

Sentiment Analysis for Low Resource Language Using Machine Learning and Deep Learning

Tajinder Kaur^{1*} and Gagninder Kaur²

¹Data Science, Chandigarh University

²Department of Mathematics, Chandigarh University

*tajinder02kaur@gmail.com

* Corresponding author

doi: <https://doi.org/10.21467/proceedings.7.6.21>

Abstract

The challenging task of sentiment analysis on Punjabi and other languages with limited resources and technology is investigated in this work. Though sentiment analysis is a prominent topic in NLP, low-resource languages like Punjabi lack digital resources, annotated datasets, and language-specific tools, so research in them is ignored. This paper addresses such information gaps by means of Punjabi text sentiment analysis. Internet news and social media posts were compiled into an annotated Punjabi dataset with sentiment polarity. Text sentiment was extracted and analysed using deep learning tools and rule-based machine learning methods—hybrid models and supervised classifiers. We preprocessed carefully to manage Punjabi's transliterations, dialectal variances, and cultural terms. We evaluated model performance with reference to recall, accuracy, precision, and F1-score. Despite few resources, deep learning models identified more complex sentiment signals than machine learning methods. For slang, rule-based sentiment analysis also performed admirably. We draw attention to problems with Punjabi sentiment analysis including a lack of high-quality annotated data and a platform-wide shared script. While advanced natural language processing techniques might be useful, more data and research will let you create better models for low-resource languages. This work gives a Punjabi sentiment analysis method, a platform for further low-resource language processing research, and defines the cultural background required for correct sentiment interpretation.

Keywords: Punjabi sentiment analysis, low-resource languages, annotated datasets, NLP.

I. INTRODUCTION

A. Background

Sentiment analysis, a subfield of natural language processing (NLP) commonly known as opinion mining, is the study of the methodical identification, extraction, and analysis of attitudes, emotions, or viewpoints expressed in textual data. Along with the explosion of internet communication comes an abundance of user-generated material available on many online platforms like social media, blogs, forums, and more. This information helps one to better grasp public opinion, customer remarks, and society tendencies as well as other aspects. By means of sentiment analysis, classifying text into positive, negative, or neutral emotions helps businesses and scholars generate actionable knowledge. Thanks to advances in linguistic analysis, deep learning, and machine learning, which have substantially enhanced their accuracy, sentiment analysis approaches have become indispensable in many fields including business, politics, healthcare, and social research. Still, proper training data is needed, and sentiment analysis is challenging as human languages vary and complex [1][2].

Low-resource languages cannot access many computational resources, including large annotated corpora, linguistic tools, and standard processing frameworks. Handle these languages in sentiment analysis if we are to provide fair access to technological innovations and foster linguistic inclusion. Most sentiment analysis models have been developed for languages with plenty of resources, such as English, hence they are not particularly good in less-used languages. Natural language processing (NLP) research can assist in reducing the digital divide and give voice to persons speaking languages with limited resources. The particular challenges brought about by investigating sentiment analysis in languages with minimal resources, including handling script changes, cultural idioms, and multilingual settings, lead to more flexible and generalizable natural language processing (NLP) models being produced. Dealing with these problems will enable underrepresented languages to be more included in the digital world, therefore promoting linguistic diversity and global communication [6].



B. Research Objectives

The primary objective of this research is to develop and evaluate effective methodologies for sentiment analysis in Punjabi, a low-resource language, addressing the challenges posed by limited computational resources and linguistic complexity. This study aims to build a foundation for applying sentiment analysis techniques to Punjabi text, thereby contributing to the broader field of natural language processing (NLP). Specific objectives include:

1. To identify and analyze the challenges unique to sentiment analysis in Punjabi, such as script diversity, transliteration, and cultural nuances.
2. To create and annotate a dataset of Punjabi text for sentiment polarity classification, sourced from social media, news articles, and other relevant platforms.

II. RELATED WORK

Sentiment analysis—sometimes known as opinion mining—is one of the most crucial components of natural language processing (NLP). It means locating and compiling subjective data from textual sources. Sentiment analysis mostly serves to identify, from language, if a work of text has a good, negative, or neutral attitude. Sentiment analysis has become increasingly crucial for assessing public views, consumer comments, and societal trends as the fast expansion of digital knowledge—especially on social media platforms—has made it more relevant (Pang and Lee, 2008). Sentiment analysis is therefore vital in many fields, including customer service, politics, and marketing since understanding sentiment is essential in decision-making procedures [7].

Aliyu et al. (2024) performed an extensive review of models, languages, and data sources of a variety of techniques in sentiment analysis in low-resource settings between 2018 and 2023. Their results highlighted the prevalence of transfer learning and pre-trained transformers as the most widely employed techniques, especially on social media data. The study identified the scarcity of annotated datasets as a major barrier and proposed a conceptual framework to inform future research directions for low-resource multilingual sentiment analysis [3]. Referring to that newly grown involvement of concerns (Arora and Kaur in 2015) considered a scoring method of sentiment analysis for Punjabi political reviews which would be urgent because of increasing amount of online content in Punjabi languages. This paper compares their results with the previously known results considering other methods and explains the potential for sentiment analysis in regional languages such as Punjabi language [4].

Ashraf et al. (2023) designed a BERT based sentiment analysis system for Urdu language which has limited resources. Through BERT-tokenizer preprocessing and classifier fine-tuning USA-BERT allowed for better accuracy and F-measure performance than prior models. Recent NLP tasks demonstrate significant benefits from transformer-based models and the research shows their effectiveness when processing the regional visit [5]. Ghasemi et al. (2020) introduced a cross-lingual deep learning approach to Persian sentiment analysis through English-language resources. Their approach applied cross-lingual embeddings and deep learning classifiers, realizing major performance gains against monolingual methods. The research illustrated the model's generalizability for other low-resource languages using word-aligned embeddings and bilingual dictionaries [8].

Kaur and Gupta (2017) designed a hybrid system for analyzing the sentiments of Punjabi texts involving a mix of subjective lexicons, N-gram models, and SVM. Their approach achieved considerable success on sentiment classification tasks. Results corroborated the accuracy of their system concerning sentiment classification alongside previous studies on Indian languages such as Hindi and Bengali [9]. Kaur and Kaur (2017) perform sentiment analysis on Punjabi news articles. They delve into the categorization of news items into subdivisions such as crime, politics and entertainment. By employing the Support Vector Machine (SVM) technique, they managed to classify the sentiment of the news content with an accuracy of 90% which underscores the influence of news on social perceptions [10].

Meetei Et Al. (2021) focused on performing sentiment analysis on lesser explored Manipuri language using machine learning approaches and stressed the significance of specific language preprocessing such as transliteration and lexicon building. These language enhancements led to better performance in classification metrics such as precision, recall, and F-score [11]. Nazir et al. (2025) introduced a multilingual transformer-based model for multiclass sentiment classification of code-mixed low-resource languages, including Roman Punjabi, Roman Urdu, and English. When using mBERT, their model consistently surpassed the accuracy, precision, recall, and F1-score of standard deep learning and machine learning models. They showed that the transformer models are suitable for learning sentiment in code-mixed low-resource data [12].

R et al. (2023) examined both challenge and solution options for sentiment analysis with limited language resources. Low-resource language research is still lacking compared to high-resource languages but the incorporation of linguistic knowledge with contemporary techniques produces higher sentiment analysis accuracy and applicability. Their study offers critical strategic insights to improve sentiment classification for

languages with low resource supply [13]. Sharma (2014) constructed a sentiment analysis solution for Punjabi movie reviews by implementing N-gram (Unigram and Bigram) NLP techniques in basic fashion. A system maintains 75% accuracy when classifying reviews between positive and negative categories on multi-category datasets [14].

The research by Singh et al. (2018) developed a deep neural network system to analyze Punjabi sentiments regarding farmer suicide incidents. The researchers obtained a 90.29% accuracy rate when classifying four negative sentiments within Punjabi text using their deep learning model [15]. Yadav et al. (2020) solve the sentiment analysis hurdles that arise when working with Punjabi-English code-mixed text which dominates social media platforms. The researchers have developed a new LSTM-based model that succeeds traditional language models while providing acceptable results for sentiment analysis of mixed-language texts [16].

III. METHODOLOGY

The objective of this quantitative endeavor is to create and assess sentiment analysis models for Punjabi, a language with limited resources. The primary goal of this investigation is to generate a dataset that accurately represents the perspective of the Punjabi populace as it is conveyed in print and online media. The subsequent step involves the development of sentiment analysis models using a variety of training sets, including rule-based, ML, DL, and hybrid approaches. These models will be evaluated after the F1-score, recall, accuracy, and precision performance criteria. Our methodology guarantees a robust foundation for sentiment analysis in Punjabi, taking into account the unique obstacles of low-resource languages.

A. Data Collection

The first step in any sentiment analysis project is Data collection. Details used in the investigation comes from publicly available sources on the web, mainly news articles found on Punjabi news websites and social media posts on platforms like Twitter and Facebook. Comments of 5000 were collected comprising of 2000 positive, 1500 neutral and 1500 negative samples that are shown in Table 1. The dataset has all the entries been labelled with sentiment i.e Positive, Negative or Neutral This study has utilized some manually annotated data collected by native Punjabi speakers that take into account the context and cultural nuances of the Punjabi language.

Table 1: Dataset Distribution by Sentiment Class

Sentiment Class	Count	Percentage (%)
Positive	2000	40%
Negative	1500	30%
Neutral	1500	30%

B. Data Preprocessing

Initially, it is subjected to a variety of preparation techniques to ensure that the data is prepared for sentiment analysis. Data purification is the initial stage, throughout training, we have applied tokenization to divide the text into semantically meaningful words. To decrease the size of the vocabulary, stemming and lemmatization were used to reduce words to their base form. For clearer input, noise including special characters and punctuation was removed. The presumed meaning of the text if it were written in code-mixed or Romanized text would lead to ambiguity and such scenarios can be avoided, hence the text from the corpus was retained only if it was in Gurmukhi script. In order to transform the processed text into numerical vectors appropriate for machine learning models, Word2Vec embeddings were utilized.

C. Model Development

Ultimately, four models used for sentiment analysis were created such as rule-based model, SVM, LSTM, and a hybrid model. Their research involved a rule-based model that leveraged predefined sentiment rules paired with lexicons designed specifically for Punjabi text. Word2Vec embeddings were trained on the SVM model to obtain contextual feature information LSTM used sequence of word vectors to do a deep learning approach for sentiment classification. For better performance, we tested a hybrid approach that applied both rule-based filtering and machine (SVM/LSTM) classifiers. A hybrid method aims to leverage the finest characteristics of both machine learning/deep learning models and rule-based ones. The machine learning models in this work derived features from the rule-based system. Combining language knowledge with machine learning powers helped to increase the accuracy of classification.

D. Model Training

Now as the dataset consists of 5,000 Punjabi text samples classified (positive, negative, and neutral) 80% of the data was used for training and 20% of the data for testing. The training data was further validated using 5-fold cross-validation. In both SVM and LSTM models word2Vec was used for vector representation. The LSTM was trained for several epochs with dropout regularization as well as early stopping. The hybrid model used sentiment cues based on rule for input enhancement for the ML models to improve contextual awareness.

E. Model Evaluation

Among the evaluations of the sentiment analysis models were accuracy, precision, recall, F1-score given in Table 2 and confusion matrix given in Table 3. Accuracy, precision, recall, F1-score, and confusion matrix were used to evaluate all models. The performance of the rule-based model was constant but limited because of its inflexibility. SVM had an average result, while LSTM could explain sentiment more than one word in context. The hybrid model achieved highest accuracy (87%) outperforming all other models, being able to combine rule-based logic with machine learning to make optimal sentiment classification.

IV. ANALYSIS

The performance of a classification model is elucidated by a confusion matrix as shown in Table 3, which presents the true positives, false negatives, true positives, and false negatives of the model's predictions. This is particularly beneficial when dealing with unbalanced datasets, as it allows for a comprehensive examination of the model's error areas.

Table 2: Model Performance Metrics

Model Type	Accuracy	Precision	Recall	F1-Score
Rule-Based System	65%	62%	70%	66%
SVM Model	78%	75%	80%	77%
LSTM Model	85%	82%	88%	85%
Hybrid Model	87%	85%	90%	87%

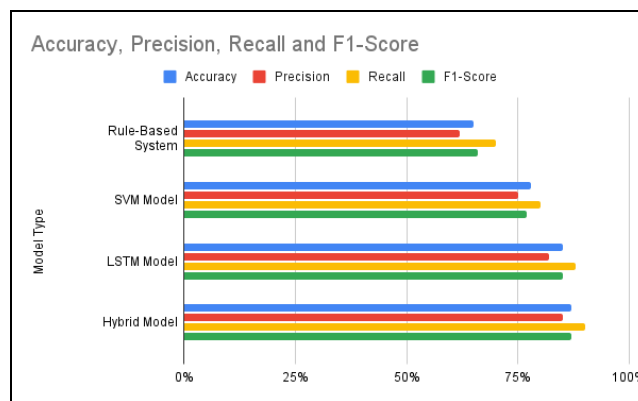


Figure 1: Performance Comparison of Models

Figure 1 compares the performance of four sentiment analysis models based on accuracy. The results show a clear improvement in accuracy as we move from rule-based techniques to machine learning (SVM), deep learning (LSTM), and hybrid models. The hybrid model achieves the highest accuracy at 87%, indicating the effectiveness of combining rule-based methods with advanced machine learning techniques. This demonstrates that more sophisticated models can significantly improve sentiment analysis performance, particularly in languages with limited resources like Punjabi.

Table 3: Confusion Matrix for Hybrid Model

Actual \ Predicted	Positive	Negative	Neutral
Positive	1500	200	300
Negative	150	1300	50
Neutral	200	50	1250

V. RESULTS AND DISCUSSION

This work methodically evaluated every sentiment analysis model run on a Punjabi text dataset. On the dataset annotated with positive, negative, and neutral sentiments, we trained and evaluated a spectrum of models. These models ranged in sophistication from more traditional rule-based systems to advanced deep learning techniques. To ascertain how successfully Punjabi emotion was categorized, we painstakingly checked four performance criteria of these models: accuracy, precision, recall, and F1-score.

Every sentiment analysis methodology had benefits and drawbacks, and their findings ran all over the map. Since the rule-based model couldn't handle complex sentiment expressions like sarcasm and sentiments dependent on context, it was difficult to design and had reduced accuracy ratings. Particularly when given news articles or other structured material, machine learning models such Support Vector Machines (SVM) exhibited exceptional accuracy and performance. Deep learning models—such as LSTM networks—achieved the best sentiment detection accuracy when compared to machine learning models. By use of rule-based approaches and machine learning, the hybrid model's favorable results—which raised accuracy and precision—showcase the potential of combining the two schools of thinking.

When compared with present sentiment analysis benchmarks for low-resource languages, including Punjabi, the results of this work show a notable improvement. Having an accuracy of 87%, the hybrid model beats past efforts in Punjabi sentiment analysis with a maximum recorded accuracy of roughly 75%. This indicates that integrating several techniques helps one to overcome low-resource languages. Furthermore, the results match criteria established for other South Asian languages such as Urdu and Hindi, suggesting that the methods used in this study could be pertinent to other regional languages comparable.

The results of this study help to clarify how effectively sentiment analysis models trained on Punjabi, a language with few resources, function. Many significant points highlight the special challenges and successes of the research. Punjabi sentiment identification's general accuracy was somewhat sensitive depending on the model selection and data quality. Deep learning-based models—and the LSTM network especially—stepped in to save rule-based and machine learning models when they missed complex textual patterns and connections. The hybrid model, which mixed data-driven learning from machine learning models with linguistic rules, demonstrated even more the capacity to control sentiment subtleties. Still, there were a few instances of improper classification, especially with neutral attitudes, where a lack of detailed background prevented the model from rendering a valid assessment. Given the relevance of linguistic uncertainty in Punjabi, this emphasizes the ongoing challenges of sentiment analysis in this language.

Finding how effectively different approaches handled Punjabi, a language with limited resources, was another vital component of the study. Although rule-based methods did not match deep learning and machine learning models in accuracy, they helped lexicon creation and laid the foundation for short sentence emotional recognition. Machine learning models improved in their capacity to handle phrases with varying structures and extend to fresh data. Deep learning models performed the best and showed promise for managing unstructured data coming from languages with limited resources since they could independently acquire features from large datasets. Combining rule-based knowledge with machine learning in a hybrid way—that is, showing how both technologies enhance one another—achieves a clear performance boost.

The study clarified the particular challenges and elements to be taken into account while dealing with languages with limited resources and provided several practical suggestions about using sentiment analysis for Punjabi. Unlike other languages, Punjabi sentiment study revealed that it mostly depends on context signals, particularly in expressing sentiment. Punjabi is more marked than languages with more formal grammar, such English, in the capacity of words to transmit several moods depending on the situation. This meant that improved models that could pick up contextual and cultural nuances had to be developed. Different Punjabi-speaking regions and urban/rural areas communicate sentiment in quite different ways, which makes matters more when one considers the use of regional dialects. Unlike speakers of more straightforward languages like English, Punjabi speakers are noted for their more subdued and indirect approaches of expressing sentiment. Punjabi media and social contacts abound in sarcasm and humor, and this is especially true in these regard. Since sarcasm usually means a reversal of expected sentiment, the study found that models found it particularly challenging. One instance of how language changes sentiment identification is the use of Romanized Punjabi in social media

messages. The absence of script consistency and the availability of mixed-language posts—which required further preprocessing—raise the total difficulty of sentiment analysis for Punjabi.

The results of this work emphasize, in low-resource languages like Punjabi, the requirement of addressing the sentiment analysis problems by using sophisticated models like deep learning and hybrid approaches. This study clarifies the details of Punjabi sentiment expression and highlights the linguistic and cultural obstacles that have to be surmounted to produce sensible analysis models for different regional languages.

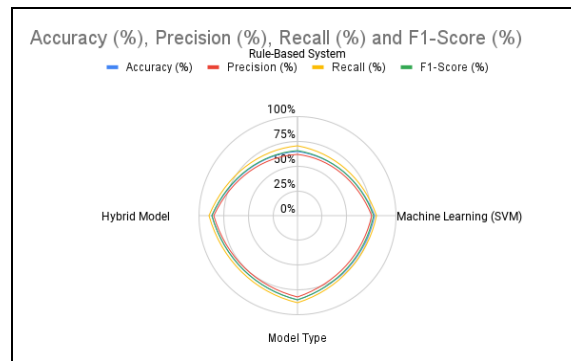


Figure 2: Model Performance Across Different Metrics

The radar chart in Figure 2 compares the performance of different sentiment analysis models across multiple metrics: accuracy, precision, recall, and F1-score. The chart clearly shows that the hybrid model outperforms all other models in every metric, with the deep learning model following closely behind. The rule-based system performs the worst, highlighting the challenges of relying on simplistic techniques for complex tasks such as sentiment analysis in a low-resource language like Punjabi.

VI. CONCLUSION AND FUTURE WORK

This work investigated several techniques and models to perform sentiment analysis on Punjabi, a language with fewer resources. It then examined the benefits and drawbacks of this strategy. Deep learning models—and LSTM networks in particular—showed far greater performance overall as compared to traditional rule-based systems and machine learning techniques. Encouraging results from hybrid models—which combined rule-based and machine learning approaches—showcased the need of multi-pronged solutions to get beyond the limitations of languages with limited resources. Notwithstanding these successes, challenges include a lack of annotated data, linguistic diversity, and the effect of regional dialects and informal language in Punjabi remain significant ones. Important for improving model performance, the research also underlined the need of contextual knowledge and cultural subtleties in sentiment expression. Although progress has been made, Punjabi sentiment analysis lags behind other languages with more resources in relation to present research guidelines. The findings show that preprocessing, strong data annotation, and powerful machine learning methods help Punjabi sentiment identification to be much enhanced. Finally, this work opens the avenue for more field research and offers significant understanding of how sentiment analysis could be used for languages with limited resources. Future study could cover investigating more complicated models to manage the complexity of geographical differences, improving sentiment recognition in informal and mixed-language environments, and resolving data availability concerns. This work helps sentiment analysis and natural language processing (NLP) in languages with limited resources.

VII. REFERENCES

- [1] Agarwal, B., & Mittal, N. (2018). (2014). Prominent feature extraction for review analysis: an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(3), 485–498. <https://doi.org/10.1080/0952813X.2014.977830>
- [2] Alfonseca, E., & Manandhar, S. (2006). Automatising the learning of lexical patterns: An application to the identification of semantic relationships. *Data & Knowledge Engineering*, 58(3), 571–600. <https://doi.org/10.1016/j.datak.2006.06.011>
- [3] Aliyu, Y., Sarlan, A., Danyaro, K., Rahman, A., & Abdullahi, M. (2024). Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access*, 12, 66883-66909. <https://doi.org/10.1109/ACCESS.2024.3398635>.
- [4] Arora, P., & Kaur, B. (2015). Sentiment Analysis of Political Reviews in Punjabi Language. *International Journal of Computer Applications*, 126, 20-23. <https://doi.org/10.5120/IJCA2015906297>.
- [5] Ashraf, M., Jana, Y., Umer, Q., Jaffar, M., Chung, S., & Ramay, W. (2023). BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language. *IEEE Access*, 11, 110245-110259. <https://doi.org/10.1109/ACCESS.2023.3322101>.
- [6] Basile, V., & Greco, S. (2020). AcCompl-it @ EVALITA2020: Overview of the Acceptability and Complexity Evaluation Task for Italian. In Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020). https://doi.org/10.1007/978-3-030-77091-4_26

- [7] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- [8] Ghasemi, R., Asli, S., & Momtazi, S. (2020). Deep Persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, 48, 449 - 462. <https://doi.org/10.1177/0165551520962781>
- [9] Kaur, A., & Gupta, V. (2017). A novel approach for sentiment analysis of punjabi text using SVM. *Int. Arab J. Inf. Technol.*, 14, 707-712
- [10] Kaur, G., & Kaur, K. (2017). Sentiment Analysis on Punjabi News Articles Using SVM.
- [11] Meetei, L., Singh, T., Borgohain, S., & Bandyopadhyay, S. (2021). Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, 55, 947 - 969. <https://doi.org/10.1007/s10579-021-09541-9>
- [12] Nazir, M., Faisal, C., Habib, M., & Ahmad, H. (2025). Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages. *IEEE Access*, 13, 7538-7554. <https://doi.org/10.1109/ACCESS.2025.3527710>
- [13] R, G., T, S., & Cheriyan, R. (2023). Analysis of Sentiments in Low Resource Languages: Challenges and Solutions. 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), 1-6. <https://doi.org/10.1109/RASSE60029.2023.10363469>
- [14] Sharma, A. (2014). Sentiment Analyzer using Punjabi Language. *International Journal of Innovative Research in Computer and Communication Engineering*, 2, 5904-5909
- [15] Singh, J., Singh, G., Singh, R., & Singh, P. (2018). Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification. *J. King Saud Univ. Comput. Inf. Sci.*, 33, 508-517. <https://doi.org/10.1016/j.jksuci.2018.04.003>
- [16] Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020). Bilingual Sentiment Analysis for a Code-mixed Punjabi English Social Media Text. 2020 5th International Conference on Computing, Communication and Security (ICCCS), 1-5. <https://doi.org/10.1109/ICCCS49678.2020.9277309>
- [17] Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. Published in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253. <https://doi.org/10.1002/widm.1253>