International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis and Opinion Mining

Kushal Gajera*, Santosh Sonone**
*(MCA, JSPM University, Pune
Email: kushalgajera2003@gmail.com)

----------**********----------

Abstract:

This study investigates the use of sentiment analysis and sentiment mining for extracting meaningful information from user-generated content (UGC) posted on social media, reviews of products or services via e-commerce sites such as Amazon, and posts on various digital forums. The goal of this research is to identify trends in public opinion, classify sentiments as positive/negative/neutral, and assess the effect of those sentiments on the decision-making processes made by companies and governments. The methodology for this research project involves using machine learning and natural language processing (NLP) to carry out sentiment analysis using datasets obtained from various social networking sites such as Twitter and review sites like Amazon. The research will employ several methods ranging from preprocessing text data, extracting features, and classification methods (e.g., Naive Bayes (NB), Support Vector Machine (SVM), deep learning), and multiple other machine-learning algorithms to accomplish its objectives. The anticipated results of this research include a prototype system for performing sentiment analysis, improved classification accuracy when performing sentiment classifications, and greater potential for Indian small and medium enterprises (SMEs) as well as Indian policymakers to adopt the research outputs.

Keywords — **Sentiment Analysis, Opinion Mining, Natural Language Processing (NLP), Machine Learning, Deep Learning, Social Media Analytics & their usefulness to businesses & their use in text classification, such as through the Naive Bayes Algorithm or Support Vector Machine Algorithm or through the mining of User Generated Content (UGC) or by analysing public sentiment using social media channels including Twitter Data Analytics & analysing customer reviews & ratings on e-commerce websites like Amazon, etc.**

----------**********----------

I. INTRODUCTION

THE RAPID GROWTH OF DIGITAL COMMUNICATION HAS OCCURRED THROUGH:

social media (Twitter, Facebook, Instagram, etc.) and e-commerce (e.g., Amazon, Flipkart) platforms that provide opportunities for individuals to express their opinions. Daily, millions of reviews, comments, and posts are generated, leading to a massive amount (unstructured) of text data.

Analyzing the vast volume of text data to identify public opinion and sentiments is impossible through manual intervention. Automated tools such as sentiment analysis allow analysts to quickly determine public opinion and emotion associated with different topics.

Automated sentiment analysis can identify if a text expresses positive, negative, or neutral opinions. Researchers conducting opinion mining can identify sentiments related to specific subject areas such as product features, services, and policies.

THE APPLICATIONS OF SENTIMENT ANALYSIS INCLUDE:

Businesses can leverage the information retrieved from automated sentiment analysis to make improvements to the way they deliver their products

or services or develop strategies used to market those products or services.

Governments can assess citizen opinions related to how effective they believe their policies are, assess the effectiveness of the government via elections, and assess social issues.

Researchers can study the trend of communication styles and how society behaves.

The problems resulting from the introduction of automated sentiment analysis include:

The sheer volume of data is produced on a daily basis through social media, blogs, and e-commerce sites. Manually analysing this type of data is impossible, creating a demand for automated sentiment analysis solutions.

Complexity of the English language is often another source of error in sentiment analysis tools. English is often used in many forms (ex. sarcasm, slang, emojis, and multiple languages), which can lead to incorrect identification of a statement's true meaning. For example, a statement like "Wow, great service!" may be identified (incorrectly) as being a positive statement due to its literal words. Research Gap

A. Limited work on multilingual sentiment analysis frameworks.

B. Few studies combine deep learning with regional language datasets.

C. Lack of robust solutions for real-world applications in Indian SMEs and governance.

II. RELATED WORK

Sentiment Analysis & Opinion Mining (also called affect analysis perception) has increasingly become a vital area of academic research over the past few years, largely due to the rise of social networking sites, the growth in online shopping, and discussion boards. Like most areas of study during the early days of Sentiment Analysis, the first attempts to perform sentiment analysis utilized lexicon-based methods to identify whether a word conveyed either a positive or negative sense by referencing pre-existing dictionaries of positive and negative words. While these methods were very simple and computationally fast, they could not adequately detect sarcasm, contextual meaning, and/or specialized word meanings.

As of today, various forms of Machine Learning (ML) techniques have emerged, including Naive Bayes Classifier (NBC), Support Vector Machines (SVM), and Decision Trees, as effective ways of conducting sentiment analysis. ML has been widely accepted by researchers who have demonstrated improvements in both the accuracy of results as well as the speed of achieving that level of accuracy by using machine learning. For example, research indicates that Support Vector Machine (SVM) outperformed other traditional ML methods when processing large amounts of high-dimensional text data. However, researchers also indicate that Naive Bayes Classifier (NBC) provides researchers with the fastest processing time and a much more straightforward implementation than other traditional ML methods. However, conducting classification using traditional ML methods requires a great deal of pre-processing and manipulation.

Recently, researchers have begun using deep learning and/or natural language processing (NLP) techniques by utilizing Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM) networks, etc., through the use of transformer-based architectures such as BERT and RoBERTa, to conduct sentiment analysis. Unlike traditional ML methods, deep learning (CNN and LSTM) allows researchers to automatically learn the context of each word as well as its relationship to surrounding words, which has led to improvements in the accuracy of sentiment prediction. Multilingual sentiment analysis is also an important focus of current research due to the globalization of business and the growing prevalence of code-switched text (e.g., Hinglish) in both India and throughout the world; however, research regarding the aforementioned topics remains limited in scope at this time.

III. SYSTEM MODEL

The Sentiment Analysis and Opinion Mining model aims to give an overview of user-generated

content on social media and e-commerce websites. The model has four key phases; data collection, preprocessing, feature extraction, and Sentiment Classification.

Phase one - Data Collection

In the data collection phase of the research model, we build a dataset from multiple sources such as Twitter, Amazon Reviews, Flipkart Reviews, and multilingual datasets from Kaggle. These sources produce unstructured/raw, noisy (text, emojis, slang), punctuation, and multi-language written-up.

Phase Two - Preprocessing

The next phase of the model involves cleaning the unstructured/raw data so that analysis can take place effectively. This cleaning includes the removal of all stop words, special characters/symbols, URL's, and punctuation symbols. The next step is to use Tokenisation, Stemming, and Lematisation to normalise the text to prepare the data for feature extraction.

Phase Three - Feature Extraction

Once the data has been cleaned, there are multiple feature extraction methods to create numerical representations of text for use by machine learning models. The techniques include TF-IDF, Word2Vec, and BERT embeddings.

Phase Four - Sentiment Classification

The identified features are passed to various Sentiment Classification Algorithms, including Naive Bayes, Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and BERT. All these methods will classify the data as either positive, negative, or neutral sentiments.

The performance of the method will be evaluated using multiple performance metrics such as Accuracy, Precision (True positives and False Positives), Recall (True Positives and False Negatives), F1 Score, and a Confusion Matrix. Finally, the analysis results will be graphically represented on dashboards to be useful to different stakeholders - businesses, researchers, and government.

IV. LITERATURE REVIEW

Summary of Reports:

Research involving evaluating opinions has greatly increased during the last decade mainly because of the growth of social networking and online review sites.

In the beginning, lexicon-based approaches were primarily used. These techniques involved utilizing pre-determined lists of words that were either positive or negative to determine the sentiment of some form of text.

Subsequent work using machine learning techniques (Naive Bayes classifier, Support Vector Machines, Decision Trees, etc.) increased the accuracy of sentiment classification through training on labeled data sets.

The most recent developments in sentiment classification have involved the use of deep learning methodologies (CNNs, RNNs, LSTMs) as well as transformer models (BERT, RoBERTa, etc.), which have set the state-of-the-art for accuracy in sentiment classification.

Traditional Methodologies:

The use of machine learning algorithms, such as Support Vector Machines and Naive Bayes, can provide better performance than lexicon-based approaches; however, they require an enormous amount of feature engineering.

Deep Learning Methodologies:

Various forms of neural networks allow for a better understanding of the context, thus minimizing the need for manual feature extraction.

Studies in Non-English and Domain-Specific Languages:

Much of the previous research has been conducted mainly on Benchmarks (i.e., datasets) in English rather than in other languages, such as Hindi or Marathi, among others.

Additionally, domain-specific sentiment classification (e.g., product review vs. political debate) is still a challenge.

Challenges in Sentiment Analysis:
Detecting sarcasm and irony: Many sentiment analysis models misinterpret when sarcasm or irony is present, as the literal words used may have a positive connotation but their actual meaning/intent is negative.

Mixed Languages: In India, users often communicate in a Hybrid Language (Hinglish) or using Marathi mixed with English. This can produce difficulties during the processes of tokenising words and classifying sentiment.

Domain-specific vocabulary: The same words can take on different meanings in different contexts. For example, Killer can be considered positive in reference to gaming but can be considered Negative when used in regard to crime.

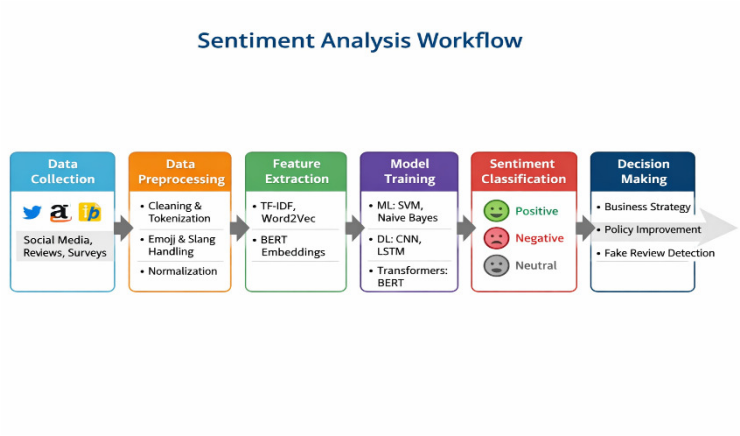
Data Collection Imbalance: Data collected may be biased towards positive reviews; therefore, the model will reflect bias in terms of the majority positive opinions and will be unable to accurately identify minority negative opinions.

Ethics: Issues of bias in the data set used to train the model can exacerbate stereotypes. As well, there may be ethical considerations related to obtaining information from social media accounts as a result of this analysis.

Research Objectives & Questions:

Purpose: To create a fast and effective way to do Sentiment Analysis on multilingual data.

Diagram:



Technologies & Tools:

Transformers: BERT, RoBERTa, DistilBERT provide deep contextual understanding.
Fine-tuning domain-specific data results in better accuracy.

Pre-trained multilingual models: mBERT or XLM-R use both Hindi and English together without difficulty.

These models can be used effectively on code-mixed datasets present in Indian social media.

Cloud platforms: AWS Comprehend offers scalable enterprise-level sentiment analysis.

Google Cloud NLP can be used in conjunction with large data pipelines.

Azure Cognitive Services provides multilingual capabilities via simple API interfaces.

Visualization applications:
Tableau and Power BI assist business leaders in constructing decision-support dashboards.

Matplotlib and Seaborn are suitable tools for visualizing academic data trends over time.

Hybrid solutions: By combining lexicon-based and deep learning, hybrid approaches provide additional robustness.

Technologies:
Python will be the programming language used for text pre-processing and building models.

The libraries utilized will include NLTK, Scikit-Learn, TensorFlow and Keras, in addition to Hugging Face Transformers.

The development environment will be Jupiter Notebook.
Supporting resources will include Kaggle datasets, Google Collab for training the models, and GitHub for version control purposes

Data Collection:
Publicly available datasets used will include:

Twitter sentiment datasets.
Amazon or Flipkart product review datasets.
Kaggle datasets for multilingual sentiment analysis.

The collected data will consist of English and Hindi text for testing the performance of the multilingual capabilities of the framework.

Data Pre-processing:
Pre-cleaning of the collected text data will involve removing all punctuation, stop words and special characters.

After cleaning, the collected data will be tokenized by splitting sentences to get words out.

Next, the collected text will be normalized (lower case letters, stemming, and lemmatization).

Additional cleaning will occur with emoji's, slang and mixed language text.

Finally, numerical features will be created from the collected text by applying one of the methods (i.e., TF-IDF, Word2Vec, or BERT embeddings) to the collected text documents.

Analysis Methods:
Machine Learning Analysis Methods: SVM, Naïve Bayes.
Deep Learning Analysis Methods: CNN, LSTM.
Transformer Model Analysis Method: BERT.

Evaluation Metrics: F1, Recall, Precision, Accuracy, Confusion Matrix.

Dataset Privacy: All datasets will be publicly available.

Anonymization: User anonymity will be accounted for.

No Personal Data/Private Information: No sensitive or private information will be collected.

V. CONCLUSIONS

An effective methodology for sentiment analysis and opinion mining can be developed using tools such as Natural Language Processing (NLP), Machine Learning, and Deep Learning. This exploration of potential study areas concentrates on extracting information from user-generated sources (e.g., Social Media, E-commerce, Online discussion forums) while assessing multiple technique(s) for classifying sentiment (e.g., Naive Bayes, Support Vector Machine (SVM), CNN, LSTM, BERT) so that an overall improvement in accuracy and efficiency with regard to predicting sentiment may be achieved.

ACKNOWLEDGMENT

This research work has been possible due to the support provided by the Department of Computer Science at JSPM University, Pune, as well as the research guidance and academic resources of the faculty members, mentors, and other researchers who assisted with this work throughout its course.

I also would like to acknowledge the contributions made by publicly available datasets, research papers, and open-source software tools (such as Python, Scikit-Learn, TensorFlow, Keras, Hugging Face Transformers,Kaggle) to this research development and analysis.

Recognizing that family and friends have encouraged me throughout the process, with their

support, patience, and motivation, I express my deepest thanks to them for their help in completing this research Paper.

REFERENCES

1. Medhat, W., Hassan, A., & Korashy, H. (2023). *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, 14(2), 101–115.
2. Zhang, L., Wang, S., & Liu, B. (2022). *Deep learning for sentiment analysis: A survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1455.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT, 4171–4186.
4. Kumar, A., & Jaiswal, A. (2024). *Multilingual sentiment analysis using deep learning: A case study on Indian languages*. International Journal of Computer Applications, 183(5), 25–32.
5. Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.
6. Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2020). *Sentiment analysis: The evolution of deep learning approaches*. IEEE Intelligent Systems, 35(5), 62–73.
7. Rana, T. A., & Cheah, Y. N. (2021). *Aspect-based sentiment analysis: Recent developments and future directions*. Artificial Intelligence Review, 54(3), 2039–2097.
8. Zhao, Z., Liu, W., & Wang, K. (2023). *Research on sentiment analysis method of opinion mining based on multi-model fusion transfer learning*. Journal of Big Data, 10(155).
9. Priya, M. L., Parimala, E. H., Janaki, M., & Likhitha, G. (2024). *Emerging approaches in sentiment analysis and opinion mining: A critical review*. Bangalore City College Journal of Research, 12(1), 45–58.
10. Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
11. Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of ACL, 417–424.
12. Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Proceedings of ICWSM, 216–225.
13. Socher, R., Perelygin, A., Wu, J., et al. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. Proceedings of EMNLP, 1631–1642.
14. Kim, Y. (2014). *Convolutional neural networks for sentence classification*. Proceedings of EMNLP, 1746–1751.
15. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.
16. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.
17. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Proceedings of EMNLP, 1532–1543.
18. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
19. Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.
20. Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Pearson Education.
21. Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.
22. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Stanford University Technical Report.
23. Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining*. Proceedings of LREC, 1320–1326.
24. Tang, D., Qin, B., & Liu, T. (2015). *Document modeling with gated recurrent neural network for sentiment classification*. Proceedings of EMNLP, 1422–1432.
25. Sun, C., Huang, L., & Qiu, X. (2019). *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence*. Proceedings of NAACL, 380–385.
26. Alharbi, A. S., de Doncker, E., & Abraham, A. (2018). *Machine learning approaches for sentiment analysis: A survey*. Journal of Computer Science, 14(9), 1231–1243.
27. Sharma, A., & Dey, S. (2021). *A comparative study of machine learning algorithms for*

sentiment analysis. International Journal of Advanced Computer Science and Applications, 12(4), 215–221.

28. Singh, V., & Gupta, R. (2022). *Sentiment analysis on multilingual Indian datasets using hybrid deep learning models*. Procedia Computer Science, 218, 543–550.
29. Balahur, A., & Turchi, M. (2014). *Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis*. Computer Speech & Language, 28(1), 56–75.
30. Khan, F. H., Bashir, S., & Qamar, U. (2014). *TOM: Twitter opinion mining framework using hybrid classification scheme*. Decision Support Systems, 57, 245–257.

