IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138
INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Study On Hybrid Twitter Opinion Mining Using Lexicon And Machine Learning Approaches

P Anitha

Assistant Professor, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi, Tamil Nadu

Abstract

Online social networks such as Twitter generate massive volumes of short, informal messages that reflect public opinion on products, events, and policies. However, the presence of slang terms, abbreviations, misspellings, hashtags, and emoticons makes accurate sentiment analysis on Twitter particularly challenging. This paper proposes a hybrid opinion mining framework that combines a lexicon-based approach with supervised machine learning to improve sentiment classification of tweets. The framework first applies a Twitter-specific preprocessing pipeline, including noise removal, hashtag normalization, slang expansion, and emoticon-to-sentiment mapping. A sentiment lexicon extended with slang and emoticon entries is then used to compute an initial polarity score for each tweet. These lexicon-derived scores, together with n-gram and TF–IDF features, are provided as inputs to a machine learning classifier to form the hybrid model. The framework is evaluated on a labeled Twitter dataset, and its performance is compared against pure lexicon-based and pure machine learning baselines using accuracy, precision, recall, and F1-score. Experimental results show that the proposed hybrid approach achieves higher F1-scores, particularly for tweets containing informal expressions and emoticons, demonstrating its effectiveness for opinion mining in noisy microblog environments.

Key Words: opinion mining, sentiment analysis, Twitter, hybrid approach, lexicon-based method, machine learning, slang, emoticons, social media analytics

1. Introduction

Twitter has become a major platform for users to express opinions about products, services, events, and public policies in real time. The short, rapid nature of tweets enables organizations and researchers to monitor public attitudes and emerging trends more quickly than with traditional feedback channels. As a result, opinion mining or sentiment analysis on Twitter data has attracted significant attention in areas such as marketing, politics, crisis management, and social monitoring [1]. However, tweets differ from well-formed text because they are limited in length and often contain slang, abbreviations, hashtags, user mentions, URLs, and emoticons, which makes automatic sentiment classification more difficult.

Traditional sentiment analysis methods can be broadly categorized into lexicon-based approaches and machine learning approaches. Lexicon-based methods rely on predefined dictionaries of sentiment-bearing words to assign polarity scores to text, making them relatively simple and interpretable. Machine learning approaches, on the other hand, learn sentiment patterns from labeled data using algorithms such as Naive Bayes, Support Vector Machines, or neural networks. While both approaches have shown strong performance across many domains, their effectiveness declines on noisy Twitter data when issues such as non-standard language, domain-specific slang, and frequent use of emoticons are not handled carefully.

Several existing studies have applied sentiment analysis to Twitter data, but many of them focus either solely on lexicon-based scoring or solely on machine learning classification, and often use generic resources that are not tailored to the unique characteristics of tweets. Pure lexicon-based methods may fail to capture context and new expressions, whereas pure machine learning models may not fully benefit from prior sentiment knowledge embedded in lexicons, especially when the labeled dataset is limited. Moreover, many works do not explicitly model the sentiment contribution of emoticons and slang, even though these elements carry strong emotional cues in microblog text.

To address these issues, this paper proposes a hybrid opinion mining framework for Twitter data that integrates lexicon-based sentiment scoring with supervised machine learning. The framework is designed to operate at sentence (tweet) level and to handle noisy, informal Twitter language [1]. A dedicated preprocessing pipeline is applied to remove noise, normalize text, expand common abbreviations, and convert emoticons into sentiment tokens. An extended sentiment lexicon, enriched with slang and emoticon entries, is used to compute an initial polarity score for each tweet. These lexicon-derived features are then combined with n-gram and TF–IDF features and fed into a machine learning classifier to improve overall sentiment classification performance. The effectiveness of the proposed hybrid approach is evaluated on a labeled Twitter dataset and compared with pure lexicon-based and pure machine learning baselines using standard evaluation metrics.

2. Review of Literature

Research on Twitter sentiment analysis has explored lexicon-based, machine learning, and hybrid approaches, each with distinct strengths and limitations. Early work typically treated these approaches separately, while more recent studies show that combining them can improve performance on noisy social media text.

Kolchyna et al.[2] presented a detailed comparison of lexicon-based and machine learning methods for Twitter sentiment classification. They showed that the accuracy of lexicon-based methods can be significantly improved by enhancing sentiment lexicons with emoticons, abbreviations, and social-media slang expressions. Furthermore, they proposed an ensemble method that uses a lexicon-based sentiment score as an input feature to a machine learning classifier, demonstrating that this combination yields more precise classifications on SemEval Twitter benchmarks. Their findings directly support the use of hybrid frameworks that integrate lexicon scores with learned features.

Maurya et al.[4] developed a hybrid sentiment analysis approach on Twitter data that combines textual features with additional modalities. Their framework uses preprocessing, feature extraction, and a hybrid classification mechanism to handle social media challenges such as informal language and high noise levels. The study reports that hybrid models can outperform pure machine learning or pure lexicon-based approaches, especially when training data is limited or heterogeneous. However, the work primarily focuses on combining different feature types and does not emphasize systematic handling of slang and emoticons as separate sentiment carriers.

Sayeedunnisa et al.[5] specifically investigated the impact of slang and emoticons in sentiment analysis of Twitter data. Using tweets related to the #MeToo movement, they showed that integrating emoticons and slang features with conventional bag-of-words significantly improves classification accuracy compared to using textual features alone. Their work uses a slang dictionary (SLANGSD) that assigns sentiment scores to slang terms and demonstrates that these additional features enhance performance on informal, subjective tweets. This study clearly supports the idea that slang and emoticons should be treated as first-class sentiment indicators in social media opinion mining.

A widely cited survey by Gayathri et al.[6] reviewed Twitter sentiment analysis methods and highlighted that most studies are based either on lexicon methods or supervised learning, with hybrid methods gaining increasing attention. The survey notes that Twitter data often contains incomplete expressions, acronyms, irregular grammar, and non-dictionary terms, which degrade the performance of both lexicon and machine learning models if not treated properly. This reinforces the need for specialized preprocessing and feature engineering for microblog data.

Kouloumpis et al.[7] and other works on Twitter sentiment analysis using machine learning emphasize the role of supervised classifiers such as Naive Bayes, SVM, and Random Forest, along with TF–IDF or n-gram features. These methods achieve good results when sufficient labeled data is available but depend heavily on the quality and size of the training set and often ignore lexicon-based prior sentiment information.

Systematic reviews of social opinion mining further show that Twitter is the most widely used platform in opinion mining research, yet issues like sarcasm, domain dependence, and informal language remain challenging.

Therefore, existing literature indicates that:

- Enhancing lexicons with emoticons, abbreviations, and slang improves lexicon-based performance on Twitter.
- Hybrid methods that feed lexicon-derived sentiment scores into machine learning classifiers provide better accuracy than using either approach alone.
- There is still a need for frameworks that explicitly combine Twitter-specific preprocessing, slang/emoticon handling, and hybrid lexicon–ML modelling in a unified, sentence-level opinion mining system. The present work aims to address this gap.

3. Proposed Methodology

This section describes the proposed hybrid opinion mining framework for Twitter data, including data collection, preprocessing, lexicon-based sentiment scoring, machine learning classification, and evaluation design [8].

3.1 Data Collection

Tweets are collected using the official Twitter API based on a predefined set of keywords and hashtags related to the chosen topic (for example, a product, event, or policy). The query specification includes:

- Relevant hashtags and phrases (e.g., “#brandname”, “brandname review”).
- Language filter (e.g., lang = "en").
- Time window (e.g., tweets posted over a period of 2–3 months).

Only original tweets are retained: retweets, quoted tweets, and replies containing no meaningful text are removed. Non-English tweets and tweets consisting solely of URLs, images, or emojis without text are discarded. A subset of the collected tweets is then manually annotated by human coders into three sentiment classes: positive, negative, and neutral, to create a labeled gold-standard dataset for training and evaluation[9].

3.2 Preprocessing of Twitter Data

Due to the noisy and informal nature of Twitter text, a specialized preprocessing pipeline is applied:

- Noise removal: Eliminate URLs, user mentions (@username), and unnecessary punctuation.
- Hashtag processing: Remove the hash symbol and split compound hashtags where possible to preserve sentiment words.
- Case normalization and character normalization: Convert all text to lowercase and reduce repeated characters (e.g., “goooood” → “good”).
- Slang and abbreviation handling: Use a curated slang dictionary to expand common social media abbreviations (e.g., “u” → “you”, “gr8” → “great”).
- Tokenization and stop-word removal: Tokenize tweets into words and remove standard stop-words that do not contribute to sentiment.
- Emoticon and emoji mapping: Map emoticons and basic emojis to sentiment tokens (e.g., “:)”, “·” → POS_EMOTICON; “:(”, “·” → NEG_EMOTICON), so they can be incorporated into sentiment scoring.

This preprocessing directly targets common Twitter challenges such as slang, abbreviations, misspellings, and heavy use of emoticons.

3.3 Lexicon-Based Sentiment Scoring

In the first analysis stage, a lexicon-based method estimates the polarity of each tweet. A base sentiment lexicon [9,10] (e.g., SentiWordNet or another polarity dictionary) is extended with:

- Entries for frequent Twitter slang and abbreviations, each with an associated sentiment score.
- Entries for emoticon tokens (POS_EMOTICON, NEG_EMOTICON, and optionally neutral).

For each preprocessed tweet, the procedure is:

1. Identify sentiment-bearing tokens (lexicon words, slang terms, emoticon tokens).
2. Retrieve their polarity scores and sum or average them to compute a raw sentiment score S.
3. Apply simple negation handling (e.g., invert scores after "not", "never" for a short window).
4. Assign a lexicon-based label: if $S > \theta_{pos}$, positive; if $S < \theta_{neg}$, negative; otherwise neutral.

This lexicon-based label and the numeric score S are used both as an independent baseline and as features in the hybrid model.

3.4 Machine Learning Classifier (Hybrid Framework)

In the second stage, a supervised machine learning classifier is trained using both standard textual features and lexicon-derived features [11]. Feature extraction includes:

- Bag-of-words and n-gram (unigram, bigram) features.
- TF–IDF weights for terms.
- Binary indicators or counts for emoticon categories and slang terms.
- The lexicon-based sentiment score S and its discrete polarity label as additional features.

A classifier such as Support Vector Machine (SVM), Logistic Regression, or Naive Bayes is used, following common practice in Twitter sentiment analysis. The classifier learns to combine contextual patterns captured in n-grams with prior sentiment information from the lexicon and emoticon/slang features, forming the proposed hybrid opinion mining framework [12,13].

3.5 Evaluation Design

The labeled dataset is divided into training and test sets or evaluated via k-fold cross-validation [14]. Performance is assessed using:

- Accuracy
- Precision
- Recall
- F1-score

Three configurations are compared:

1. Lexicon-based only: sentiment assigned purely from extended lexicon scores.[15,16]
2. Machine learning only: classifier trained on textual features without lexicon score or emoticon/slang features.[16,17,18]]
3. Proposed hybrid: classifier trained on textual features plus lexicon score and emoticon/slang features [19,20,21].

This comparison allows quantifying the contribution of lexicon integration and Twitter-specific preprocessing over standard baselines.

4. Conclusion and Future Work

This study presented a hybrid opinion mining framework for Twitter that combines lexicon-based sentiment scoring with supervised machine learning. The framework is specifically designed to handle the noisy and informal nature of tweets by incorporating Twitter-oriented preprocessing steps such as hashtag normalization, slang expansion, and emoticon-to-sentiment mapping. At the core of the approach, an extended sentiment lexicon enriched with slang and emoticon entries is used to compute an initial polarity score for each tweet, which is then integrated as an additional feature into a machine learning classifier alongside n-gram and TF–IDF features. Based on findings from existing literature, such hybrid designs have been shown to outperform purely lexicon-based or purely learning-based methods on social media sentiment tasks, especially in environments with limited labeled data and high linguistic variability.

The proposed framework contributes conceptually by emphasizing three elements often treated separately in prior work: Twitter-specific preprocessing, explicit modeling of slang and emoticons as sentiment carriers, and systematic integration of lexicon-derived scores into a supervised classification pipeline. This design can be adapted to different domains (e.g., product reviews, political discourse, event monitoring) by changing the keyword set and updating the domain-specific lexicon. It also provides a clear foundation for later empirical work, where the relative contributions of each component—preprocessing, lexicon extension, and hybrid feature design—can be quantitatively evaluated.

Future work will focus on implementing and validating the framework on one or more real Twitter datasets with manual annotations. Planned extensions include: (i) detailed experiments comparing multiple machine learning models (e.g., SVM, Logistic Regression, Random Forest) and, optionally, deep learning models such as LSTM or transformer-based architectures as alternatives to classical classifiers; (ii) automatic or semi-automatic construction of domain-specific sentiment lexicons using distributional semantics or word embeddings, reducing manual lexicon maintenance; and (iii) incorporation of more advanced features to handle sarcasm, negation scope, and code-mixed language. Further research could also explore multimodal Twitter opinion mining by integrating text with images or metadata, as well as developing visualization tools and dashboards to support real-time monitoring of public sentiment.

5. References

[1] Angelpreethi, A., & Ramesh Kumar, S. B. (2019). Dom_Classi: An enhanced weighting mechanism for domain specific words using frequency-based probability. International Journal of Applied Engineering Research, 14(1), 140-148, [Online]. Available: ijaerv14n1_21.pdf

[2] Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T., "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," arXiv preprint, arXiv:1507.00955, 2015.

[3] Angelpreethi, A., & Britto Ramesh Kumar, S. (2018). Opinion mining using hybrid approach on big data: A perspective. International Journal of Scientific Research in Computer Science and Management Studies, 7(5). https://doi.org/10.5281/zenodo.17399553

[4] Maurya, C. G., Jha, S. K., & others, "Sentiment Analysis: A Hybrid Approach on Twitter Data," *Procedia Computer Science*, vol. 235, pp. 429–436, 2024. DOI: 10.1016/j.procs.2024.04.094.

[5] Sayeedunnisa, F., Hegde, N. P., & Khan, K.-U.-R., "Using Slang and Emoticon for Sentiment Analysis of Social Media Data," *International Journal of Engineering Research & Technology (IJERT)*, NCAIT – 2020, Vol. 8, Issue 15, 2020. DOI: 10.17577/IJERTCONV8IS15024

[6] Gayathri, M., Shajun Nisha, S., & Mohamad Sathik, M., "Twitter Sentiment Analysis: Survey," *International Journal of Engineering Research & Technology (IJERT)*, ICATCT – 2020 (Volume 8, Issue 03), 2020. DOI: 10.17577/IJERTCONV8IS03029

[7] Kouloumpis, E., Wilson, T., & Moore, J. D., "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pp. 538–541, 2011.

[8] Angelpreethi, A. (2023). Fuzzy based sentiment classification using fuzzy linguistic hedges for decision making. Mapana Journal of Sciences, 22(Special Issue 2), 63–79. . [Online]. Available: https://doi.org/10.127v23/mjs.sp2.4

[9] Dao, P. Q., Roantree, M., Nguyen-Tat, T. B., & Ngo, V. M. (2024). Exploring multimodal sentiment analysis models: A comprehensive survey. Preprints. https://doi.org/10.20944/preprints202408.0127.v1

[10] Velmurugan, S., Prakash, M., Neelakandan, S., & Martinson, E. O. (2024). An efficient secure sharing of electronic health records using IoT-based hyperledger blockchain. International Journal of Intelligent Systems, 2024.

[11] Andrew, J., Isravel, D. P., Sagayam, K. M., Bhushan, B., Sei, Y., & Eunice, J. (2023). Blockchain for healthcare systems: Architecture, security challenges, trends and future directions. Journal of Network and Computer Applications, 215, 103633.

[12] Angelpreethi, A. (2025). OPINE_NEG: An approach to detecting negations and intensifiers using social media data. International Multidisciplinary Research Journal Reviews, 2(8), 13–17. [Online]. Available: IMRJR.2025.020803.pdf

[13] Parimala, S., Kumutha, R., Iram, F., Sunena Rose, M. V., N. R., & Angelpreethi, A. (2024, July). Improved moth flame optimization with deep convolutional neural network for colorectal cancer classification using biomedical images. In Proceedings of 2024 Second International Conference on Advances in Information Technology (ICAIT), IEEE. .[Online].Available: 10.1109/ICAIT61638.2024.10690631

[14] Xu, C., Qu, Y., Eklund, P. W., Xiang, Y., & Gao, L. (2021). Bafl: An efficient blockchain-based asynchronous federated learning framework. In IEEE Symposium on Computers and Communications, 1–6. https://doi.org/10.1109/ISCC53001.2021.9631405

[15] Malik, S., Dedeoglu, V., Kanhere, S. S., & Jurdak, R. (2019). TrustChain: Trust management in blockchain and IoT supported supply chains. In IEEE International Conference on Blockchain (Blockchain), 184–193.

[16] Aleisa, M. A. (2025). Blockchain-Enabled Zero Trust Architecture for Privacy-Preserving Cybersecurity in IoT Environments. IEEE Access.

[17] Angelpreethi, A., & Ramesh Kumar, S. B. (2019). NIC_LBA: Negations and intensifier classification of microblog data using lexicon based approach. Journal of Emerging Technologies and Innovative Research, 6(6),753 759.[Online].Available: https://www.jetir.org/papers/JETIR1906R05.pdf.

[18] Angelpreethi A. Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic. IJIEMR Transactions. 2023;12(2):380–387. Available from: https://ijiemr.org/public/uploads/paper/590451676720775.pdf. 10.48047/IJIEMR/V12/ISSUE02/60.

[19] Bae, J., & Lim, H. (2018). Random mining group selection to prevent 51% attacks on Bitcoin. In 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, 81–82. https://doi.org/10.1109/DSN-W.2018.00040

[20] Wang, K., & Kim, H. S. (2019). FastChain: Scaling blockchain system with informed neighbor selection. In International Conference on Blockchain (Blockchain), 376–383. https://doi.org/10.1109/Blockchain.2019.00058

[21] Niranjanamurthy, M., Nithya, B. N., & Jagannatha, S. (2019). Analysis of blockchain technology: Pros, cons and SWOT. Cluster Computing, 22(6), 14743–14757. https://doi.org/10.1007/s10586-018-2387-5