STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 16, Issue 2, 2025
May–August
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Review JoCTA

A Review on Punjabi Language Sentiment Analysis Using Machine Learning

Surbhi Sekhri[1,*], Khushboo Bansal[2]

Abstract

In this computer era, everyone is expressing their opinions or sentiments not only physically but over the social media platforms very frequently. Earlier, English was the only language for exchanging thoughts over the web, however, in the recent times, feedbacks, surveys or comments by the users are expressed in regional languages too. This leads to the need of analyzing sentiments for better communication on the digital platforms. For example, if someone wants to book a hotel for a day, he may check for feedbacks and ratings for making his final decision. Sentiment analysis is the study that clarifies and categorizes views, feelings, and judgements in written information. Over the last decade, research into semantic analysis in the English language has expanded rapidly. Additionally, the number of online users who create content has been rising, engaging not only in English but also in their native languages. Most of the paperwork completed there to date has been in English, but there has been limited research in the field of Indian regional languages, particularly Punjabi. In this study, we explored different methods employed to carry out opinion research and investigative work for vernacular languages like Hindi, Bengali, Punjabi and many more. Numerous deep learning and lexicon-based strategies have been suggested in the literature to automate the sentiment analysis process.

Keywords: Lexicon-based, deep learning, transformers, NLP, polarity, negation handling

INTRODUCTION

"Sentiment analysis refers to the use of characteristic natural language processing, computational linguistics and content analytics to distinguish and extract subjective data from source materials". Sentiment analysis, also known as emotions mining, is a multidimensional problem in artificial intelligence [1]. Its goal is to minimize the gap between human minds and computers as much as possible. Views are typically conveyed through a positive, negative, or neutral stance of the author. Sentiment analysis focuses on classifying content according to its subjective and intended nature [2]. The task is typically described as categorizing a given content (typically a sentence) into either an objective or subjective category. The text may contain some objective facts or subjective opinions. It is essential to distinguish between them. SA aids in selecting the textual components and topics from which the slant and sentiment are drawn [3].

***Author for Correspondence**
Surbhi Sekhri
E-mail: surbhisekhri_19@yahoo.com

[1]Research Scholar, Department of Computer Science and Engineering, Desh Bhagat University, Mandi Gobindgarh, Punjab, India
[2]Assistant Professor, Department of Computer Science and Engineering, Desh Bhagat University, Mandi Gobindgarh, Punjab, India

Received Date: May 28, 2025
Accepted Date: June 23, 2025
Published Date: July 12, 2025

Citation: Surbhi Sekhri, Khushboo Bansal. A Review on Punjabi Language Sentiment Analysis Using Machine Learning. Journal of Computer Technology & Applications. 2025; 16(2): 116–122p.

Subjectivity

Sentiment can be powerfully classified into three distinct categories: objective, which represents undeniable truths; positive, reflecting a state of happiness, bliss, fulfillment, or satisfaction; and negative, capturing feelings of anguish, misgiving, bitterness, or discontent. Understanding these classifications can enhance our comprehension of emotional experiences and their impact on our lives. The sentiment can be scored according to its degree

of polarity. There are two study strategies in this regard: first, categorize clearly defined sentiment of the text to express the feeling as positive, negative, or neutral.

(Zara brand clothes are great) ਜ਼ਾਰਾ ਬ੍ਰਾਂਡ ਦੇ ਕੱਪੜੇ ਬਹੁਤ ਵਧੀਆ ਹਨ।

This sentence has a sentiment and talks about clothes with positive opinion (**ਵਧੀਆ**) of Zara clothing brand and hence these types of sentences are subjective in nature.

Objectivity

It demonstrates the content that lacks feeling [4]. It is basically verifiable facts and evidences, number based data which is reliable and consistent.

(The clothes are of Zara brand) ਕੱਪੜੇ ਜ਼ਾਰਾ ਬ੍ਰਾਂਡ ਦੇ ਹਨ।

This sentence gives a general information about the Zara brand rather than about a sentiment and hence it is objective in nature.

CLASSIFICATION OF SA

Sentiment analysis (SA) is structured around three pivotal phases of classification: document-level, sentence-level, and aspect-level. Every level plays a key role in extracting meaningful insights from text, ultimately enhancing our understanding of sentiments and opinions in various contexts [5].

1. *Document-level:* The entire document is assessed, prioritizing a comprehensive understanding of the sentiments over the examination of individual components. It is used to capture important aspects of the document likewise detecting the entire book's reviews. However, it is the least-used sentiment analysis.
2. *Sentence-level*: SA wants to group the conclusions stated in each sentence. The first stage is to identify the nature of a sentence that is subjective or objective. If the language is subjective, SA will determine whether it conveys favorable or unfavorable judgements. The ability to identify whether sentences remain on-topic, a complex co-reference challenge, poses significant hurdles to effective sentiment categorization at the sentence level. Addressing this issue is important for enhancing the correctness of sentiment analysis [6–9].
3. *Aspect-level*: Applications that order sentiment writing at the sentence or document level usually fall short because these levels do not comprehensively analyze the sentiment of text. To conduct a thorough study, we must identify the points of view and determine whether they are favorable or unfavorable from each perspective. Sentiment analysis is based on aspects that include both entities and aspects [10–13]. It is the most widely used sentiment analysis type.

Sentiment analysis works on statements that are of basically of two types:

1. *Regular statement*: ਇਹ ਐਪਲ ਫੋਨ ਬਹੁਤ ਵਧੀਆ ਹੈ।
2. *Comparative statement*: ਇਹ ਐਪਲ ਫੋਨ ਸੈਮਸੰਗ ਫੋਨ ਨਾਲੋਂ ਵਧੀਆ ਹੈ।

Sentiment Analysis (SA) refers to a collection of functions, tools, and techniques used to recognize and interpret emotions, views, personalities, and attitudes expressed in text. Understanding customer emotions is essential for businesses since customers tend to provide honest feedback in online settings. By automatically analyzing customer reviews and survey responses, companies can gain insights into their customers' preferences and make adjustments to their services to better meet their needs. In the past, sentiment analysis was primarily utilized to identify opinion polarity, classifying sentiments as positive, neutral, or negative. However, recent research has expanded the applications of this technique, integrating it with artificial intelligence and machine learning [14–17]. SA can be applied across various domains, including customer feedback, political commentary, and movie or product reviews. This technique is also capable of processing multiple languages, including Punjabi, which will aid in extracting sentiments from reviews in that language.

TYPES OF APPROACHES USED

Rule-based (Lexicon-based) Approaches

This method is both straightforward and highly effective, leveraging carefully curated lists of words (lexicons) that are highly associated with positive or negative sentiments. We can gain profound insights into emotional responses and enhance our understanding of sentiment analysis. Algorithms analyze text by matching words to these lexicons and calculating a sentiment score. Different rules can be added to handle negations, intensifiers, and other linguistic nuances.

Rules are typically expressed in if-then statement where if means condition and then means conclusion. Example, to move a robot, the condition that can be given is:

If (no-obstacle) Then Move forward;

Else Turn-right.

Machine Learning Approaches

These methods involve training machine learning models on labelled datasets (text with known sentiment). Algorithms learn patterns in the data and use these patterns to predict the sentiment of new text. These algorithms achieve high accuracy but a bit expensive. Additionally, the black box nature of the model significantly hinders our ability to comprehend its internal behavior, making it challenging to trust and interpret its decisions effectively.

Common Algorithms

- *Naive bayes:* It is a straightforward probability classification algorithm derived from Bayes' Theorem. Its working assumes that the features are conditionally independent, which makes it fast and efficient, particularly for large datasets. The term “naive” means here to the idea that the all features are independent of each other in a specific class.
- *Support vector machines (SVM):* This powerful method is ideal for classification as well as regression tasks. It effectively identifies the optimal line that not only separates the data into distinct groups but also maximizes the distance between the nearest points, referred to as support vectors, of each group. By leveraging this approach, we can achieve superior accuracy and robustness in our predictive models.
- *RNN*: This method has a significant role while dealing with the sequential nature of language. Semantic analysis aims to know the meaning of given data. RNNs excel at capturing the relationships between words in a sequence, allowing them to understand how preceding words influence the meaning of subsequent words. This is vital for tasks like Understanding how words like “not” or “very” modify the sentiment of other words. RNNs often work in conjunction with word embeddings (e.g., Word2Vec, GloVe). Recurrent Networks can process these word embeddings sequentially to understand the meaning of a sentence.
- *Transformers (like BERT):* The groundbreaking innovation lies in the attention mechanism, a powerful feature that empowers the model to discern the importance of many words in a sentence. This capability enhances its understanding, allowing for a deeper understanding of meaning and context. This approach is far more effective than previous recurrent neural network (RNN) approaches. Transformers can process entire sequences of words simultaneously, which makes them much faster and more efficient.

Hybrid Approaches

These innovative methods seamlessly blend the strengths of rule-based systems with the power of machine learning. For example, by using a rule-based system for text pre-processing, we can effectively enhance the input quality, allowing the machine learning model to excel in accurately classifying sentiment in the final stages. This integrated method improves efficiency as well as guarantees a more reliable and nuanced analysis of textual data.

LITERATURE REVIEW

According to Dadvar *et al*., a person's personality can be characterized as a set of requirements that demand a propensity on their behavior; this tendency is constant over time and situations [6]. It is important since it refers to a person's priorities in a properly established manner. The method that aids in characterizing people is sentiment analysis. In this study, we used personality keywords to extract the personality traits from user-submitted reviews written in Hindi about any individual.

According to Yi *et al.*, sentiment analysis is a powerful technique that uncovers and interprets subjective information from diverse sources [7]. By leveraging computational linguistics, text analysis, and natural language processing, it reveals valuable insights that can drive informed decisions and improve public sentiment's understanding. Better products have been created, user opinions have been understood, and business decisions have been implemented and managed, thanks to sentiment analysis. It is a method that aids in identifying consumer reviews of any product. Positive or negative feedback might be found in the reviews of the people. Sentiment analysis is a field that is constantly developing with several applications. Its primary goal is to enable computers to recognize and produce human-like emotions. The increase in user-generated content (UGC) in Hindi on websites, in business, academia, and other contexts has made it possible to effectively study and mine the data. They put out an algorithm that takes the review out of a review-sentence and also determines the subject of the opinion. Additionally, it utilizes a database dictionary to determine the review's sentiment score.

The consequences of negative emotion detection in SA in movie reviews were investigated by Singh *et al.* [8]. When negative terms, for instance, do not appear in the sentences, its polarity is investigated. To determine how it might affect the polarity identification of the sentences, several negation scopes that influence classification accuracy are explored. The findings demonstrate that using different window sizes does not significantly alter categorization accuracy.

The Sentiment Analyzer (SA) is a tool designed to extract sentiment about a specific topic from various online text sources, as outlined by Kaur and Gupta [9]. It employs NLP techniques to identify sentiment related to subject in each reference, rather than categorizing the sentiment of the respective document concerning that topic. The SA relies on two key language resources for its analysis: a sentiment lexicon and a sentiment pattern database. Its algorithms were tested using online product reviews, as well as a range of broader documents such as news articles and general websites [18–20].

Today, everyone on social media is using more than one language to convey their message for their ease, likewise English-Punjabi mixed code language, according to Soumya and Pramod [10]. So, there is a need to analyze this data appropriately to make the online communication better, they used trigram and five-gram approach to predict the performance of the system. Trigram shows 83% of accuracy to predict the mix-code data while five gram has 82% accuracy [11].

Analysis of Literature Review

The literature review for the comparative analysis of previous studies is presented in Table 1.

MODEL EVALUATION PARAMETERS

In the world of AI, there are metrics to test and evaluate a developed AL model to assess if it is accurate and efficient enough. A confusion matrix is a powerful device that utilizes a visual chart to effectively summarize the performance of a classification-based AI model [12]. By clearly juxtaposing the predicted values generated by the model with the actual, correct outcomes, it enables a complete assessment of the model's accuracy and effectiveness. This insightful analysis is essential for understanding how well your AI system performs and where improvements can be made (Figure 1).

Accuracy

Accuracy assesses the proportion of correct predictions relative to the total number of instances, providing a clear measure of performance. Conversely, error, which represents the complement of accuracy,

Table 1. Comparative study of previous study.

Researcher name	Data set used	Techniques used	Accuracy	Feature extracted
Mukherjee and Bhattacharyya [2]	Hindi WordNet	Unigram method with simple scoring method	47–48%	Positive and negative polarity
Sharma [14]	Movie reviews from web pages	Unigram and bigram method	75%	Positive and negative polarity
Jain and Sandu [15]	Punjabi news articles	Support vector machine	90%	Positive or negative sentiment
Singh *et al.* [8]	10 Lakh sentences from tweets, Watsapp text, Facebook comments	n-gram approach	82%	The identification of code-mixed text encompasses various elements, including phonetic typing, abbreviations, clever wordplay, intentional misspellings, and slang.
Chopra and Bhatia [17]	Self collected	Survey approach	-	Positive or negative sentiment
Das and Bandyopadhyay [18]	Punjabi websites, newspaper	Supervised SVM	-	Linguistic: sadness, fear, surprise, anger
Soumya and Pramod [10]	Retrieving tweets using Twitter API	SVM, Naïve Bayes	95.6%	Positive and negative emotions

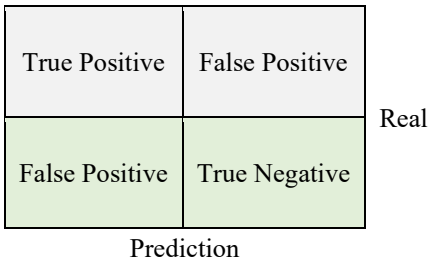

Figure 1. Diagram flowchart.

can be easily determined by subtracting accuracy from one. Understanding these metrics is crucial for evaluating and improving predictive models effectively.

Accuracy = True Positive + True Negative/True Positive + True Negative +

False Positive + False Negative

Precision

Precision quantifies the proportion of true positive instances relative to all instances identified as positive. It serves as a critical metric for assessing the accuracy of specific predictions, highlighting the reliability of results in high-stakes scenarios where certainty is essential. Emphasizing precision is vital, particularly in contexts where confident predictions can make a significant difference.

Precision = True Positive/True Positive + False Positive

Recall

It calculates fraction of positive cases that are correctly identified. It is the proportion of correctly predicted positive cases to the total number of positive instances in the dataset.

Recall = True Positive/True Positive + False Negative

F1-score

The F1-score ranges from 0 to 1 and helps us balance Precision and Recall. It addresses the challenge of comparing classifiers that might have high Recall but low Precision, or vice versa. Therefore, the F1-score provides a solution to this dilemma.

F1-score = 2×Precision×Recall/Precision + Recall

Table 2. Accuracy results.

Method type	Proposed algorithm	Unigram method	Simple scoring method
Lexicon based	56%	53%	54.50%
Negation handling	69.40%	60.25%	66.85%

RESULT EVALUATION

Results of our approach are compared with different available techniques that are as follows:

- *Unigram method*: Number of positive and negative polarity words are counted in a sentence and choose the highest polarity count for evaluating sentence's sentiments.
- *Simple scoring method*: We add positive and negative score of the polarity of a sentence and choose the one with dominant polarity.

We have tested data with our approach with unigram and simple methods (Table 2).

CHALLENGES

1. *Handling emoji based text*: It is of no doubt that emoji enhances the meaning of text but it becomes a challenging task to interpret around 3000+ emojis using sentiment analysis.
2. *Long and complex sentences*: Analyzing long sentences is a major issue in today's techniques hence new techniques should be used for interpreting complex sentences.
3. *Grammatical mistakes:* Sometimes, grammar mistakes can lead to misinterpretation of the sentence.
4. *Word order:* Punjabi is free-order language in which a subject, object and verb may come in any order. Hence, a slight variation in the order of the word may alter the entire meaning of the sentence and also changes its polarity.

CONCLUSION

In the present times, sentiment analysis is one of the fastest growing domain of artificial intelligence. To access people's view on web space manually is an impossible task, however with different involvements of new techniques makes it possible now to evaluate sentiments of digital users. This study focuses on the examination of the various sentiment analysis methods. Our algorithm works on 56% accuracy in subject lexicon and 69.40% accuracy in negation handling. Theoretically, machine learning (supervised) techniques have outperformed lexicon-based (unsupervised) methods in sentiment analysis in respect of accuracy. In this research we find polarity of different words by identifying features and negation in the sentence. In future, researchers may have scope to work on conjunctive and disjunctive nature of words to improve results.

REFERENCES

1. Kaur A, Gupta V. A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence (JETWI). 2013 Nov 1; 5(4): 367–71.
2. Mukherjee S, Bhattacharyya P. Sentiment analysis: A literature survey. arXiv preprint arXiv:1304.4520. 2013 Apr 16.
3. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr. 2008 Jul 6; 2(1–2): 1–35.
4. Karamibekr M, Ghorbani AA. Sentiment analysis of social issues. In 2012 IEEE international conference on social informatics. 2012 Dec 14; 215–221.
5. Sandhu JK, Garg P. Quality Evaluation of Product Reviews using Sentiment Analysis. Int J Latest Trends Eng Technol. 2022; 20(2): 018–022.
6. Dadvar M, Hauff C, De Jong FM. Scope of negation detection in sentiment analysis. In 11th Dutch-Belgian Information Retrieval Workshop, DIR 2011; University of Amsterdam. 2011 Feb 4; 16–20.
7. Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Third IEEE international conference on data mining. 2003 Nov 22; 427–434.

8. Singh M, Goyal V, Raj S. Sentiment analysis of english-punjabi code-mixed social media content to predict elections. In Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019. Singapore: Springer; 2021; 81–90.
9. Kaur A, Gupta V. Proposed algorithm of sentiment analysis for Punjabi text. Journal of Emerging Technologies in Web Intelligence (JETWI). 2014 May 1; 6(2): 180–3.
10. Soumya S, Pramod KV. Sentiment analysis of Malayalam tweets using machine learning techniques. ICT Express. 2020 Dec 1; 6(4): 300–5.
11. Arora P, Kaur B. Sentiment analysis of political reviews in Punjabi language. Int J Comput Appl. 2015 Jan 1; 126(14): 20–23.
12. Kaur G, Kaur K. Sentiment detection from Punjabi text using support vector machine. Int J Sci Res Comput Sci Eng. 2017 Dec; 5(6): 39–46.
13. Chopra FK, Bhatia R. A critical review of sentiment analysis. Int J Comput Appl. 2016; 149(10): 37–40.
14. Sharma A. Sentiment analyzer using Punjabi language. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE). 2014 Sep; 2(9): 5904–9.
15. Jain EU, Sandu A. Emotion detection from Punjabi Text using hybrid support vector Machine and maximum entropy algorithm. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE). 2015 Nov; 4(11): 89–93.
16. Deepali NG. Movie review mining in Punjabi. International Journal of Application or Innovation in Engineering and Management (IJAIEM). 2013; 2(12): 372–375.
17. Chopra FK, Bhatia R. Sentiment analyzing by dictionary based approach. Int J Comput Appl. 2016 Oct; 152(5): 32–4.
18. Das A, Bandyopadhyay S. Sentiwordnet for bangla. Knowledge Sharing Event-4: Task. 2010 Feb; 2: 1–8.
19. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intell Syst. 2018 Jan 24; 32(6): 74–80.
20. Ganeshbhai SY, Shah BK. Feature based opinion mining: A survey. In 2015 IEEE International Advance Computing Conference (IACC). 2015 Jun 12; 919–923.

