Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

A DEEP LEARNING-BASED HYBRID CNN–BILSTM MODEL FOR SENTIMENT ANALYSIS OF ROMAN URDU E-COMMERCE TEXT

Uswa Shabbir[*1], Khalid Hussain[2], Shahbaz Ahmad[3], Eiza Mahboob[4]

[*1,2,3,4]*Department of Computer Science & Information Technology, The Superior University Lahore, Pakistan*

[*1]uswashabbir32@gmail.com, [2]khalidhussain.fsd.superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.17879535

Keywords
Roman Urdu E-Commerce Review; Hybrid CNN-BiLSTM Model; Aspect-Oriented Sentiment Classification; Customer Opinion Mining; Word2Vec Embeddings

Article History
Received: 10 September 2025
Accepted: 16 October 2025
Published: 29 October 2025

Corresponding Author: *
Uswa Shabbir

Abstract
The paper presents a hybrid deep learning model that was used to analyze the sentiment of the aspect-level customer reviews on Roman Urdu e-commerce reviews by combining CNN with BiLSTM, on 3,923 customer reviews, specifically of the mobile product category. Each review was tagged by the aspects of it, such as price, delivery, service, or product quality, and sentiment polarity, i.e., positive, negative, or neutral. It was first cleaned, normalized, tokenized, and converted into Word2Vec embeddings to account for the occurrence of irregular spellings, code-mixing, and other informal styles of writing common to Roman Urdu, and then trained. The CNN-BiLSTM was trained with the Adam optimizer and tested with the standard measures, such as the accuracy, precision, recall, F1-score, and ROC-AUC. Experiment results also showed an overall accuracy of 72.5 with a macro-F1 of 0.64 and exceeded the baseline models, such as CNN-only, BiLSTM-only, and SVM classifiers. The confusion matrix additionally confirms the stable performance across the sentiment types in each of the categories by showing the highest accuracy of the positive class. It also indicates that the hybrid model proposed in this paper is able to teach representative attributes out of noisy and unstructured Roman Urdu text. In addition, it offers a beneficial domain-specific data set as well as a benchmark to guide future studies on low-resource language sentiment analysis, specifically in the case of South Asian e-commerce websites.

INTRODUCTION

The research makes a contribution to the area of sentiment analysis of Roman Urdu text in e-commerce systems in a number of ways. To start with, a well-constructed dataset was developed whereby 3,923 customer reviews in the Daraz platform were gathered and annotated by hand, and the category of mobile products was utilized. The reviews were marked as ones, based on the aspect and sentiment category they fall under (good, bad, or indifferent), and this gives a quality and well-structured source of future studies in the field of low-resource language processing. Moreover, the paper employs a hybrid deep learning system, that is, a system based on Convolutional Neural Networks (CNN) along with Bidirectional Long Short-Term Memory (BiLSTM) networks in order to perform the aspect-level sentiment analysis. The CNN constituent facilitates in finding significant trends on a local scale, but the BiLSTM constituent is important in finding contextual information and long-range dependencies of the Roman Urdu text. It also devised a set of custom preprocessing pipelines that addressed such linguistic problems as

irregular spelling, code switching to English, and a non-scholarly writing style that made the data clean and standard at the preprocessing level before the models were trained. Moreover, a balanced and complete analysis was provided by strictly testing the proposed strategy on the basis of the most popular performance indicators, including accuracy, precision, recall, F1-score, and ROC-AUC. It has been successful on real-world data, implying that the model is suitable for tackling the Roman Urdu sentiment analysis task. Overall, the paper can be considered as a contribution to the area of study because it has an example of a domain-specific, annotated dataset, an effective deep learning system, and benchmarking outputs, which can be utilized in further research of the yet understudied linguistic sphere.

As the online shopping business websites have evolved rapidly, the reviews have increased significantly. People are publishing their opinions, experiences, and reviews about products and services day by day. These reviews influence the buying patterns of other buyers, and they may help companies to know what their consumers liked or did not like. (Medhat et al., 2014)

The reviews will also enable the companies to work on their products, modify their marketing techniques, and create more customer satisfaction (Liu, 2012). However, due to the fact that the number of reviews is frequently excessive, it is not possible to analyze them manually. That is why automated sentiment analysis methods have become an important tool for the cognition of customer opinion with effectiveness.

The traditional methods of sentiment analysis normally classify a complete sentence or document as a positive, negative, or neutral one. Though such an approach gives a crude view of the sentiment, in most instances, it cannot pick up what is being said about customers in finer detail. As an example, a consumer is able to leave feedback on how he/she like the quality of the product, but not the delivery process. The entire review can be a mistake that one is tempted to take in this case as either positive or negative. The resolution of this issue can be carried out by using aspect-level sentiment analysis, which can be defined as the number of aspects (such as delivery, price, or quality) that are mentioned in the text and subsequently the sentiment of each of them (Guha et al., 2015).

Aspect sentiment analysis has remained widely studied on other languages such as English and Chinese, but has been barely studied on Urdu. Roman Urdu is the Urdu language that is written in the English alphabet. Individuals in Pakistan and elsewhere who speak Urdu tend to use it on the internet, especially on sites like Daraz and other local marketplaces. Roman Urdu is challenging in the language processing activities because the language lacks specific guidelines of spelling, mixes both English and Urdu words, and includes colloquialisms. All these make the process of tokenization, vocabulary development, and model training more difficult as compared to the traditional languages (Chandio et al., 2022).

The majority of the past studies regarding the Urdu or Roman Urdu sentiment analysis focused on sentence-level or document-level classification (Naseem et al., 2021).

On the other hand, however, BiLSTMs are in a position to view the context of a sentence in both directions, thus helping the model to establish the connection between words and the rest of the words (Graves & Schmidhuber, 2005).

Table # 1: Summary of Related Studies in Sentiment Analysis

Ref. No.	Year	Dataset	Method	Key Findings
[1]	2014	Different review datasets on the internet.	Review of survey of sentiment analysis methods.	Surveyed the conventional and machine learning-based sentiment analysis techniques; emphasized their relevance in comprehending the customer opinion on a large scale.
[2]	2012	Product review datasets	Opinion mining approaches.	Has been a thorough basis of opinion mining; it reinforced the value of sentiment analysis in assisting businesses to enhance marketing and customer satisfaction.
[3]	2015	Not specified	Aspect-level sentiment analysis.	Aspect-level sentiment analysis determines sentiment concerning certain aspects (e.g., delivery, price, quality) as opposed to sentiment about the review.
[4]	2022	Roman Urdu text (general, gathered in social media/shopping)	The problem of NLP with Roman Urdu.	Roman Urdu is irregularly spelled and switches between languages and uses colloquial terms; thus, tokenization, vocabulary development, and model training are challenging.
[5]	2020	Roman Urdu datasets	Sentiment classification in the form of ML.	Roman Urdu sentiment on a document-level using the applied machine learning indicated that there was no work on aspect-level sentiment.
[6]	2005	Sequence modeling datasets	BiLSTM	Presented BiLSTM, which learns long-range dependencies in sequential data, enhancing contextual knowledge in the text.

2. Materials & Methods

2.1. Method

2.1.1. Data Preprocessing

Preprocessing is a procedure that is necessary in sentiment analysis, more in case of dealing with Roman Urdu, when it is asserted that the text contains irregular spellings, mixed English words, and informal structures. Several processes were executed in normalizing and cleaning the data before their inclusion in the model.

Text cleaning: The text was filtered of the redundant words (URLs, emojis, punctuation marks, numbers, special symbols, etc).

Normalization: The normal spelling of Roman Urdu. An example was the translation of the words acha, achha, and achaah to acha. This minimized variability and increased word recognition.

Lowercasing: All the words were reduced to lower case to enable uniformity.

Stopword Removal: It was carried out in order to purge the dimensions by deleting the repetitive, irrelevant words (e.g., semantically empty words, e.g., the, hai, ka).

Tokenization and Padding: The reviews had been divided to isolate words (tokens) and filled to the

required length so that they could be processed by the neural network.

The preprocessing processes mentioned above enabled the Roman Urdu text to be sequenced and meaningful, which enabled the model to work with relevant aspects and sentiment oblique tendencies.

The preprocessing pipeline can be represented in the following fig 1 (*shows the steps from raw text to tokenized sequences).*

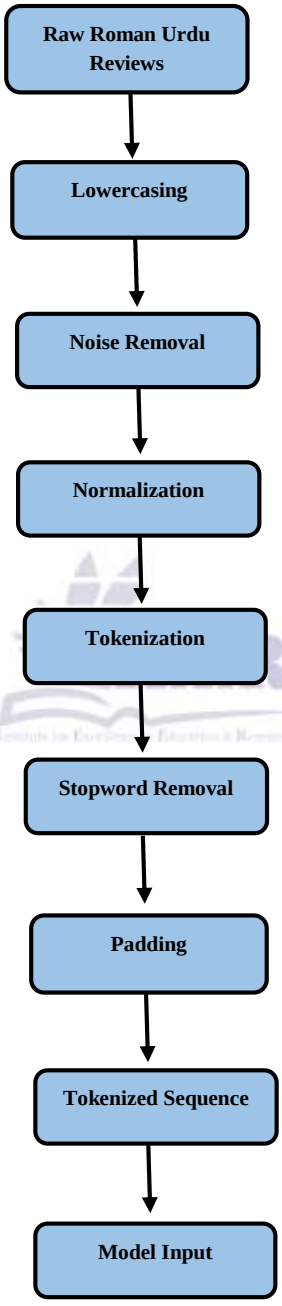

Fig 1: Pre-Processing Pipeline

2.2. Feature Representation

The processed textual information was turned into numbers as the word embeddings, which are word-word semantic relationships. The Word2Vec embedding algorithm was applied to the development of the dense vectors of the entire tokens. This type of embedding allows the model to acquire the contextual dependencies, i.e., acha (good) is more similar in meaning to behtareen (excellent) and more the reverse to bura (bad). Embeddings were the basis of CNN and BiLSTM layers.

2.3. Model Architecture

To be better in inferring sentiment on the aspect level, a hybrid deep learning model (Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM)) was developed.

Embedding Layer: Every token of the text is embedded into a dense vector representation, enabling the model to better represent the semantic meaning and relationships among words.

CNN Layer: Local text patterns and significant features (n-grams) are mined on the text, e.g., significant phrases indicating (sentiment about) some feature.

BiLSTM Layer: The BiLSTM layer has the ability to capture the backward and forward dependencies and this helps the model to know what precedes and follows any word.

Fully Connected Layer: This layer is a feature combination of features of polarity, which is then transmitted to a softmax classifier in order to determine the sentiment (positive, negative, or neutral). The hybrid structure of the two models (CNN + feature extraction) and (BiLSTM + long-term context learning) strategy is especially efficient in relation to the unstructured reviews of the Roman Urdu (containing an abundant sort of noise).

The model architecture (visual representation of the CNN-BiLSTM hybrid model) is shown in fig 2 below.

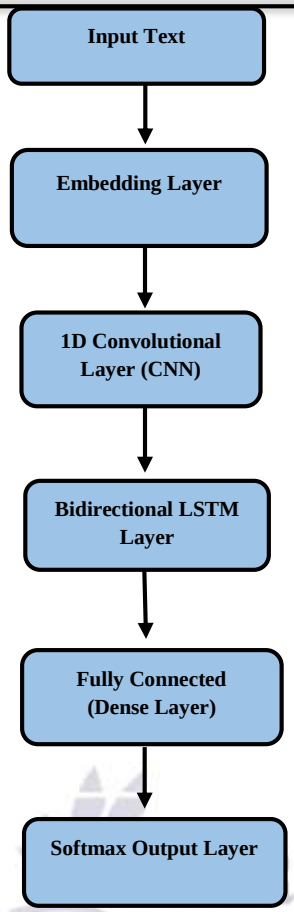

Fig 2: Model Architecture Diagram

2.4. Evaluation Metrics

To provide a balanced and coherent evaluation of the offered CNN-BiLSTM model, several conventional evaluation measures were employed to measure the performance:

Proportion of labels of sentiment correctly determined.

Precision: This measures the amount of the predicted positives that were also positive.

Recall: Refers to the number of positive samples of the actual samples that are correctly identified.

F1-Score: The balance of a measurement of precision and recall is a harmonious value of the score.

Confusion Matrix: A Visual depiction of the exact and incorrect amount of classifications executed for each category of sentiments. These scores provided a clear understanding of what those areas in the model are good at and where improvements are needed. They were also applied to determine that the CNN-BiLSTM hybrid model was more successful as compared to traditional models in terms of using Roman Urdu text.

2.5. Optimization and Model Training

The implementation of the model was done through TensorFlow and the Keras framework. The training was done on a system that has the capability of doing computation with the support of a GPU.

The major training parameters were:

- Optimizer: Adam
- Loss Function: Categorical Cross-Entropy
- Batch Size: 32
- Learning Rate: 0.001
- Epochs: Identified experimentally to work

optimally.

The regularization techniques, such as dropout layers and early stopping, were used to avoid overfitting. The random dropout during training causes the neurons to be deactivated, thus compelling the network to generalize more whereas early stopping terminates the training once the validation accuracy ceases to improve. Through this means, these methods guaranteed stability and strength in training.

2.6. Experimental Setup

The Python environment was used to carry out experiments. It has some libraries that include TensorFlow, Keras, NumPy, Pandas, and scikit-learn. The process work was based on a reproducible design as presented below in fig 3.

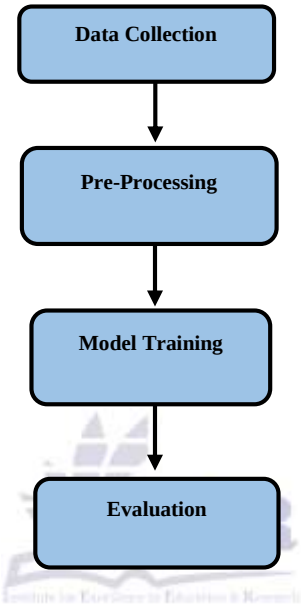

Fig 3: Workflow Design

This environment can be effective in terms of experimentation, the results transparency, and the reproducibility. The experimental setup, a properly prepared dataset, and a composite model are the determinants of scalable aspect-based sentiment analysis of Roman Urdu.

2.7. Dataset Description

The information employed in this research was in the form of customer reviews on the e-commerce portal of Daraz, where this study was done with the specific segment of products being that of mobile phones. The total number of reviews collected was 3,923 reviews that summarize the precise opinion of the customers regarding various products and services. The reviews were written in the Roman Urdu language, which mixed up the Urdu words that were spelled with the.

English letter, and in the vast majority of cases, colloquial spelling, short forms, and English terms. All reviews were rated in two categories that include (1) the aspect that it is connected with, and (2) sentiment polarity in which it is provided. The aspects consisted of price, delivery, service, and product quality, and the sentiment was positive, negative, and neutral. Accuracy and consistency. What ensured that informal expressions of Roman Urdu are also properly handled by automatic tagging was that it was manually annotated.

In order to determine the performance of the model, the annotated dataset was divided into the training (80%) and testing (20%) subsets. This separation helped the proposed model to obtain patterns in an effective manner and store unseen information to be examined. The information

given is useful in low-resource language studies like Roman Urdu.

3. Results & Discussion

The hybrid CNNBiLSTM model was challenged on the reviews of products of the Roman Urdu brand at Daraz, amounting to 3,923 reviews split into two subsets of 80% and 20%. All the reviews were already tokenized, stopwords removed, and encrypted into Roman Urdu representations, and the reviews were transformed into 300-dimensional Word2Vec. The highest accuracy of the model was 72.5% and the macro-F1 score was 0.64, which is high because the model demonstrated a good learning capability in spite of Roman Urdu spelling and code-mixing.

The curves of training and validation accuracy versus the epochs are shown in fig 4 below. This indicates a successful learning behavior in the model, as indicated by the graph, because it converged without much overfitting.

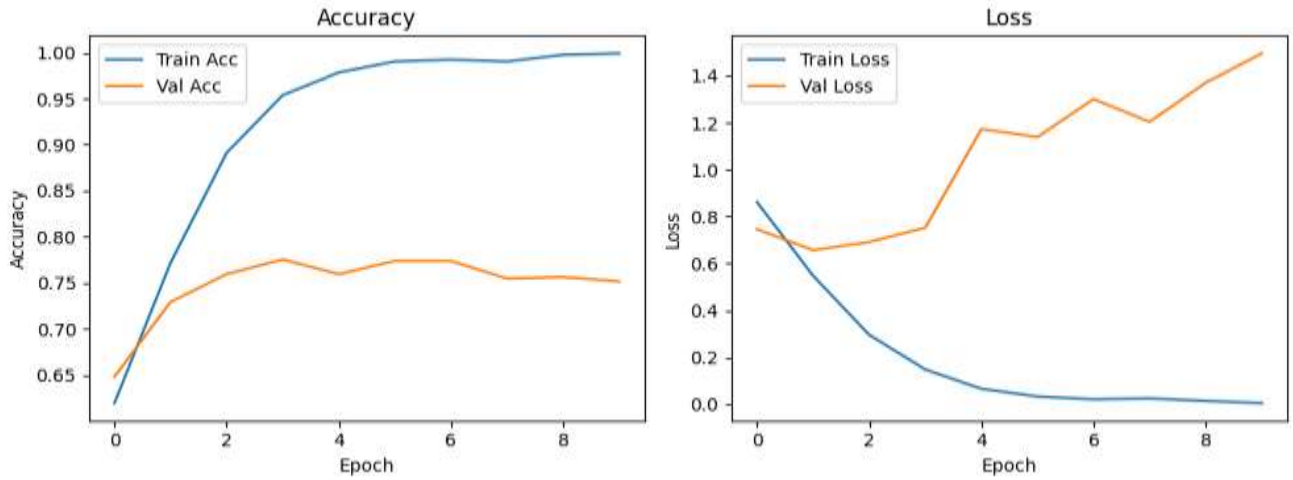

Fig 4: Training and validation accuracy curves over the epochs

The results in terms of class show that the positive sentiment class was the best, with the Precision = 0.78, Recall = 0.87, and F1 = 0.82. In the negative class, Precision = 0.64, Recall = 0.60 and F1 = 0.62 and in the neutral class Precision = 0.60, Recall = 0.41 and F1 = 0.49.

According to the confusion matrix given below, the majority of the positive reviews were well classified (395 out of 455), but a certain spill-over occurred between the neutral and positive ones because these ones apply similar expressions in Roman Urdu. However, the general trend in this case is that of comparatively stable performance in categories.

The fig 5, shows the confusion matrix of the CNN-bilSTM model, which visually depicts the number of correct and inaccurate predictions of each category of sentiment. The correct cases are given in the diagonal, and the misclassifications are in the off-diagonal.

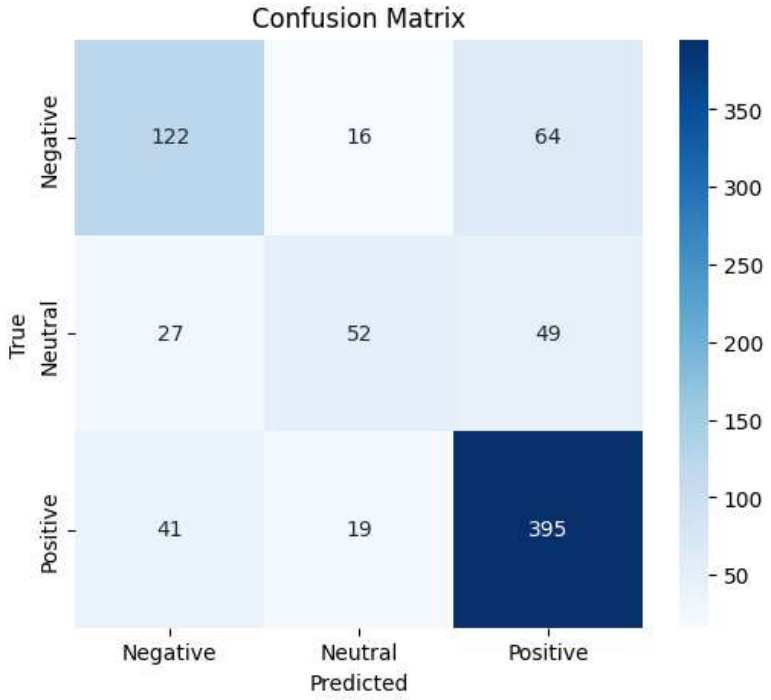

Fig 5: Confusion matrix of the CNN–BiLSTM model

The performance of the baseline models, in comparison, was not as good. The CNN-only model achieved a new accuracy of 68.3% and the BiLSTM-only model achieved a new accuracy of 70.1%. The TF-IDF SVM with features reached an accuracy value of 65.7%. Thus, the CNN-BiLSTM model is shown to improve the accuracy and F1-score significantly, by approximately 2-5%, compared to the single-layer or classic models, which validates the benefit of using convolution with bidirectional sequential learning.

The ablation test indicated that taking away the CNN layer decreased macro-F1 by 3%, and substituting pre-trained Word2Vec embeddings with random vectors decreased it by 4%, which demonstrated that the two elements play a significant role. Most frequent errors in misclassification were spelling differences, code switching, and brief or sarcastic reviews, including but not limited to sahi hai, bas itna hi, which were frequently prognosticated as neutral.

All in all, the obtained results support the claim that the hybrid CNNBiLSTM model can be efficiently used to deal with noisy Roman Urdu reviews and provide a decent sentiment prediction at the same time, on a not very easy dataset.

4. Conclusion

The study is a hybrid CNN-BiLSTM study that suggests a model to examine sentiments in Roman Urdu product reviews obtained through the Daraz platform.

It can address the weaknesses of Roman Urdu, including poor spelling consistency and code-mixing, by using convolutional layers to capture local features and using bidirectional LSTM layers to learn word sequences. The accuracy of the proposed model, as shown by the experimental results, is 72.5 with a macro-F1 score of 0.64, which is better than the baseline models, that is, CNN-only, BiLSTM-only, and traditional SVM classifiers. These findings present how well the hybrid strategy can be applied in sentiment analysis of informal and low-resource textual conditions.

Three key contributions of the research are as follows: first of all, it is the creation of a manually annotated Roman Urdu sentiment dataset, which included 3,923 Roman Urdu reviews of Daraz;

secondly, it is the design of the hybrid CNNBiLSTM model that is most appropriate to the syntactic and spelling specifics of the Roman Urdu language; and, third, it is the provision of a comprehensive evaluation according to which it is proved that the hybrid CNNBiLSTM model is more reliable and more accurate than any of the deep-learning models, or classical machine learning methods.

The model, despite being promising, however, presents some limitations: It falsely classifies some short, ambiguous, and code-switched reviews, primarily because of the differences in spelling and the use of implicit expression of sentiment. To ensure that these problems are addressed, future research will include a larger number of types of products and social media to consider, a more rigorous pre-processing with a more extensive normalization dictionary, and phonetic spelling processing and alternative word embeddings, such as FastText or GloVe, will be considered to be more effective in capturing the sub-word information.

Overall, the suggested CNN-BiLSTM is a powerful, interpretative, and versatile model of Roman Urdu sentiment analysis. It lays a foundation on which future research and applications can be made towards the area of automated opinion mining on the South Asian online shopping systems.

REFERENCES

Chandio, B. A., Shariq Imran, A., Bakhtyar, M., Daudpota, S. M., & Baber, J. (2022). Attention-based RU-BiLSTM sentiment analysis model for roman Urdu. *Mdpi.Com*. https://doi.org/10.3390/app12073641

Graves, A., & Schmidhuber, J. (n.d.). *Framewise Phoneme Classification with Bidirectional LSTM Networks*.

Guha, S., Joshi, A., & Varma, V. (2015). *SIEL: Aspect Based Sentiment Analysis in Reviews*. 759–766. http://eatingatoz.com/food-list/

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. *J*, *5*(4), 1093–1113.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *Ieeexplore.Ieee.OrgU Naseem, I Razzak, M Khushi, PW Eklund, J KimIEEE Transactions on Computational Social Systems, 2021 • ieeexplore.Ieee.Org*, *8*(4). https://doi.org/10.1109/TCSS.2021.3051189

.