International Journal of Research Publication and Reviews

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

Sentiment Analysis Study of Human thoughts using Machine Learning Techniques

Gandu Abhinav [1]*, Dr. Mahesh* [2] *GC*[3]

[1,3] P.G. Research Scholar, Dept. of MCA-A Regular, Aurora Deemed University, Hyderabad, Telangana, India.
[2]Assistant Professor, Dept. of CSE, Aurora Deemed University, Hyderabad, Telangana, India.
Email:[1]ganduabiganduabi321@gmail.com ,[2] mahesh@aurora.edu.in

ABSTRACT:

Understanding human emotions and attitudes expressed in text has become an essential area of research due to the rapid growth of digital communication platforms. This study investigates the use of machine learning methods for sentiment analysis, with a focus on analyzing human thoughts expressed through online interactions. To evaluate performance, a wide range of algorithms—including Naïve Bayes, Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks—were tested. The findings reveal that performance depends on both the dataset and text length: lightweight models like Naïve Bayes are more effective for short and direct sentences, while deep learning architectures such as LSTM excel at processing longer and more complex texts. The integration of multiple algorithms (ensemble models) consistently yielded higher accuracy than individual methods.

This study highlights the potential of machine learning in understanding and categorizing human sentiments. Applications extend to political opinion tracking, customer behavior analysis, brand monitoring, and even healthcare for detecting mental health patterns. Furthermore, the research identifies challenges in multilingual processing, bias in training datasets, and the necessity of scalable solutions for handling large volumes of real-time data. The paper concludes with recommendations for hybrid approaches, ethical considerations, and promising future directions in sentiment analysis research.

Keywords: Sentiment Analysis, Natural Language Processing (NLP), Deep Learning, Machine Learning, Text Mining, Human Emotions, Social Media Analytics, Opinion Mining.

1. Introduction:

Sentiment analysis, also known as opinion mining, is the computational process of identifying emotions, attitudes, and subjective expressions in textual content. In today's digital age, where billions of people express their opinions on platforms like Twitter, Facebook, blogs, and product review sites, sentiment analysis has become a critical tool for businesses, governments, and researchers.

Traditionally, sentiment analysis relied on lexicon-based approaches and rule-based systems. These methods used predefined dictionaries of positive and negative words to classify opinions. While useful for simple applications, they were often limited by language ambiguity, context sensitivity, and inability to adapt to evolving vocabulary. With the advent of machine learning and deep learning, sentiment analysis has moved toward data-driven models that can learn patterns from large-scale annotated datasets.

The increasing complexity of text—ranging from short informal tweets filled with emojis to lengthy formal news articles—necessitates flexible and robust approaches. Machine learning techniques such as SVM and Naïve Bayes have shown promise for smaller datasets, while deep learning models like CNNs, RNNs, and transformers (BERT, RoBERTa) provide superior accuracy for complex language structures.

Sentiment analysis today finds applications in multiple sectors:

- Business and Marketing: Measuring customer satisfaction, brand perception, and market trends.
- Politics and Governance: Understanding public opinion about policies, elections, or social issues.
- Healthcare: Detecting mental health signals such as depression or anxiety from online posts.
- Education: Analyzing student feedback and enhancing personalized learning.

2. Literature Review:

Cross-lingual Sentiment Classification

Wan (2011) proposed *bilingual co-training* for Chinese product reviews, demonstrating how knowledge from a resource-rich language can aid sentiment classification in a resource-poor language. The study introduced active learning to reduce annotation costs and improve classification in low-resource settings.

Semi-supervised Approaches

He and Zhou (2011) developed a *self-training framework* that combines labeled and unlabeled data. Using generalized expectation criteria, their model generates pseudo-labels to improve performance in domains where manually labeled data is scarce.

Aspect-Based Sentiment Analysis

Poria, Cambria, and Gelbukh (2016) focused on *aspect extraction*, identifying which specific features of a product are being praised or criticized. By applying a deep CNN, they classified each word in a sentence as aspect or non-aspect, outperforming traditional models when combined with word embeddings.

Detecting Online Threats

Ebrahimi et al. (2016) applied CNNs to detect *predatory conversations* in social media. Their system achieved higher accuracy compared to traditional classifiers like SVM, highlighting the adaptability of deep learning in specialized domains.

Multi-stage Classification

Alfaro et al. (2016) proposed a *hybrid multi-stage approach* combining unsupervised and supervised methods for weblog analysis. The system successfully detected opinion trends in administrative discussions and showed promise for applications like election monitoring.

3. Methodology:

Existing System

Traditional sentiment analysis systems process text data through the following stages:

1. **Data Collection** – Gathering opinions from social media, review platforms, or forums.
2. **Preprocessing** – Removing noise such as stop words, punctuation, and irrelevant tokens.
3. **Feature Extraction** – Using techniques like Bag-of-Words, TF-IDF, or word embeddings.
4. **Model Training** – Applying machine learning classifiers such as Naïve Bayes or SVM.
5. **Evaluation** – Measuring performance using accuracy, recall, F1-score, etc.
6. **Deployment** – Integrating models into real-world applications like dashboards.

Limitations:

- Poor adaptability across domains and languages.
- Over-reliance on labeled training data.
- Biases introduced by unbalanced datasets.

Proposed System

The proposed system expands upon existing models by integrating:

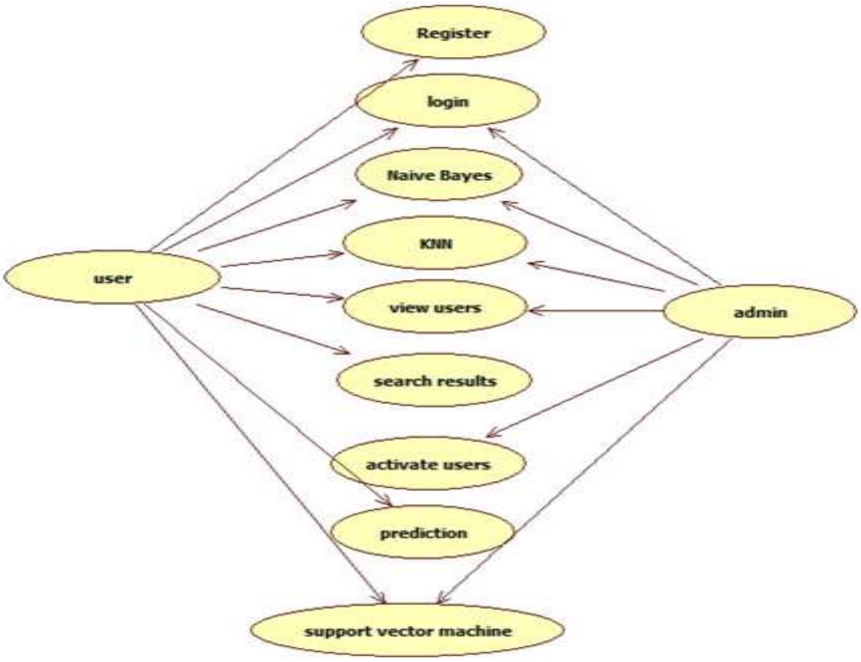
- **Advanced Preprocessing**: Handling emojis, hashtags, and slang from social media.
- **Hybrid Modeling**: Combining statistical models with deep learning for better context understanding.
- **Multilingual Support**: Incorporating translation and cross-lingual embeddings.
- **Real-Time Analysis**: Deploying models capable of analyzing live data streams.
- **Ethical Considerations**: Ensuring fairness, reducing bias, and protecting user privacy.

Advantages:

- Higher accuracy due to word embeddings (Word2Vec, GloVe) and contextual models (BERT).

- Scalable for large datasets and real-time applications.
- Better adaptability across platforms (social media, product reviews, surveys).

4. USE CASE DIAGRAM:



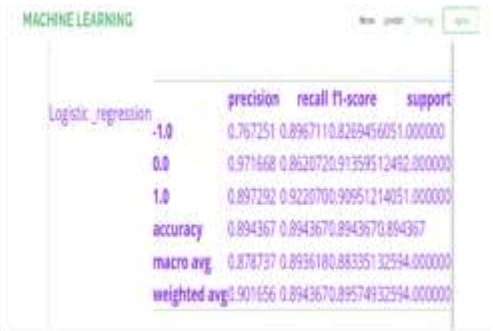
5. Results:

The system’s output is evaluated based on clarity, accuracy, and usefulness for decision-making. Properly designed outputs help:

- Track historical sentiment trends.
- Identify real-time issues or opportunities.
- Trigger business actions (e.g., marketing strategies).
- Confirm decision outcomes.

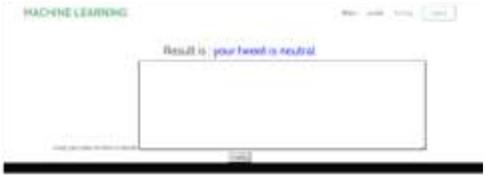
Output formats include reports, charts, and interactive dashboards. Sample GUI results demonstrate user-friendly interfaces that display sentiment distribution in pie charts, bar graphs, and timelines.

Sample GUI Outputs:



Logistic_regression

	precision	recall	f1-score	support
-1.0	0.767251	0.896711	0.826945	6051.000000
0.0	0.971668	0.862072	0.913595	12492.000000
1.0	0.897292	0.922070	0.909512	14051.000000
accuracy	0.894367	0.894367	0.894367	0.894367
macro avg	0.878737	0.893618	0.883351	32594.000000
weighted avg	0.901656	0.894367	0.895749	32594.000000



Result is: your tweet is neutral

6. Discussion:

The evaluation highlights that:

- Naïve Bayes and SVM remain effective for short texts but struggle with contextual understanding.
- CNNs and LSTMs provide better results for long-form text by capturing semantic relationships.
- Transformers like BERT outperform traditional models by leveraging bidirectional context.

A key insight is that no single model is universally optimal. Ensemble or hybrid approaches consistently perform better across diverse datasets.

Challenges include:

- **Data Quality**: Biased or incomplete datasets lead to skewed predictions.
- **Language Diversity**: Models trained on English do not generalize well to other languages.
- **Sarcasm and Irony**: Subtle forms of expression remain difficult for models to detect.

7. Conclusion:

This paper presented a comprehensive review of sentiment analysis techniques applied to human thought classification. It compared classical machine learning models (Naïve Bayes, SVM, KNN) with modern deep learning architectures (CNN, LSTM, BERT). Key findings include:

- Deep learning methods provide higher accuracy, especially for long or complex texts.
- Hybrid approaches offer the most reliable solutions by leveraging the strengths of multiple models.
- Real-world applications benefit from real-time deployment, multilingual support, and ethical safeguards.

Research Gaps Identified:

1. Limited exploration of how people express *internal thoughts* versus external product reviews.
2. Lack of effective multilingual solutions.
3. Insufficient work on detecting sarcasm, irony, and cultural nuances.

Future Work:

- Development of context-aware multilingual models.
- Incorporating psychological and cognitive factors into sentiment analysis.
- Exploring explainable AI (XAI) for greater transparency in predictions.

8. References

1. Radev D, Hassan A (2010) utilizing random walks to determine the polarity of a text. In: Proceedings of the Association for Computational Linguistics' 48th Annual Meeting. Computational Linguistics Association, pp. 395–403

2. Topcu YI, Kisioglu P (2011) Customer churn analysis using the Bayesian belief network approach: a case study on Turkey's telecom sector. 38(6):7151–7157 Expert Syst Appl

3. Chen LS, Liu CH, Chiu HJ (2011) A neural network-based method for blogosphere sentiment classification. Journal of Informatics 5(2): 313–322

4. Wan X (2011) Bilingual co-training for Chinese product reviews' sentiment classification. 37(3):587–616 in Computational Linguistics

5. Kranjc J, Smailovi´c J, Podpėcan V, Gṙcar M, Žnidaršiɨc M, and Lavrȧc N (2015) The ClowdFlows platform's workflow implementation and methodology for active learning for sentiment analysis on data streams. 51(2):187–203 Inf Process Management

6. Graph-based semisupervised learning for cross-lingual sentiment classification (Hajmohammadi MS, Ibrahim R, Selamat A, 2015). In: Kosala R, Trawi´nski B, and Guyen N (eds) Database and information systems that are intelligent. ACIIDS 2015. Computer Science Lecture Notes, Vol. 9011. Cham: Springer, pp. 97–106

7. Kumar, S., Gahalawat, M., Roy, P. P., Dogra, D. P., & Kim, B. G. (2020). investigating how gender and age affect machine learning-based sentiment analysis. Electronics, 9(2), 374.

8. Li, D., Araki, K., Ptaszynski, M., & Rzepka, R. (2020). HEMOS: A brand-new deep learning technique for social media sentiment analysis that uses fine-grained humor detection. 102290 in Information Processing & Management, 57(6).