decision-making, and offering valuable insights across domains like social media monitoring, brand analysis, political forecasting, and crisis management. By demonstrating the efficacy of this hybrid model, the research contributes to advancements in sentiment analysis techniques and sets the stage for further explorations into leveraging deep learning for real-time social media insights.

**I**

**TABLE OF CONTENTS**

[**ABSTRACT**](#_bookmark1) **I**

[**LIST OF FIGURES**](#_bookmark2) **IV**

**LIST OF TABLES V**

[**ABBREVIATIONS**](#_bookmark3) **VI**

**CHAPTER TITLE PAGE**

**NO. NO.**

[**1 INTRODUCTION**](#_bookmark4) **1**

* 1. Introduction 2
  2. Problem Statement 3
  3. Objectives of the Project 4

[**2**](#_bookmark4) **LITERATURE REVIEW 5**

[**3 SYSTEM ARCHITECTURE AND DESIGN**](#_bookmark20) **6**

**4 RESEARCH METHODOLOGY 21**

4.1 Architecture Document 40

4.2 Outcome of objectives/ Result Analysis 41

[**5**](#_bookmark4) **SPRINT RETROSPECTIVE** 42

[**6**](#_bookmark4) **EXPERIMENTAL SETUP**

**7 RESULTS AND DISCUSSIONS 40**

7.1 Project Outcomes 41 7.2 Sub Title 42

**8 CONCLUSION AND FUTURE ENHANCEMENT 45**

**9 REFERENCES 46**

**APPENDIX**

**A CODING 48**

**B CONFERENCE PUBLICATION 50**

**C JOURNAL PUBLICATION 51**

**D PLAGIARISM REPORT 52**

**LIST OF FIGURES**

**CHAPTER TITLE PAGE**

**NO. NO.**

##### [2.1 Thresholding segmentation in action on the skin lesion image input](#_bookmark12) . . 4

[2.2 **Computer Vision Pipeline**](#_bookmark16). . . . . . . . . . . . . . . . . . . . . . . . . . 5

vii

**LIST OF TABLES**

**CHAPTER TITLE PAGE**

**NO. NO.**

##### [2.1 Thresholding segmentation in action on the skin lesion image input](#_bookmark12) . . 4

[2.2 **Computer Vision Pipeline**](#_bookmark16). . . . . . . . . . . . . . . . . . . . . . . . . . 5

* 1. [**ROC curve CNN and dermatologists**](#_bookmark22) . . . . . . . . . . . . . . . . . . . . 8
  2. [**confusion matrix with CNN vs doctors**](#_bookmark26) . . . . . . . . . . . . . . . . . . . 10

**viii**

**ABBREVIATIONS**

**AES** Advanced Encryption Standard

**ANN** Artificial Neural Network **CNN** Colvonutional Neural Network **CSS** Cascading Style Sheet

**CV** Computer Vision

**DB** Database

**DNA** Deoxyribo Neucleic Acid

**GCP** Google Cloud Platform

**HAM** Human Against Machine **HTML** Hyper Text Markup Language **HTTP** Hyper Text Transfer Protocol **JS** Javascript

**KNN** K Nearest Neighbours

**MNIST** Modified National Institute of Standards and Technology

**PWA** Progressive Web App

**RNA** Ribo Neucleic Acid

**ROC** Receiver Operating Characteristic

**SASS** Syntactically Awesome Style Sheets **SMOTE** Synthetic Minority Oversampling Technique **SQL** Structured Query Language

**SVM** Support Vector Machine

**UI** User Interface

**UV** UltraViolet

**UX** User Experience

**YOLO** You Only Look Once

ix

**CHAPTER 1**

**INTRODUCTION**

* 1. **Introduction**

In the era of digital communication, social media platforms have become pivotal in shaping public discourse, influencing societal trends, and even impacting political and economic decisions. Twitter, with its microblogging format and real-time interactions, has emerged as a significant platform for capturing public sentiment on diverse topics ranging from social and political issues to product reviews and personal opinions. Sentiment analysis, the computational task of determining whether a piece of text expresses positive, negative, or neutral sentiment, has thus become a crucial tool for researchers and analysts seeking to extract actionable insights from large volumes of social media data. However, the nature of Twitter data presents unique challenges for sentiment analysis, as it is characterized by short, informal text that often includes slang, abbreviations, emojis, and misspellings. Additionally, the complex and dynamic language use on Twitter, including sarcasm, irony, and cultural references, complicates the task of accurately determining sentiment.

Traditional sentiment analysis approaches have primarily relied on machine learning models such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees, often combined with basic word embeddings like Word2Vec and GloVe. While these methods have achieved reasonable results in some text classification tasks, they are limited in their ability to capture nuanced contextual information and sequential dependencies, especially in short texts like tweets. To address these limitations, recent research has focused on deep learning architectures, particularly those based on Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and transformers such as BERT (Bidirectional Encoder Representations from Transformers). These models are capable of capturing intricate language patterns and context, which are essential for more accurate sentiment classification in challenging environments like Twitter.

* 1. **Problem Statement**

The increasing reliance on social media platforms like Twitter as sources of public opinion and sentiment presents a significant challenge for traditional sentiment analysis methods. Twitter’s informal language structure, including abbreviations, slang, emojis, and rapid shifts in topics, makes it difficult to capture and interpret sentiment accurately. Conventional sentiment analysis models, which often use basic machine learning algorithms and static word embeddings, struggle to process these complexities, resulting in reduced accuracy when dealing with the nuances of social media text. These models fail to capture contextual dependencies, limiting their ability to interpret the sentiment in posts with subtle or ambiguous expressions such as sarcasm or idioms, both of which are common on Twitter.

To address these challenges, there is a need for more sophisticated approaches that combine advanced deep learning models capable of understanding complex and context-sensitive language. By integrating CNN, BERT, and BiLSTM, this research seeks to create a hybrid model that leverages each component’s strengths in capturing both local features and contextual embeddings, as well as sequential dependencies. This approach is designed to enhance the robustness and accuracy of sentiment analysis on Twitter, providing a more reliable tool for extracting actionable insights from real-time public sentiment across various applications, including market analysis, political forecasting, and social research.

* 1. **Objectives of the Project**

The primary objective of this project is to develop a robust and accurate sentiment analysis model for Twitter feed data using a hybrid approach that combines Convolutional Neural Networks (CNN), Bidirectional Encoder Representations from Transformers (BERT), and Bidirectional Long Short-Term Memory (BiLSTM) networks. By integrating these models, the project aims to address the limitations of traditional sentiment analysis techniques in handling the nuanced and context-dependent language typically found on social media platforms like Twitter. This objective is driven by the need for a model that not only captures individual word meanings but also understands context, tone shifts, and language complexities, such as sarcasm and idioms, which are common in tweets.

A key objective is to leverage BERT’s transformer-based architecture to generate context-aware embeddings that retain the relative meaning of each word within its sentence. This allows the model to handle linguistic variability, understanding how the sentiment associated with a word or phrase can change based on its context within a sentence. By using BERT as a foundational layer, the project seeks to improve the accuracy of sentiment detection in tweets, where the intended sentiment may not be obvious from a single word or phrase but becomes clear within the broader textual context.

Another objective is to incorporate CNN for feature extraction at a granular level. The CNN layer is expected to capture localized features within tweets, such as phrases or expressions that convey strong sentiment, either positive or negative. CNN’s ability to identify patterns in short sequences makes it particularly useful for detecting sentiment cues in the compact text structure of tweets. By capturing these localized features, the project aims to enhance the model’s ability to identify sentiment in short, informal text formats without relying on extensive sentence structure or grammar.

The final objective is to apply BiLSTM for capturing sequential dependencies across the tweet, allowing the model to understand how sentiment flows forward and backward through the text. Given that sentiment in tweets can often shift between positive and negative tones or build on preceding words, the BiLSTM layer is intended to enhance the model’s sensitivity to such temporal dependencies. This will enable the model to provide more accurate sentiment classification by recognizing patterns of sentiment change within tweets.

Through these objectives, the project aims to create a comprehensive sentiment analysis model capable of producing high accuracy and reliability on Twitter data, with the potential for application across domains where real-time sentiment tracking and analysis are valuable, including market analysis, public opinion research, and crisis management.

**CHAPTER 2**

**LITERATURE SURVEY**

* 1. **LITERATURE SURVEY FOR RESEARCH**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Title**  *(Name of the journal, author and publication details)* | **Methodology**  *(Provide a Summary of key studies and their findings)* | **Identification of gaps and limitations.** *(Identify the limitations of the Research Paper)* |
| 1 | Agarwal, A., Xie,, Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, pp.30-38, Portland, Oregon | * Used distant learning for sentiment data acquisition using emoticons for positive and negative sentiment. Compared Naive Bayes, MaxEnt, and SVM classifiers, with SVM performing best. Found that unigram models outperformed bigram and POS features. * Focused on subjective vs. objective classification. Found that both POS and bigrams helped, contradicting Go et al.'s findings. * Used noisy labels from websites for model training and manually labeled tweets for tuning and testing. Employed features like retweets, hashtags, and punctuation. Their approach showed that combining prior polarity with POS was effective. * Emphasized the role of linguistic features like POS tags. Demonstrated that abstract linguistic analysis features contributed   significantly to classifier accuracy. | * **Data Bias**: The training and testing data were collected via search queries, introducing potential bias. * **Limited Feature Exploration**: Emoticons and hashtags showed only marginal improvements, suggesting a need for further exploration of other Twitter-specific features. * **Manual Annotation**: Although extensive, manual annotation can be subjective and inconsistent. * **Generalizability**: The models and features tested might not generalize well to other types of social media or different languages. |

|  |  |  |  |
| --- | --- | --- | --- |
| 2 | ” P. Chakraborty, U. S. Pria, M. R. A. H. Rony and M. A. Majumdar, "Predicting stock movement using sentiment analysis of Twitter feed," 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology  (ICIEV-ISCMHT), Himeji, Japan, 2017, pp. 1-6, doi: 10.1109/ICIEV.2017.8338584. | * Used polarity predictions from three websites as noisy labels and syntax features of tweets, achieving improved classifier performance by integrating syntax and polarity features. * Showed that linguistic features like POS tags contribute significantly to classifier accuracy in sentiment analysis on feedback data from Global Support Services surveys. | * The paper does not mention how the test data was collected. * The study is limited by the potential inaccuracies introduced by using Google Translate for non-English tweets. * The "junk" labeled tweets indicate translation issues, leading to data quality concerns. * The classifier's performance improvement using Twitter- specific features is marginal, suggesting limited utility in practical applications. * The paper relies heavily on manually annotated data, which may not scale well for   larger datasets. |
| 3 | R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 208-211, doi:  10.1109/ICECA.2018.8474783 | * Sentiment analysis of Twitter data is conducted using supervised machine learning techniques such as Naive-Bayes, SVM, and Maximum Entropy . * Different sentiment analysis methodologies were compared, showing varying accuracy rates based on techniques and datasets used . * Semantic analysis with WordNet followed   by machine learning techniques increases accuracy by 4-5% . | * The study requires extensive data cleaning, scraping, and integration, which increases overhead . * Inefficiency in real-time analytics and the time- consuming process for analyzing large datasets were noted as major limitations . |

|  |  |  |  |
| --- | --- | --- | --- |
| 4 | V. Prakruthi, D. Sindhu and D. S. Anupama Kumar, "Real Time Sentiment Analysis Of Twitter  Posts," *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2018, pp. 29-34, doi: 10.1109/CSITSS.2018.8768774 | * Various papers discuss systems for retrieving and classifying Twitter data based on semantics . * Real-time Twitter data stream frameworks provide quick feedback through opinion mining . * The sentiment analysis of tweets using R language and Rhadoop Connector handles large data volumes effectively . * Emoji and slang detection techniques improve sarcasm detection algorithms in sentiment analysis . * A pixel cell-based sentiment calendar visualizes opinions, displaying positive, neutral, and negative sentiments with color- coded cells . | * The high complexity of data extraction and cleaning processes poses significant challenges . * Integration of qualitative analysis with large-scale data mining techniques needs further enhancement . |
| 5 | H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, India, 2016, pp. 416- 419, doi: 10.1109/ICATCCT.2016.7912034 | * Twitter sentiment analysis utilizes Hadoop for processing large movie review datasets . * Different sections present positive, negative, and neutral sentiments extracted from Twitter data . * Twitter's popularity and data mining capabilities make it a prime source for opinion mining . | * Traditional databases are inadequate for processing large amounts of unstructured data in a specified time . * Existing systems for real-time analytics are inefficient and time-consuming, requiring better handling of big data   complexities . |
| 6 | Chandra Gupta Maurya, Sudhanshu Kumar Jha, Sentiment Analysis: A Hybrid Approach on Twitter Data, Procedia Computer Science, Volume 235, 2024, Pages 990-999, ISSN 1877-0509,  https://doi.org/10.1016/j.procs.2024.04.094 | * Achieved 92.66% accuracy using CNN, LSTM, FastText, and GRU on streaming data but faced data bias issues. * Found SVM (85%), KNN (64%), and Naïve Bayes (80%) accuracy on a Twitter dataset,   only examined association, not causation. | * Limited set of features, excluding commonly used emojis on Twitter . * Narrow scope, findings not generalizable to other text   data. |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | * Attained 95% accuracy using MLTSA and SVM on multilingual Twitter data, highlighted need for advanced feature engineering. * Achieved 63% accuracy with Naïve Bayes, Decision Tree, SVM on small Twitter   datasets, limiting generalizability. | * Neglected model sensitivity to hyperparameters, important for replication. * Small dataset limited generalizability to other datasets or domains. |
| 7 | V. Mahalakshmi, P. Shenbagavalli, S. Raguvaran, V. Rajakumareswaran, E. Sivaraman, Twitter sentiment analysis using conditional generative adversarial network, International Journal of Cognitive Computing in Engineering, Volume 5, 2024, Pages 161-169, ISSN 2666-3074,  https://doi.org/10.1016/j.ijcce.2024.03.002 | * Analyzed Twitter data to understand leadership prioritization, beneficial to society. * Used Twitter-based mining for content management, aiding in extracting useful information. * Focused on classifying social media sentiments, highlighting rapid information spread. * Built sentiment analysis models using machine learning techniques, found Bag of Words more accurate than TF-IDF for tweet classification . | * Limited by the scope of analyzing leadership prioritization only. * Limited to extracting and managing Twitter content, may not be generalizable. * Focused on classification, may not address underlying reasons for sentiment spread. * Limited by the keywords used, potentially missing relevant tweets not containing the specified keywords; lacked computational resources to process a larger number of   tweets . |
| 8 | Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* **55**, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1 | * The article discusses various methods of performing sentiment analysis and details several applications. * Some of the key findings are that supervised machine learning methods are often used for sentiment analysis, and that there are several challenges associated with sentiment analysis, such as sarcasm and   irony. | There are a number of challenges associated with sentiment analysis, including:   * Sarcasm and irony * Multiple meanings of words * Contextual dependence * Different languages * Different cultures |

|  |  |  |  |
| --- | --- | --- | --- |
| 9 | Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi, Sentiment analysis and classification of Indian farmers’ protest using twitter data, International Journal of Information Management Data Insights, Volume 1, Issue 2, 2021, 100019, ISSN 2667-0968,  https://doi.org/10.1016/j.jjimei.2021.100019 | * **Farmers' protest sentiment analysis:** Majority of tweets analyzed were neutral, with positive sentiments close behind and negative sentiments last. * **Random Forest model:** Found to be the most effective in classifying sentiment during the farmers' protest among the models tested . | * **Keywords limitation:** Relevant tweets might be missed if specific keywords are not used in the messages. * **Sample size:** The study could have analyzed a larger number of tweets to uncover more insights but was limited   by computational resources . |
| 10 | Veny Amilia Fitri, Rachmadita Andreswari, Muhammad Azani Hasibuan, Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm, Procedia Computer Science, Volume 161, 2019, Pages 765-  772, ISSN 1877-0509,  https://doi.org/10.1016/j.procs.2019.11.181 | * Researchers preferred general sentiment lexicons over domain-specific ones for SA tasks due to easier construction and universal applicability . * Conventional machine learning, especially SVM, is frequently used for SA tasks due to its high accuracy with high-dimensional text features . * Bi-LSTM is commonly used in SA tasks because of its ability to capture long-term dependencies and effective training on large datasets . * Transformer-based pre-trained models like BERT perform well in SA tasks due to pre- training on large-scale data, reducing training time, and capturing contextual dependencies . | * Establishing robust datasets in multiple languages with ethical standards is necessary for better research . * The detection of hidden emotions, sarcasm, and irony remains an open research question in SA . * Challenges include context dependency, high dimensionality, redundancy, and handling slang in feature extraction for SA . * Addressing cross-domain accuracy improvement and handling multilingual data are significant open research   topics in SA . |
| 11 | Hanhoon Kang, Seong Joon Yoo, Dongil Han,Senti- lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, Expert Systems with Applications, Volume 39, Issue 5, 2012, Pages 6000-6010, ISSN 0957-4174,  https://doi.org/10.1016/j.eswa.2011.11.107 | * Different methods and applications of SA are categorized into document, phrase, and aspect levels to provide a comprehensive understanding . | * Aspect-based SA models often perform poorly on datasets from different domains due to domain adaptation issues . * Invalid aspect extraction can decrease performance, |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | * Document-level SA treats each document as an independent object with one sentiment polarity, useful for coarse-grained analysis . * Sentence-level SA aims to classify sentences into positive, negative, or neutral, helping in detailed sentiment categorization . * Aspect-level SA focuses on fine-grained sentiment analysis, achieving significant performance improvements despite domain   adaptation issues . | highlighting the need for better contextual understanding .   * Collecting correlation information between sentences and context is crucial for accurate aspect term extraction . |
| 12 | Yanying Mao, Qun Liu, Yu Zhang, Sentiment analysis methods, applications, and challenges: A systematic literature review, Journal of King Saud University - Computer and Information Sciences, Volume 36, Issue 4, 2024, 102048, ISSN 1319-1578,  https://doi.org/10.1016/j.jksuci.2024.102048 | * Aspect-based SA models often perform poorly on datasets from different domains due to domain adaptation issues . * Invalid aspect extraction can decrease performance, highlighting the need for better contextual understanding . * Collecting correlation information between sentences and context is crucial for accurate aspect term extraction . | * There is a need for better datasets, especially those that are well-annotated, finely graded, and comply with ethical standards . * Identifying sarcasm and ridicule in text remains a complex challenge due to the implicit nature of these emotions . * Context dependency and handling contextually charged words are significant   challenges in SA . |
| |  |  |  |  | | --- | --- | --- | --- | |  |  | * The study provides insights into public perceptions of Turkish higher education by analyzing tweets about universities and their educational quality. * Previous studies evaluated the performance of Word2Vec, BiLSTM, SVM-RF hybrid techniques, and other models for Turkish tweet analysis, achieving varying degrees of accuracy . | generalizability of the findings.   * The research is largely focused on a specific dataset and domain (Turkish universities), potentially limiting the broader applicability of the findings. * The paper does not thoroughly address the challenges in data collection and preprocessing, which could impact the model's   performance and reliability. |   13 | Abdulfattah Ba Alawi, Ferhat Bozkurt, A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities sing Twitter data, Decision Analytics Journal, Volume 11, 2024, 100473, ISSN 2772-6622,  https://doi.org/10.1016/j.dajour.2024.100473 | * The paper proposes a novel hybrid model for analyzing Turkish texts using BiLSTM, CNN, and BERT-based transformers, outperforming state-of-the-art models with accurate predictions. * It compares various machine learning classifiers and deep learning models like LSTM, BiLSTM, CNN, and Hybrid CNN-   BiLSTM. | * The study does not provide a comprehensive exploration of the limitations specific to the hybrid model's application in various contexts . * There is limited discussion on the potential biases in the data or the models used,   which might affect the |

### **CHAPTER 3**

### **SYSTEM ARCHITECTURE AND DESIGN**

### 

**1. Input Layer: Preprocessing Twitter Data**

The input layer handles the initial data preprocessing, which is essential for normalizing Twitter’s unique text structures, such as abbreviations, slang, emojis, and hashtags. Text preprocessing involves several key steps:

* **Tokenization**: Each tweet is broken down into tokens, which can be single words, hashtags, emojis, or punctuation marks.
* **Noise Removal**: Unnecessary symbols, user mentions, and URLs are removed to focus on meaningful text.
* **Emoji and Slang Handling**: Emojis and slang words are either retained for sentiment emphasis or translated into textual descriptions, as they are often critical in conveying tone on Twitter.

Once the text is preprocessed, it is ready to be passed to the BERT model for context-rich embedding generation.

**2. BERT Embeddings Layer: Context-Aware Feature Generation**

BERT (Bidirectional Encoder Representations from Transformers) is the first major layer in the model. It generates high-dimensional, context-aware embeddings for each token in the text. BERT operates bidirectionally, meaning it processes each word in both directions (left-to-right and right-to-left), allowing it to capture nuanced dependencies between words. For example, in a phrase like "not very good," BERT recognizes the impact of "not" on "good," capturing the phrase’s negative sentiment despite the presence of a typically positive word.

BERT embeddings are pre-trained on large corpora, which enables it to understand language context exceptionally well, even in cases of informal or ambiguous expressions. These embeddings are crucial for Twitter sentiment analysis, as they provide a richer representation of each token, especially when handling contextually charged phrases and sarcasm.

**3. CNN Layer: Local Feature Extraction**

Following BERT, the embeddings are passed through a Convolutional Neural Network (CNN) layer, which identifies localized features within the text. CNN is particularly effective at detecting patterns within fixed-sized chunks of data, known as **n-grams**. For example, CNN can capture phrases like "highly recommend" or "not satisfied," which often convey strong sentiment. By sliding its filters across the BERT embeddings, CNN extracts sentiment-specific patterns that highlight important phrases and clusters of words that contribute to sentiment.

This layer consists of several convolutional filters that apply different levels of scrutiny to the text, generating feature maps that capture critical sentiment cues. For short-form text like tweets, CNN’s localized feature extraction is effective in capturing sentiment-related information without requiring an understanding of the entire sentence structure, which is often fragmented or non-standard on Twitter.

**4. Pooling Layer: Dimensionality Reduction**

After CNN processes the text, a max-pooling layer is applied to the feature maps. Max-pooling reduces the dimensionality of the data by retaining only the most prominent features from each feature map. This reduction process simplifies the data while preserving the most critical sentiment-related features, enhancing the model’s computational efficiency and preventing overfitting.

**5. BiLSTM Layer: Sequential Dependency Learning**

The pooled features are then passed into a Bidirectional Long Short-Term Memory (BiLSTM) layer. BiLSTM is designed to capture sequential dependencies in text by processing it in both forward and backward directions, which is particularly useful for understanding how sentiment evolves throughout the sentence. Unlike traditional LSTM, which only processes data in one direction, BiLSTM’s bidirectional nature allows it to retain information from both preceding and succeeding words, enabling the model to better interpret nuanced sentiment.

For example, in a sentence like “The product is good, but customer service is poor,” BiLSTM captures the sentiment shift from positive to negative, which a unidirectional model might miss. The BiLSTM layer provides a more comprehensive view of the sentiment flow across the text, essential for handling Twitter’s mixed-tone and short-form sentences.

**6. Fully Connected Layer: Higher-Level Feature Integration**

After BiLSTM refines the sentiment context, the output is passed to a fully connected (dense) layer. This layer combines the features extracted from BERT, CNN, and BiLSTM, generating a high-level feature representation that is suitable for classification. The fully connected layer integrates these diverse features, enhancing the model’s understanding of the overall sentiment conveyed by the tweet.

**7. Output Layer: Sentiment Classification**

The final layer is a softmax output layer, which assigns a probability to each sentiment category (e.g., positive, negative, neutral). The softmax function transforms the dense layer’s output into a probability distribution, allowing the model to make a definitive sentiment classification. This classification output is then used to label the tweet’s sentiment based on the probability scores for each category.

**Summary of Architectural Advantages**

**1. Context-Awareness through BERT**

BERT (Bidirectional Encoder Representations from Transformers) provides the model with an in-depth, context-aware understanding of language, crucial for Twitter’s informal, abbreviated, and sometimes ambiguous text. Unlike traditional word embeddings, which assign a fixed meaning to each word, BERT generates embeddings that depend on each word's context within the sentence. This capability is essential for sentiment analysis, as the sentiment of a word can vary greatly depending on its surrounding words. For example, in the phrase “not very helpful,” BERT understands that "helpful" carries a negative sentiment due to the preceding “not.” This context-awareness is particularly useful for interpreting Twitter language, where sentiment is often implied through word placement and surrounding terms. By using BERT, the model gains a significant edge in handling the subtleties of Twitter data, capturing even sarcastic or ironic sentiments that may be challenging for traditional models.

**2. Localized Feature Detection with CNN**

Convolutional Neural Networks (CNNs) excel at identifying localized features in text, which is especially beneficial for short and concise formats like tweets. Sentiment on Twitter is often conveyed through specific phrases or word clusters, such as “love it” or “absolutely terrible.” CNN’s sliding filters capture these sentiment-laden patterns by scanning BERT-generated embeddings for clusters of words that convey strong opinions. This ability to detect sentiment within short, specific segments of text makes CNN an ideal component for Twitter sentiment analysis, where expressions are often brief and impactful. By focusing on local features, CNN layers help the model quickly identify and emphasize key phrases that reveal sentiment, allowing for more efficient and accurate analysis.

**3. Sequential Dependency Capture with BiLSTM**

The inclusion of a Bidirectional Long Short-Term Memory (BiLSTM) layer enables the model to understand sequential dependencies in both directions (forward and backward), which is crucial for accurately capturing the sentiment flow within a tweet. Unlike CNN, which captures only localized patterns, BiLSTM processes the full text sequence, identifying how words influence each other across the sentence. For instance, in the sentence “The product is great, but the service is terrible,” BiLSTM captures the shift in sentiment from positive to negative, something a unidirectional model might miss. This bidirectional processing is particularly valuable for handling Twitter’s mixed-tone sentences, where opinions may shift from positive to negative or vice versa. As a result, BiLSTM enhances the model’s accuracy in classifying tweets that contain complex sentiment expressions.

**4. Dimensionality Reduction and Computational Efficiency**

The use of a max-pooling layer after the CNN layer helps to reduce dimensionality, focusing the model’s attention on the most prominent sentiment features. Max-pooling retains the highest-value features from each feature map, thereby reducing the data’s complexity while preserving essential information. This not only decreases the model’s computational load but also helps prevent overfitting, which is particularly valuable when dealing with high-dimensional BERT embeddings. The pooling layer’s ability to streamline data processing makes the model more efficient without sacrificing accuracy, allowing it to handle large volumes of Twitter data.

**5. Comprehensive Feature Integration with the Dense Layer**

After extracting and refining features from BERT, CNN, and BiLSTM, the dense (fully connected) layer integrates these diverse features, creating a high-level feature representation that the model can use for classification. This layer combines local patterns from CNN, context-aware embeddings from BERT, and sequential information from BiLSTM, producing a unified representation that is highly informative for sentiment classification. By combining these insights in a fully connected layer, the model can achieve a more holistic understanding of each tweet, accounting for both localized sentiment cues and overall sentiment context. This integration maximizes the model’s performance by ensuring that no single aspect of the sentiment expression is overlooked.

**6. Improved Accuracy in Complex Sentiment Detection**

The combination of context-aware, localized, and sequentially dependent layers results in a model that is highly adept at handling complex sentiment expressions, including sarcasm, mixed tones, and idiomatic phrases. On Twitter, sentiment is often expressed through subtle cues or evolving statements, which can make classification difficult for simpler models. The hybrid architecture, however, is designed to manage these complexities by combining layers that focus on context, local features, and the overall sentence structure. As a result, the model is more capable of accurately identifying positive, negative, and neutral sentiments, even in challenging cases.

**7. Versatility and Applicability Across Domains**

The CNN-BERT-BiLSTM architecture is not only effective for Twitter sentiment analysis but is also versatile enough to be applied across other domains that involve informal or contextually rich language, such as product reviews, feedback analysis, and customer service data. The model’s ability to capture subtle linguistic patterns, combined with its efficiency in processing large datasets, makes it adaptable to various real-world applications. This versatility is particularly advantageous for domains where sentiment insights are critical for decision-making, such as brand monitoring, political analysis, and public opinion research.

**REFERENCES**

1. Wang, Yili et al. (2022). Sentiment Analysis of Twitter Data.
2. Wankhade, K. et al. (2019). Sentiment Analysis on Twitter Data Using Machine Learning Techniques.
3. S. Symeonidis et al. (2018). Feature Engineering for Twitter Sentiment Analysis.
4. N. Jain et al. (2020). Deep Learning for Twitter Sentiment Analysis.
5. Bansal, S. et al. (2020). Sentiment Analysis on Twitter Data Using Bi-LSTM.
6. Silva, M. et al. (2020). Twitter Sentiment Analysis Using BERT.
7. Zhang, Y. et al. (2021). Attention Mechanisms in Sentiment Analysis for Tweets.
8. Patel, H. et al. (2020). Hybrid Machine Learning Approaches for Twitter Sentiment Analysis.
9. Gupte, M. et al. (2014). The Impact of Hashtags on Twitter Sentiment Analysis.
10. Go, A. et al. (2016). Twitter Sentiment Classification Using Distant Supervision.
11. Abdullah, M. et al. (2019). SVM and Neural Networks for Real-Time Twitter Sentiment Analysis.
12. Mittal, A. et al. (2019). Social Media Monitoring Using Sentiment Analysis of Tweets.
13. Agarwal, A. et al. (2011). Sentiment Analysis of Twitter Data Using Tree Kernels.
14. Dai, X. et al. (2019). Preprocessing Techniques for Twitter Sentiment Analysis.
15. Zubiaga, A. et al. (2018). Analyzing Sentiments in Real-Time Events via Twitter.
16. Wang, C. et al. (2016). Twitter Sentiment Analysis Using Recurrent Neural Networks.
17. Adwan, O. et al. (2018). Sentiment Analysis of Political Tweets.
18. Oliveira, A. et al. (2020). Twitter Sentiment Analysis for Brand Monitoring.
19. Mekala, R. et al. (2021). Transfer Learning for Twitter Sentiment Analysis.
20. Li, J. et al. (2019). Multi-Aspect Sentiment Analysis of Tweets Using CNN.
21. Asghar, M. et al. (2020). Comparative Analysis of ML Algorithms for TSA.
22. Cambria, E. et al. (2016). Affective Computing for Twitter Sentiment Analysis.
23. Liu, B. (2020). Lexicon-Based Sentiment Analysis in Twitter.
24. Hutto, C. et al. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.
25. Yadav, V. et al. (2021). TSA Using Ensemble Learning.
26. Siddharth, N. et al. (2021). Long-Short Term Memory Networks for TSA.
27. Chowdhury, S. et al. (2019). Emotion Detection in Tweets Using BERT.
28. Habimana, T. et al. (2020). Twitter Sentiment Analysis for Disaster Response.

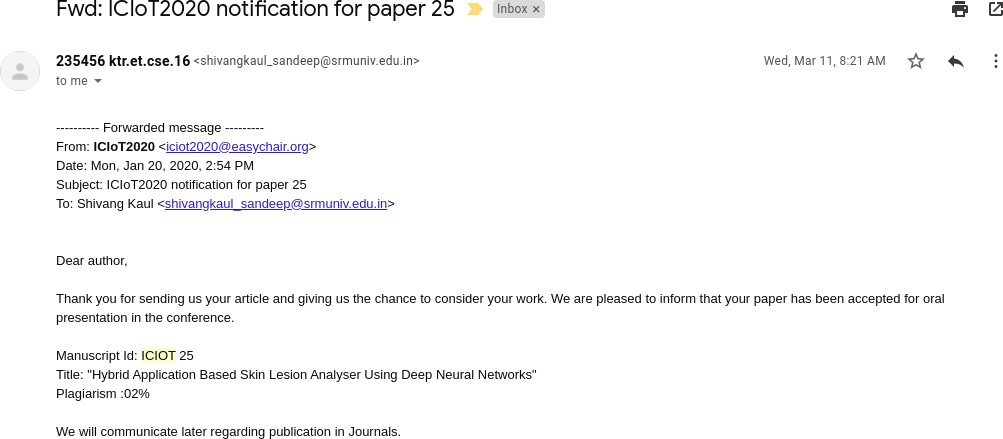
**APPENDIX A**

**CODING**

**APPENDIX B**

**CONFERENCE PRESENTATION**

Our paper on **Hybrid application based skin lesion analyzer using deep neural networks** was presented at ICIOT 2020 conference held at SRM. 200+ shortlisted teams presented their papers on various fields in the conference. Our paper got accepted as paper id : 25 with a plagiarism of just 2 %.



##### Figure A.1: ICIOT 2020 Acceptance

On presenting the paper in this international conference held at SRM KTR campus, we received positive remarks and suggestion from the judging panel. We were then awarded the best paper award at the same conference.

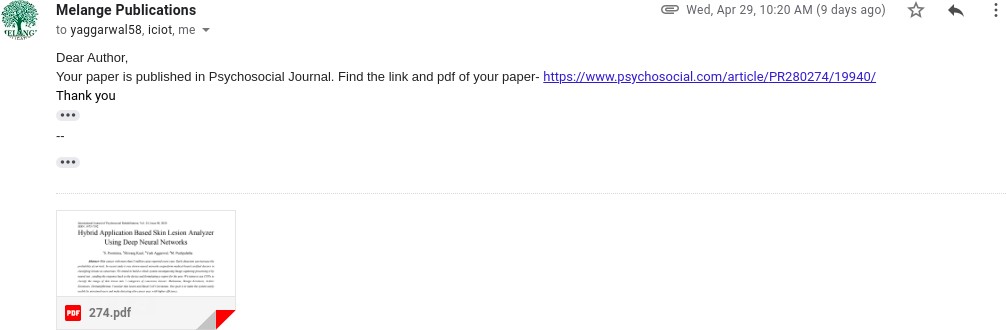
##### Figure A.2: ICIOT 2020 Best Paper award

83

**APPENDIX C**

**PUBLICATION DETAILS**

We submitted our research paper for publication at IJPR publication house puducherry. We had selected the journal **International Journal of Psychosocial Rehabilitation (ISSN: 1475- 7192)**. We got the acceptance notification from the IJPR stating our paper has been published in the April Issue of the same journal. Proof of publication is attached in figure [B.1](#_bookmark123) The research



##### Figure B.1: Publication Notification

paper cover page has been attached below.

International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020 ISSN: 1475-7192

Hybrid Application Based Skin Lesion Analyzer Using Deep Neural Networks

1S. Poornima, 2Shivang Kaul, 3Yash Aggarwal, 4M. Pushpalatha

***Abstract--****Skin cancer with more than 5 million cases reported every year. Early detection can increase the probability of survival. In recent study it was shown neural networks outperform medical board certified doctors in classifying lesions as cancerous. We intend to build a whole system encompassing Image capturing processing it by neural net , sending the response back to the device and formulating a report for the user. We intent to use CNNs to classify the image of skin lesion into 7 categories of cancerous lesions: Melanoma, Benign Keratosis, Actinic Keratoses, Dermatofibroma, Vascular skin lesion and Basal Cell Carcinoma. Our goal is to make the system easily usable by untrained users and make detecting skin cancer easy with higher efficiency.*

***Key words--****Neural Networks, Image Processing, Convolu-tional Neural Networks, Skin Cancer Detection, Skin Lesion Imaging, App Development, Localization Algorithms, Cloud Computing, GCP, Compute Engine, App Engine.*

1. **INTRODUCTION**

Skin Cancer is a major kind of cancer with around 5 million reported cases worldwide every year. The major cause of skin cancer is exposure to UV rays. Diagnosing skin cancer generally included the skin lesion being examined by a doctor. Recent studies have shown neural networks to be more efficient in classifying lesion as cancerous as compared to trained doctors. Misdiagnosing or late detection of cancer can lead to a higher mortality rate and less chance of cure. The goal of this project is making detection and classification of lesions on the skin easier. Not all the marks on skin are a matter of concern but early detection and treatment of cancer can save lives. So this gives the user a way to check if there’s a chance of the mark on your skin being cancerous. The aim of this project is to detect and analyse such a correlation using neural networks. It is expected that the outcome of this project will lead to automated classification of skin lesions.

1. **LITERATURE SURVEY**

The following papers were read and analysed for the refer-ence of this paper. A brief image has been presented here.

1) Andre Esteva et al. 2017,” Dermatologist-level classification of skin cancer with deep neural networks.” Contribution: Claimed to classify skin lesions at par with board trained dermatologists. Methodology used:

*1Assistant Professor, CSE Department, SRMIST, Chennai, India*

*2Assistant Professor, CSE Department, SRMIST, Chennai, India,* [*shangkaul@gmail.com*](mailto:shangkaul@gmail.com) *3Assistant Professor, CSE Department, SRMIST, Chennai, India,* [*yaggarwal58@gmail.com*](mailto:yaggarwal58@gmail.com) *4Assistant Professor, CSE Department, SRMIST, Chennai, India*

**DOI: 10.37200/IJPR/V24I8/PR280274**

**Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020 2545**

**APPENDIX D**

**PLAGIARISM REPORT**