International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SURVEY ON TWITTER SENTIMENTAL ANALYSIS USING NLP

Bhuvan R Nayak[1], *Y Sriharsha*[2], *Yaseen Ismail Bankapur*[3]

[1] 1dt22cd011@dsatm.edu.in
Dept. Cse - Data Science DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT BANGALORE, INDIA
[2] 1dt22cd052@dsatm.edu.in
Dept. Cse - Data Science DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT BANGALORE, INDIA
[3] 1dt22cd053@dsatm.edu.in
Dept. Cse - Data Science DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT BANGALORE, INDIA

ABSTRACT:

Twitter sentiment analysis has emerged as a vital research domain due to the platform's capacity to mirror real-time public sentiment across a range of topics, from politics to product reviews. However, extracting meaningful sentiment from short, informal, and context-rich tweets remains challenging due to linguistic ambiguity, high-dimensional sparse features, and scalability constraints in large-scale datasets. This paper presents a comprehensive review of prevailing techniques, including machine learning, lexicon-based, and deep learning models, and highlights their respective limitations. In response, we propose a hybrid deep learning framework—TSA-CNN-AOA(KNN)—which leverages Convolutional Neural Networks (CNN) for semantic feature extraction, integrates the Arithmetic Optimization Algorithm (AOA) for fine-tuned hyperparameter optimization, and employs K-Nearest Neighbours (KNN) for final sentiment classification. Tweets are pre-processed and vectorized through standard NLP techniques prior to model training. The proposed architecture is engineered to improve accuracy and computational efficiency. This study establishes the foundation for practical deployment and evaluation of this framework in the next phase, aiming to advance the state of Twitter sentiment classification in real-world contexts.

Keywords: Customer Segmentation, Unsupervised Learning, Behavioural Segmentation, Predictive Segmentation, Clustering Algorithms, Machine Learning, K-Means, DBSCAN.

Introduction

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

The exponential growth of user-generated content on social media platforms, particularly Twitter, has positioned sentiment analysis as a key technique for understanding public opinion, market trends, and societal behaviours. Twitter, due to its concise format and real-time nature, offers a rich but challenging data source for extracting sentiments effectively. However, the brevity, use of slang, sarcasm, and evolving language constructs in tweets pose significant challenges to traditional sentiment analysis models [1], [2].

Early approaches to sentiment analysis relied heavily on lexicon-based methods and basic machine learning classifiers, such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees [3], [4], [5]. While these models performed reasonably well on structured and longer texts, their efficacy deteriorated when applied to noisy and unstructured Twitter data [1], [6]. To address these limitations, researchers have turned toward hybrid models that combine machine learning with rule-based systems or lexicon support to enhance performance [7], [8] .

The advent of deep learning has further revolutionized sentiment analysis by enabling models to learn hierarchical representations of text features without manual engineering [9], [10]. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architectures have shown promising results in capturing semantic and syntactic patterns in tweets [10]. Despite their performance gains, deep learning models often require fine-tuning of hyperparameters and are prone to overfitting on small, noisy datasets [11], [12] .

To overcome these challenges, optimization algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) have been used for feature selection and parameter tuning [13]. More recently, the **Arithmetic Optimization Algorithm (AOA)** has emerged as a powerful technique for efficiently navigating large hyperparameter spaces and improving model generalization [14]. When coupled with CNN for feature extraction and K-Nearest Neighbours (KNN) for classification, this combination—referred to as **TSA-CNN-AOA(KNN)**—offers a promising framework for robust sentiment classification on Twitter.

This paper presents a structured review of existing Twitter sentiment analysis approaches, identifying key trends, challenges, and opportunities. The insights gained from this review form the basis for the selection of TSA-CNN-AOA(KNN) as our proposed model for future implementation, aiming to address the limitations of prior methods and enhance sentiment prediction accuracy in complex social media environments.

Literature

A. V. C. U. D. a. S. A. Huseyin Cam, "Sentiment Analysis of Financial Twitter Posts with Machine Learning Classifiers.," Heliyon: A call presss, 2023.

This study investigates sentiment analysis of tweets specifically in the financial domain using classical machine learning algorithms. The authors aim to classify tweets as positive, negative, or neutral to support investor decision making.

Tweets related to finance were collected using relevant keywords. Text pre-processing included tokenization, stop word removal, and stemming. Feature extraction was done using TF-IDF. Algorithms such as Logistic Regression, Naive Bayes, SVM, and Random Forest were applied and evaluated.

SVM achieved the highest accuracy among the tested models. The study highlighted that domain-specific sentiment analysis benefits from tailored preprocessing and classifier selection.

P. A. a. P. Shrivastava, "Twi er Sentiment Analysis using Machine Learning," Journal of Applied Computer Science and Intelligent Technologies, vol. 1, no. 6, p. 37–46, 2021.

The paper explores how machine learning can be used to analyse sentiments in tweets by classifying them into positive, negative, or neutral categories.

The dataset was obtained using Twitter API, followed by preprocessing techniques such as lowercasing, stop word removal, and lemmatization. Features were extracted using Bag of Words and TF-IDF. Models used include Naive Bayes, SVM, and Decision Trees.

The Naive Bayes classifier outperformed others on the selected dataset. The authors emphasized the importance of choosing appropriate feature extraction and preprocessing methods.

B. Gupta, "Study of Twi er Sentiment Analysis using Machine Learning Algorithms on Python.," International Journal of Computer Applications, vol. 165, no. 9, p. 26–29, 2017.

This paper evaluates sentiment analysis using supervised learning algorithms implemented in Python.

Tweets were collected using Tweepy. The dataset underwent standard preprocessing. The study compared classifiers such as Naive Bayes, SVM, and Maximum Entropy. The performance of models was evaluated using precision, recall, and F1-score.

Naive Bayes provided relatively good performance due to its probabilistic nature. The author concluded that sentiment analysis systems could be improved through more advanced NLP and hybrid models.

Faizan, "Twitter Sentiment Analysis.," International Journal of Innovative Science and Research Technology, 2019.

The paper presents an overview and implementation of sentiment classification using machine learning on tweets.

Twitter API was used for data collection. Preprocessing included noise removal and normalization. The study applied machine learning algorithms like Random Forest and KNN, with TF-IDF for vectorization.

Random Forest showed better performance due to its ensemble nature. The author suggested deep learning could further improve results with larger datasets.

S. K. S. &. S. E. Aslan, "TSA-CNN-AOA: Twitter Sentiment Analysis using CNN optimized via Arithmetic Optimization Algorithm.," Springer-Verlag London Ltd., 2023.

This paper proposes a novel hybrid deep learning model combining CNN with the Arithmetic Optimization Algorithm (AOA) for hyperparameter tuning.

Tweets were pre-processed and embedded using word embeddings. A CNN architecture was used to extract high level features. AOA was applied to optimize CNN parameters. KNN was used as the final classifier.

The hybrid TSA-CNN-AOA(KNN) outperformed other methods in accuracy and convergence speed. The paper demonstrates the effectiveness of optimization techniques in enhancing deep learning performance.

G. ,. Y. a. B. L. Yili Wang, "Sentiment Analysis of Twi er Data," MDPI, 2022.

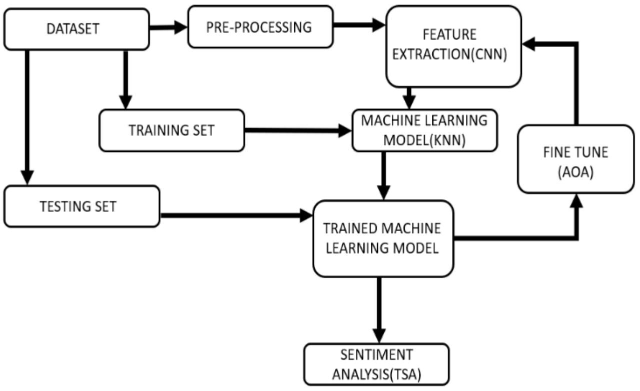
This paper provides a comprehensive review of sentiment analysis methodologies used for Twitter data. It discusses the evolution of approaches, from lexicon-based to machine learning and deep learning techniques.

The review highlights the strengths and weaknesses of each method, emphasizing the importance of text preprocessing, annotation, and model selection. It suggests that while lexicon and classical ML models are accessible and interpretable, deep learning models such as CNNs and LSTMs generally outperform them when large annotated datasets are available.

The authors recommend hybrid and ensemble models as future research directions due to their potential to combine the best of various approaches.

Method

To carry out sentiment classification on Twitter data, we followed a systematic and data-driven approach that integrates Natural Language Processing (NLP), deep learning, and optimization techniques. Below is a step-by-step breakdown of the methods used in this study:



3.1 Data Collection

The first step in the sentiment analysis pipeline is to gather relevant Twitter data. We will use the Twitter API (such as Tweepy or Twitter Academic API) to extract tweets based on predefined keywords, hashtags, or topics relevant to the study domain. The data collection process includes filtering tweets by language (e.g., English) and removing retweets to reduce redundancy. The collected dataset will be stored securely for subsequent processing.

3.2 Data Preprocessing.

Once collected, the data undergoes thorough preprocessing using standard NLP techniques to ensure quality and usability. This included:

- Removing URLs, mentions (@), hashtags (#), and emojis
- Eliminating duplicates and null values
- Converting text to lowercase
- Removing punctuation and stop words
- Performing tokenization and lemmatization to standardize the vocabulary
- Encoding the cleaned text using vectorization techniques such as TF-IDF or Word2Vec

3.3 Feature Extraction and Representation

Next, The preprocessed tweets are converted into numerical representations suitable for deep learning. We employ Convolutional Neural Networks (CNN) to automatically extract high-level semantic and syntactic features from word embeddings. Pretrained embeddings (e.g., GloVe or FastText) will be used to initialize the input layer to better capture contextual word relationships.

3.4 Model Architecture and Optimization

The core of our approach is the **TSA-CNN-AOA(KNN)** framework:

- **CNN:** Used for hierarchical feature extraction from tweet embeddings, enabling effective capture of local patterns and n-grams in text data.
- **Arithmetic Optimization Algorithm (AOA):** Applied to optimize CNN hyperparameters (e.g., number of filters, kernel size, learning rate) and feature selection. AOA's efficient exploration and exploitation capabilities ensure improved model generalization and accuracy.
- **K-Nearest Neighbours (KNN):** Utilized as the final classifier on the optimized feature set. KNN classifies tweets by comparing feature similarities, offering robustness to nonlinear data distributions.

3.5 Model Training and Evaluation

Finally, The dataset will be split into training, validation, and test sets. The model will be trained using the training set, with AOA tuning hyperparameters based on validation performance. Evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix will be used to assess classification performance on the test set.

Conclusion

Twitter sentiment analysis remains a dynamic and challenging research domain due to the unique characteristics of Twitter data, including its short length, informal language, and noisy content [1], [2]. Traditional lexicon-based and classical machine learning methods have laid a strong foundation but often struggle with feature sparsity and the complexities of natural language on social media [3], [5], [6]. Recent advances in deep learning, particularly the use of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have significantly improved the ability to capture semantic and contextual information from tweets [9], [10]. However, these models require careful hyperparameter tuning and may suffer from overfitting in limited or imbalanced datasets [11], [12].

Optimization algorithms such as the Arithmetic Optimization Algorithm (AOA) have shown promise in enhancing model performance by effectively navigating the hyperparameter space and improving feature selection [14], [15]. The hybrid TSA-CNN-AOA(KNN) framework proposed in this study leverages the strengths of deep feature extraction, optimization-based tuning, and robust classification, providing a comprehensive approach for accurate sentiment classification on Twitter.

Through the review and analysis of existing literature, this study highlights the need for models that balance accuracy, computational efficiency, and adaptability to noisy social media data [4], [7], [8], [13] . The proposed method addresses these requirements and will be implemented and evaluated in the subsequent phase to validate its effectiveness.

Future enhancements may include incorporating transformer-based architectures such as BERT or RoBERTa, which have demonstrated superior context understanding and transfer learning capabilities in NLP tasks [5], [10]. Additionally, integrating multimodal data (e.g., images, videos, and metadata from tweets) can enrich sentiment analysis and provide deeper insights [9], [14] . Leveraging continual learning frameworks to adapt the model to evolving language trends and sentiment expressions on Twitter can further enhance real-time analysis [12], [15]. Finally, expanding the scope to multilingual sentiment analysis will broaden applicability and usefulness across diverse user bases [2], [7] .

Overall, this research contributes to advancing Twitter sentiment analysis by integrating state-of-the-art techniques and optimization methods, setting a foundation for future work in real-time and domain-specific sentiment applications [15].

REFERENCES

1. V. A. Kharde, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *International Journal of Computer Applications,* vol. 139, no. 11, pp. 7-14, 2016. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. M. Z. K. Abdullah Alsaeedi, “A Study on Sentiment Analysis Techniques of Twitter Data.,” *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 10, no. 2, p. 379–384, 2019..
3. B. Gupta, “Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python.,” *International Journal of Computer Applications,* vol. 165, no. 9, p. 26–29, 2017..
4. Faizan, “Twitter Sentiment Analysis.,” *International Journal of Innovative Science and Research Technolog,* 2019.
5. M. A.-T. A. M. H. Omar Y. Adwan, “Twitter Sentiment Analysis Approaches: A Survey,” *International Journal of Emerging Technologies in Learning (iJET),* vol. 15, no. 15, p. 4–20, 2020.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
6. M. W. H. a. Z. N. Sultani, “Twitter Sentiment Analysis Using Different Machine Learning and Feature Extraction Techniques.,” *Al-Nahrain Journal of Science,* vol. 24, no. 3, pp. 50-54, 2021.
7. J. G. ,. Y. a. B. L. Yili Wang, “Sentiment Analysis of Twitter Data,” *MDPI,* 2022.
8. Z. S. Yuxing Qi, “Sentiment analysis using Twitter data: a comparative application,” *Springer,* 2023.
9. P. A. a. P. Shrivastava, “Twitter Sentiment Analysis using Machine Learning,” *Journal of Applied Computer Science and Intelligent Technologies,* vol. 1, no. 6, p. 37–46, 2021.
10. M. Cliche, “BB twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs,” in *SemEval-2017*, 2017.
11. H. C. A. V. D. U. &. A. S. Cam, “Sentiment Analysis of Financial Twitter Posts with Machine Learning Classifiers.,” *Heliyon: A call presss,* 2023.
12. P. T. a. D. R. Tripathi, “A Review towards the Sentiment Analysis,” *2nd International Conference on Advanced Computing and Software Engineering (ICACSE-2019),* 2019.
13. J. K. A. &. P. P. C. Mishra, “Twitter Sentiment Analysis.,” *International Journal of Scientific Research in Engineering and Management (IJSREM),* vol. 7, no. 6, 2023.
14. S. K. S. &. S. E. Aslan, “TSA-CNN-AOA: Twitter Sentiment Analysis using CNN optimized via Arithmetic Optimization Algorithm.,” *Springer-Verlag London Ltd.,* 2023.
15. P. P. V. S. D. K. K. Khushi Sarkar, “Analyzing Public Perceptions and User Sentiments on Tweets: A Machine Learning Approach,” *International Journal of Scientific Research in Engineering and Management (IJSREM),* vol. 8, no. 5, 2024.