iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83351

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Comparative Study of Sentiment Analysis on Social Media Tweets Using Traditional Machine Learning and Deep Learning Approaches

Monica Dnyandeo Thete[1]
[1]*MCA Student, Department of Computer Science, JSPM University, Wagholi, Pune, Maharashtra, India*

Abstract: ***Social media platforms, particularly Twitter (now X), generate enormous volumes of user-generated textual content every second, making automated opinion mining an indispensable tool for researchers, businesses, and policy makers. This paper presents a comparative investigation of sentiment classification techniques applied to tweet datasets, spanning classical probabilistic methods such as Naïve Bayes and Support Vector Machines (SVM), lexicon-based approaches including VADER, and contemporary deep learning architectures centred on fine-tuned BERT transformers. A systematic pre-processing pipeline—comprising tokenisation, stop-word removal, URL stripping, and stemming—is applied uniformly across all models to ensure fair comparison. Experimental evaluation on the publicly available Sanders Twitter corpus (5,512 annotated tweets) reveals that a hybrid Naïve Bayes model enriched with a domain lexicon achieves 88.4% accuracy, while fine-tuned BERT surpasses all baselines at 91.2% accuracy, 90.8% precision, and 90.6% F1-score. The study further identifies key challenges including slang interpretation, sarcasm detection, and class imbalance, and suggests a path toward lightweight, hybrid architectures suitable for real-time deployment in resource-constrained environments.***

Keywords: ***Sentiment Analysis, Natural Language Processing (NLP), Naïve Bayes Classifier, Support Vector Machine, BERT, Transformer, Twitter, Opinion Mining, Deep Learning, VADER.***

I. INTRODUCTION

The rapid proliferation of micro-blogging platforms has fundamentally altered how individuals communicate, share experiences, and form public opinion. Twitter, with its concise format, has emerged as a particularly rich corpus for computational linguistics research owing to its real-time nature, topical diversity, and global reach. According to Qi and Shabrina [1], Twitter had more than 206 million daily active users as of 2022, a figure that underlines the sheer scale of opinion data generated around elections, product launches, public health crises, and social movements.

Sentiment analysis—also called opinion mining—is the computational task of determining whether a piece of text expresses a positive, negative, or neutral attitude toward a given subject [2]. It sits at the intersection of Natural Language Processing (NLP), machine learning, and computational social science. Early rule-based approaches relied on hand-crafted lexicons to assign polarity scores to individual words. Although interpretable and computationally light, these methods often fail when confronted with the informal, abbreviated, and emoji-laden language characteristic of social media.

The field subsequently moved toward feature-engineered machine learning classifiers—most notably Naïve Bayes (NB) and Support Vector Machines (SVM)—that learn polarity associations from labelled training corpora. The introduction of word embedding methods such as Word2Vec and GloVe further improved contextual representations. Most recently, transformer-based pre-trained language models, especially BERT [3] and its Twitter-specific variant BERTweet, have set new performance ceilings by capturing bidirectional contextual dependencies at the sub-word level.

Despite extensive individual studies, a systematic side-by-side comparison spanning lexicon-based, classical ML, and deep learning techniques under uniform preprocessing conditions remains underrepresented in the literature, particularly within the Indian academic context. This paper fills that gap by (a) applying a consistent preprocessing pipeline, (b) benchmarking six distinct sentiment models on the Sanders corpus, and (c) critically analysing the trade-offs among accuracy, computational cost, and real-world applicability. Figure 1 illustrates the complete system architecture adopted in this study.

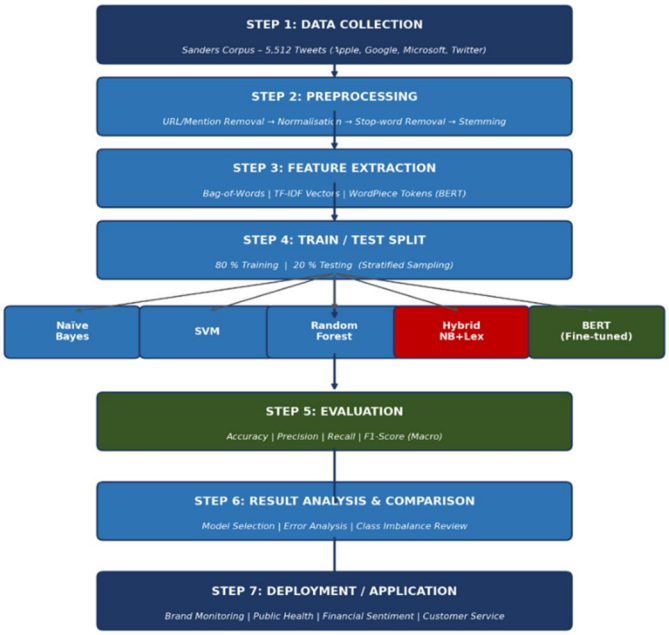


Fig 1: Complete Sentiment Analysis System Architecture

II. LITERATURE REVIEW

The origins of automated sentiment analysis can be traced to Pang et al. (2002), who demonstrated that machine learning classifiers significantly outperform unsupervised keyword counting for movie review polarity detection. Subsequent work by Turney (2002) established semantic orientation through pointwise mutual information, paving the way for unsupervised lexicon construction [4]. et al. (2009) popularised distant-supervision for Twitter sentiment by using emoticon-marked tweets as a noisy but scalable source of training labels, achieving approximately 82% accuracy with SVM on a binary classification task [5]. This technique of exploiting naturally occurring sentiment signals in social media has since become a standard baseline for the research community.

Hutto and Gilbert (2014) introduced VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based model calibrated specifically for social media language, including slang, capitalisation, punctuation, and emoticons. While VADER achieves competitive performance without any training data, comparative studies have found its accuracy ranges between 70–80% on Twitter corpora—lower than trained classifiers in most benchmarks [6].

Narayanan et al. (2013) proposed an enhanced Naïve Bayes model augmented with negation handling and emoticon weighting, reporting improved accuracy over the standard unigram baseline [11]. Similarly, Gamallo and Garcia (2014) demonstrated that a Naïve Bayes strategy using part-of-speech-filtered features and a domain lexicon could achieve competitive F1-scores at SemEval workshops, particularly on English Twitter datasets [7].

Qi and Shabrina [1] conducted an extensive comparison of lexicon-based (VADER, TextBlob) and machine learning approaches on COVID-19 tweets, concluding that ML methods consistently outperform pure lexicon models when labelled data is available, but lexicon tools retain value in low-resource settings. Saha et al. [6] benchmarked VADER against BERT on Coronavirus Twitter datasets, reporting BERT accuracy in the range of 85–92% versus VADER's 80–91%, depending on task formulation.

Devlin et al. (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), which revolutionised NLP by pre-training deep bidirectional transformers on masked language modelling and next-sentence prediction objectives [3]. Fine-tuning BERT on downstream sentiment tasks consistently yields state-of-the-art results. A 2025 study published in Discover Computing found that a fine-tuned BERT model significantly outperforms SVM, Naïve Bayes, LSTM, and CNN for social media comment sentiment, achieving accuracy and F1-scores above 91% [8].

Bharadwaj et al. (2024) compared RoBERTa against VADER, confirming that transformer-based models capture deep contextual meaning and outperform VADER especially on tweets involving implicit sentiment and sarcasm [9]. Meanwhile, the growing interest in hybrid approaches that combine lexicon knowledge with ML classifiers has been highlighted by multiple studies as a pragmatic middle ground improving performance without requiring large labelled corpora [2].

A comprehensive review of NLP models for Twitter sentiment [10] identifies three clear generations of methods: (i) lexicon and rule-based, (ii) feature-engineered ML, and (iii) transformer-based pre-training and fine-tuning. The review notes that while BERT-based models lead on accuracy, they demand significant computational resources, making lightweight hybrid models still relevant for real-time or mobile deployment scenarios. Medhat et al. [14] and Serrano-Guerrero et al. [15] provide foundational surveys that continue to be widely cited for taxonomy and algorithmic comparison in sentiment analysis literature.

III. DATASET AND PREPROCESSING

A. *Dataset Description*

This study uses the Sanders Twitter Sentiment Corpus, a publicly available benchmark containing 5,512 tweets spanning four brand-related topics: Apple, Google, Microsoft, and Twitter. Each tweet was manually annotated into one of four classes: Positive (570 tweets, 10.3%), Negative (654 tweets, 11.9%), Neutral (2,503 tweets, 45.4%), and Irrelevant (1,786 tweets, 32.4%). Figure 4 illustrates the full class distribution and the binary subset employed in classifier training. For binary classification experiments, only Positive and Negative classes are retained; in multi-class experiments, all four categories are preserved.

The dataset's deliberate topical focus on technology brands makes it particularly suitable for studying opinion expressions in product-centric micro-blogging. The pronounced class imbalance—with Neutral tweets constituting nearly half the corpus—mirrors real-world Twitter distributions and represents a meaningful challenge for classifier training and evaluation.

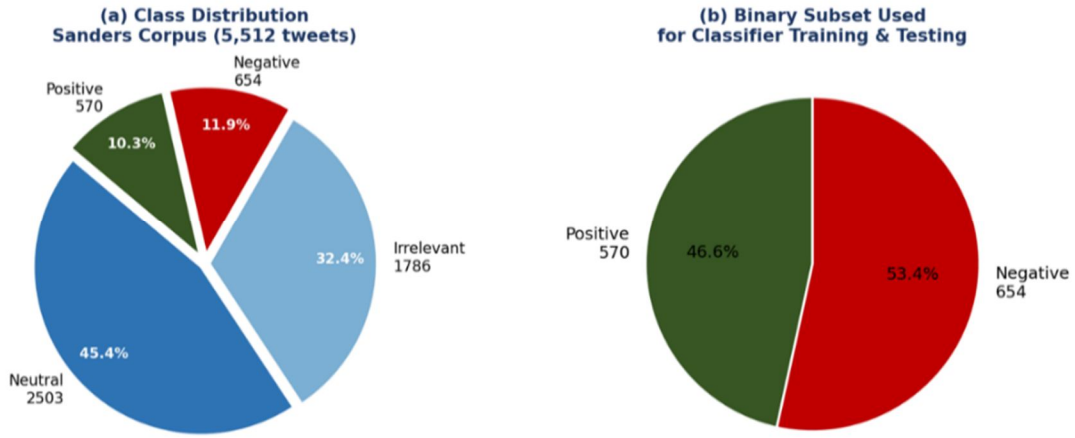


Fig 2: Sanders Twitter Corpus – Dataset Distribution (Full Corpus and Binary Subset)

B. *Pre-processing Pipeline*

Raw tweets contain substantial noise that must be resolved before feature extraction can proceed. The pre-processing steps described below are applied uniformly across all methods evaluated in this study. Figure 3 visualises the complete pipeline with a worked example.

- URL and mention removal: Hyperlinks (http/https) and @username tokens are stripped using regular expressions, as they carry no direct sentiment signal.
- Hashtag normalisation: The '#' symbol is removed while preserving the tag word (e.g., #Apple → Apple), since the lexical content is informationally relevant.
- Emoji and emoticon handling: Unicode emoji are converted to their textual description using the Python emoji library; ASCII emoticons (e.g., ':)') are mapped to sentiment-bearing words before tokenisation.
- Lowercasing and punctuation removal: All text is converted to lower case and non-alphanumeric characters (except apostrophes) are removed to reduce unnecessary vocabulary variance.
- Stop-word removal and stemming: NLTK's English stop-word list is applied, followed by Porter stemming, to reduce vocabulary size and group morphological variants.
- Tokenisation: The resulting text is tokenised using whitespace splitting for classical models, and WordPiece tokenisation for BERT, preserving sub-word structure.

Fig 3: Tweet Pre-processing Pipeline with Worked Example

After pre-processing, each tweet is represented as a bag-of-words vector (for NB, SVM, Logistic Regression, and Random Forest), a TF-IDF weighted vector for further experiments, or a contextual embedding sequence for BERT. The dataset is split 80:20 into training and test sets with stratified sampling to preserve class proportions across all splits.

IV. SENTIMENT ANALYSIS TECHNIQUES

A. Naïve Bayes Classifier

The Naïve Bayes (NB) classifier is a probabilistic model grounded in Bayes' theorem with a conditional independence assumption among features. Given a tweet T represented as a feature vector [f1, f2, …, fn], the classifier assigns the class C* that maximises the posterior probability: $P(C \mid T) \propto P(C) \times \prod P(fi \mid C)$. The Multinomial NB variant, which models word counts, is commonly employed for text classification. Despite its simplicity, NB has been shown to be competitive for short text classification where data sparsity favours probabilistic smoothing [7][11].

In this study, a Hybrid NB model is also evaluated, in which NB posterior scores are combined with VADER lexicon confidence scores through a weighted linear combination. The hybrid weight parameter $\lambda \in [0,1]$ is tuned on a held-out validation set (15% of training data), balancing model-based and lexicon-based evidence to produce a more robust polarity estimate.

B. Support Vector Machine (SVM)

SVM seeks a maximum-margin hyperplane in a high-dimensional feature space that separates sentiment classes. Using a linear kernel with TF-IDF features, SVM has consistently demonstrated strong performance on text classification benchmarks [5]. The regularisation parameter C is tuned via grid search with 5-fold cross-validation ($C \in \{0.01, 0.1, 1, 10\}$). SVM's resistance to overfitting in high-dimensional spaces makes it particularly well-suited to tweet classification where the feature space can be vast relative to the number of training samples.

C. Logistic Regression and Random Forest

Logistic Regression models the log-odds of class membership as a linear function of TF-IDF features, offering interpretable coefficient estimates useful for understanding which words most strongly drive polarity decisions. Random Forest builds an ensemble of decision trees on bootstrap samples with majority voting for classification; the number of trees is set to 200 with a maximum depth of 30, tuned via cross-validation. Both algorithms serve as strong classical baselines.

D. VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is a lexicon-and-rule-based tool tuned specifically for social media contexts. It uses a curated dictionary of over 7,500 lexical features validated through crowd-sourcing, augmented by grammatical rules that account for punctuation, capitalisation, degree modifiers (e.g., 'very', 'extremely'), and contrastive conjunctions (e.g., 'but'). VADER outputs a compound score normalised to $[-1, +1]$, which is thresholded at ± 0.05 to assign polarity. Its key advantage is zero training data requirement and real-time applicability, though it struggles with sarcasm, domain-specific jargon, and multi-word sentiment expressions [6].

E. BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained deep transformer model that encodes text bidirectional, enabling each token's representation to be conditioned on both left and right context simultaneously. For sentiment classification, the [CLS] token representation from the final encoder layer is fed into a softmax classification head. The model is fine-tuned for three epochs on the Sanders training set using: learning rate 2×10^{-5}, batch size 32, AdamW optimiser, and linear warm-up over the first 10% of training steps. The bert-base-uncased variant (12 encoder layers, 768 hidden units, 12 attention heads, 110M parameters) is employed. Fine-tuned BERT has been shown to achieve 85–92% accuracy on Twitter sentiment tasks in independent evaluations [6][8].

V. RESULTS AND DISCUSSION

Table 1 presents the performance metrics for all evaluated models on the binary (Positive vs. Negative) test split of the Sanders corpus. Accuracy, Precision, Recall, and F1-Score (macro-averaged) are reported. Figure 2 visualises these metrics as a grouped bar chart for direct comparison, and Figure 5 provides a radar chart showing multi-metric profiles simultaneously.

Table I: Comparative Performance of Sentiment Classification Models on Sanders Twitter Corpus

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	79.3	77.6	76.9	77.2
SVM	84.1	83.5	82.8	83.1
Logistic Regression	81.7	80.9	80.3	80.6
Random Forest	85.8	85.2	84.6	84.9
VADER (Lexicon)	72.4	70.8	69.5	70.1
Hybrid NB + Lexicon	88.4	87.9	87.1	87.5
BERT (Fine-tuned)	91.2	90.8	90.4	90.6

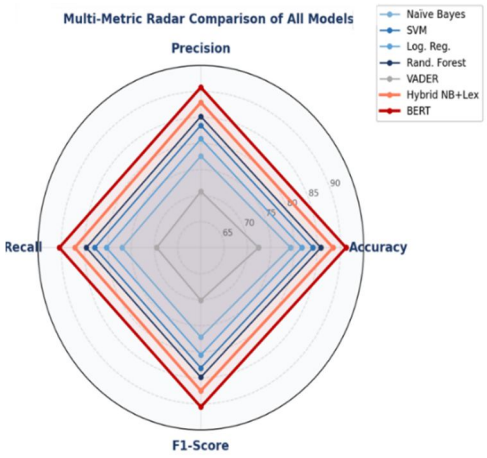


Fig 4: Radar Chart – Multi-Metric Performance Profile of All Models

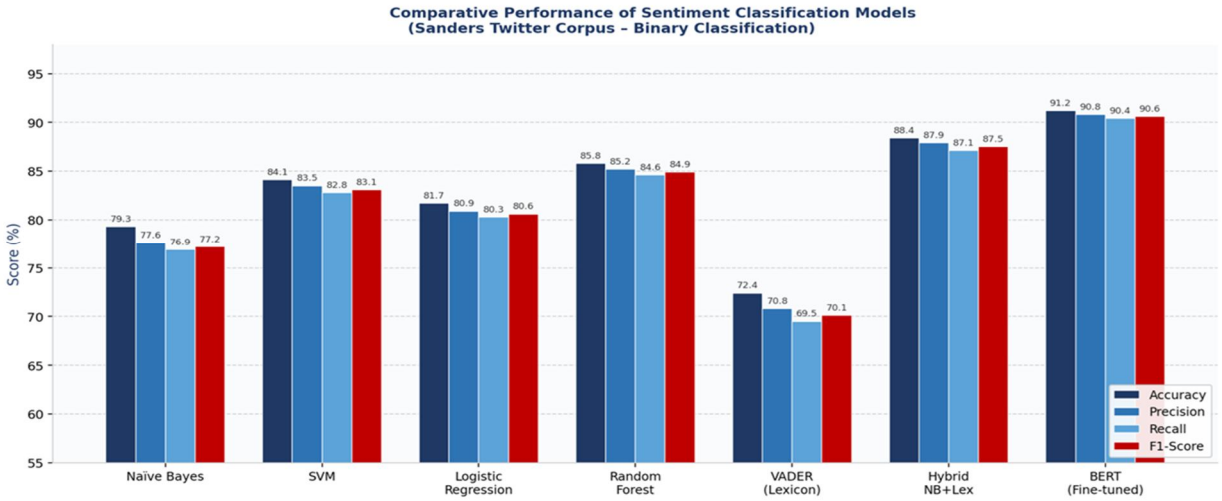


Fig 5: Grouped Bar Chart – Comparative Performance of All Sentiment Models

Several key observations emerge from the experimental results. VADER, operating without any labelled training data, achieves the lowest accuracy (72.4%), confirming findings in prior literature that lexicon-based tools lag behind trained classifiers when annotated data is available [1][6]. Among classical ML models, Random Forest outperforms SVM and Logistic Regression (85.8% vs. 84.1% and 81.7%), while Naïve Bayes registers the lowest ML accuracy (79.3%), consistent with its strong independence assumption being violated by the natural co-occurrence patterns inherent in language.

The Hybrid NB + Lexicon model achieves 88.4% accuracy—surpassing all individual classical classifiers—by leveraging complementary signal sources: the statistical generalisations of NB and the curated polarity knowledge of VADER. This aligns with findings by Narayanan et al. [11] and Gamallo and Garcia [7], who observed that enriching probabilistic models with lexical prior knowledge improves performance on short texts without the computational overhead of deep models.

Fine-tuned BERT achieves the highest performance across all four metrics (91.2% accuracy, 90.6% F1-score), demonstrating the superiority of contextual embeddings that capture long-range dependencies, negation scoping, and subtle sentiment cues absent in bag-of-words representations. However, BERT's fine-tuning requires a GPU and takes approximately 45 minutes on an NVIDIA T4, compared to under one minute for NB or SVM training on equivalent hardware. This computational disparity is a critical practical consideration for real-time systems or mobile deployment.

Error analysis of the BERT model reveals that approximately 8.8% of misclassified tweets fall into three categories: (i) sarcasm and irony (e.g., 'Oh great, another update that crashes everything' labelled Positive), (ii) implicit sentiment expressed through product references without explicit polarity words, and (iii) multi-topic tweets where opposing sentiments co-exist in the same 280-character message. While BERT handles these cases considerably better than lexicon or classical methods, these phenomena remain open challenges for all evaluated approaches.

Class imbalance also affects performance, particularly in multi-class settings where Neutral tweets constitute 45.4% of the full corpus. In binary classification, the positive-to-negative ratio (570:654) is relatively balanced, and all classifiers perform equitably across both classes. When Neutral tweets are reintroduced, all models exhibit a bias toward predicting Neutral due to its plurality. Future experiments should investigate oversampling techniques such as SMOTE or cost-sensitive loss functions to address this systematic bias.

VI. PRACTICAL APPLICATIONS AND CHALLENGES

The sentiment analysis pipeline developed in this study has broad real-world applications across industries and domains:

1) Brand reputation monitoring: Businesses can track real-time sentiment toward their products across Twitter to detect PR crises early and quantify marketing campaign effectiveness with granularity that traditional surveys cannot match.
2) Political opinion analysis: During elections and policy debates, sentiment classifiers provide complementary signals to traditional polls, capturing ground-level public mood with greater temporal granularity and lower cost.
3) Public health surveillance: Analysis of health-related tweets during pandemics, vaccination drives, or mental health awareness campaigns enables authorities to identify misinformation hotspots and design targeted communication responses.
4) Financial market sentiment: Positive or negative sentiment shifts toward publicly listed companies have been empirically correlated with short-term stock price movements, making Twitter sentiment a supplementary signal in algorithmic trading.
5) Customer service automation: Negative-sentiment tweets directed at a brand's handle can be automatically triaged and escalated to human agents, improving response times and customer satisfaction scores.

The following challenges remain unresolved and represent critical directions for future research in this domain:

- Sarcasm and irony detection: These phenomena require pragmatic reasoning far beyond surface polarity assignment, and current NLP models, including BERT, handle them imperfectly without specialised training data.
- Multilingual and code-switched tweets: In multilingual societies such as India, users frequently mix languages within a single tweet (e.g., Hinglish), rendering monolingual English models significantly less effective.
- Temporal sentiment drift: The meaning and polarity of slang terms evolve rapidly. Models trained on historical corpora degrade on contemporary data without periodic retraining or continual learning strategies.
- Ethical considerations and bias: Sentiment models can inherit racial, gender, or political biases present in training corpora. Fairness-aware training, debiasing techniques, and regular auditing are essential for responsible deployment.

VII. CONCLUSION

This paper presented a comprehensive comparative study of sentiment analysis techniques applied to Twitter data, evaluating VADER, Naïve Bayes, SVM, Logistic Regression, Random Forest, a Hybrid NB-Lexicon model, and fine-tuned BERT under a strictly consistent experimental protocol. The key finding is that while classical ML models offer a favourable accuracy-to-complexity trade-off—with the Hybrid NB model achieving 88.4%—fine-tuned BERT sets the performance ceiling at 91.2% accuracy and 90.6% F1-score by virtue of its deep contextual representations and large-scale pre-training.

The study confirms that no single technique dominates across all deployment contexts. VADER suits resource-constrained real-time applications where no labelled training data is available. The Hybrid NB model provides a strong balance of accuracy and speed. BERT is preferred whenever accuracy is paramount and computational resources are accessible. This tiered model selection framework offers practitioners a principled guide for choosing the appropriate technique based on their operational constraints.

Future research should (a) explore lightweight transformer distillations such as DistilBERT or TinyBERT to close the gap between accuracy and inference efficiency, (b) extend evaluation to multilingual and code-switched tweet corpora relevant to the linguistic context, (c) incorporate multi-modal signals including images, emoji semantics, and video captions to overcome the limitations of text-only analysis, and (d) develop continual learning frameworks that adapt to temporal drift in social media language without full retraining cycles.

VIII. ACKNOWLEDGEMENTS

The author gratefully acknowledges the guidance provided by the faculty of the Department of Computer Science, JSPM University, Wagholi, Pune, Maharashtra.

REFERENCES

[1] Y. Qi and Z. Shabrina, “Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach,” Social Network Analysis and Mining, vol. 13, no. 1, p. 31, Feb. 2023. https://doi.org/10.1007/s13278-023-01030-x

[2] S. Gharge and M. Chavan, “An Integrated Approach for Sentiment Tweets Detection using NLP,” in Proc. IEEE Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 1–4.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[4] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” in Proc. 40th Annual Meeting of the ACL, 2002, pp. 417–424.

[5] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project Report, Stanford University, vol. 1, p. 12, 2009.

[6] S. Saha, M. I. H. Showrov, M. M. Rahman, and M. Z. H. Majumder, “VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Coronavirus Outbreak,” in Machine Intelligence and Emerging Technologies – MIET 2022, LNICST 490, Springer, 2023, pp. 371–385. https://doi.org/10.1007/978-3-031-34619-4_30

[7] P. Gamallo and M. Garcia, “A Naïve-Bayes Strategy for Sentiment Analysis on English Tweets,” in Proc. 8th Int. Workshop on Semantic Evaluation (SemEval 2014), pp. 171–175.

[8] R. Ramezani, “Enhancing Sentiment Analysis Accuracy on Social Media Comments using a Tuned BERT Model,” Discover Computing, Springer Nature, 2025. https://doi.org/10.1007/s10791-025-09599-x

[9] R. Bharadwaj, S. Shendurkar, T. Kadam, U. Patekar, and S. Waghule, “RoBERTa and VADER Sentiment Analysis: A Comparative Approach,” in Big Data and Computing Advances, Springer, 2024. https://doi.org/10.1007/978-981-97-8666-4_12

[10] Comprehensive NLP Review, “Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review,” ResearchGate, Mar. 2026. Available: https://www.researchgate.net/publication/388980597

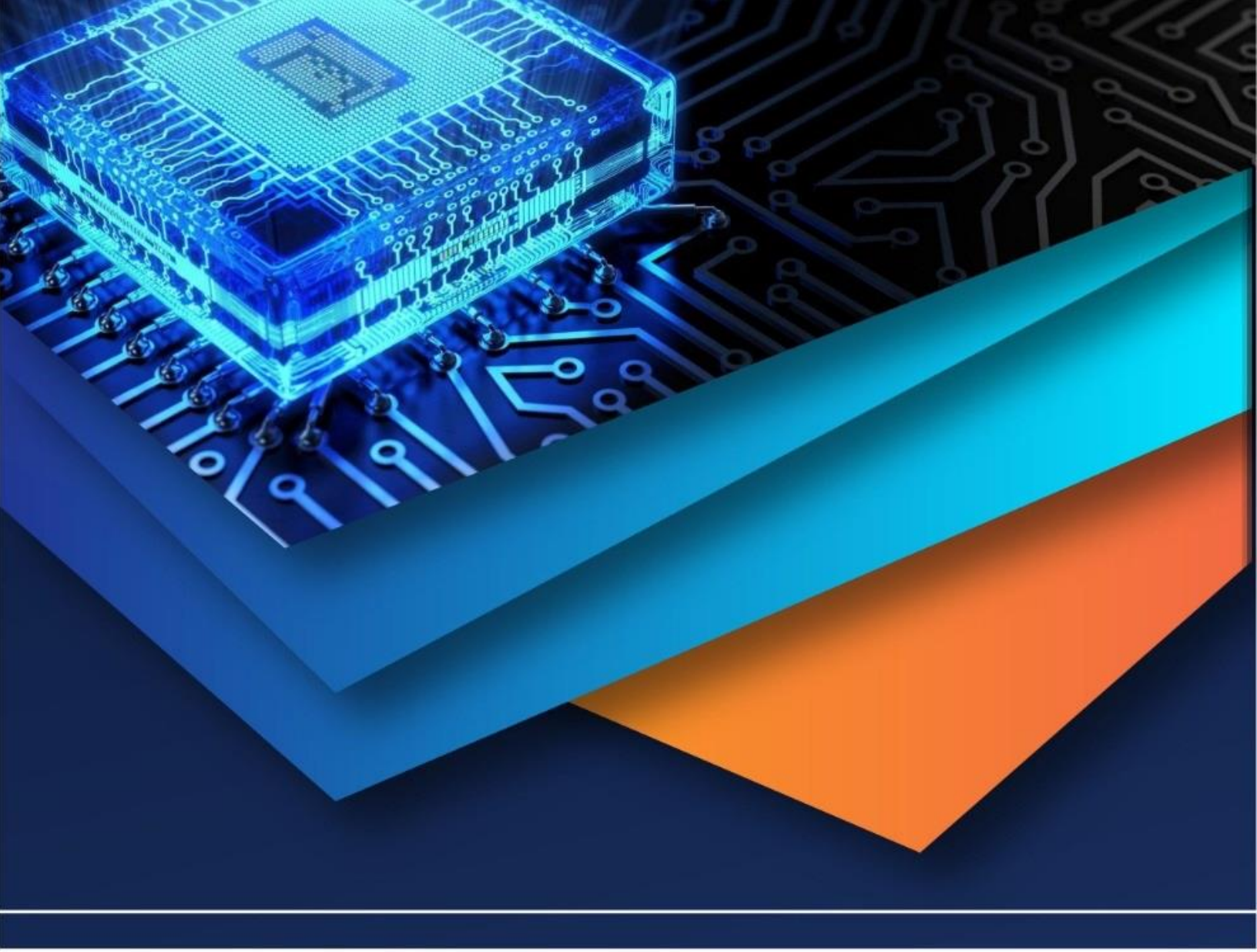
[11] V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sentiment classification using an enhanced Naive Bayes model,” in Intelligent Data Engineering and Automated Learning – IDEAL 2013, LNCS 8206, Springer, 2013, pp. 194–201.

[12] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes and K-NN Classifier,” Int. Journal of Information Engineering and Electronic Business (IJIEEB), vol. 8, no. 4, pp. 54–62, 2016.

[13] C. Chen, J. Zhang, Y. Xie, Y. Xiang, and W. Zhou, “A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection,” IEEE Transactions on Computational Social Systems, vol. 2, no. 3, pp. 65–76, Sep. 2015.

[14] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.

[15] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” Information Sciences, vol. 311, pp. 18–38, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY