**Deep Learning-Based Sentiment Analysis for Roman Urdu Text**

**Syeda Noor e Sehar Ali**

**Superior University,Lahore ,Pakistan**

**Abstract:**

Sentiment analysis is receiving a lot of attention because of its adaptable method ssfor analysing users' sentiments on various social networks, forums, e-commerce websites, and blogs. Information on the web that is related to emotions is very important and has an impact on clients, readers, and corporate organisations. Due to its efficient modelling of sequential input, the Reccurent Neural Network has been widely used to carry out Natural Language Processing tasks. We employed a deep neural long-short temporal memory model in this paper (LSTM). It possesses exceptional ability to capture long-range information, resolve gradient attenuation problem, and brilliantly depict future contextual information and word sequence semantics. This work serves as the basis for modifying deep learning techniques to carry out Roman Urdu Sentiment Analysis. Our test results demonstrate our model's substantial accuracy, which is higher than that of industry-standard machine learning techniques.

Keywords:Recurrent Neural Network(RNN);Long Short-Term Memory(LSTM);Roman Urdu Sentiment Analysis;Word Embedding

1. **Introduction:**

Sentiment analysis is a crucial component of Natural Language Processing (NLP), which collects and processes relevant information from user reviews expressed through online communities and collaborative media. In order to advance and increase the quality of scientific and commercial intelligent systems' services and goods, the opinionated data is becoming increasingly significant from an application point of view.[1] [2].

We tested our model using the Roman Urdu Corps[3].Roman Urdu is a widely used dialect with an easy-to-read and write script that non-native Urdu speakers can use everywhere in the world. Sentiment analysis in Urdu, both romanized and traditional, has been employed in a number of research. The primary technique of work has been to implement fundamental methodologies based on lexicon and machine learning. All of these methods have limitations, such as an inability to do sentiment analysis on huge datasets, a restriction on the use of new data, and a reliance on lexicons with fixed word counts and assignation numbers. In recent years, deep learning has advanced significantly in a number of fields, such as image identification, self-driving cars, and other intelligent systems[4]. Recurrent Neural Network (RNN) has outperformed in Natural language processing Tasks (NLP) because of its memory-capable ability to grasp long-term dependencies in sequential input. Recurrent neural networks (RNN) outperform traditional methodologies like the Lexicon, Machine Learning, and statistical algorithms like the Hidden Markov Model in terms of performance (HMM) [5] The LSTM network [6] in RNN is regarded as a significant attempt to describe sequential data, such as text and speech[7].

]In our suggested strategy, we employed long short-term memory LSTM networks [6] to word segmentation job in order to improve the state of the art performance of the Roman Urdu Sentiment Analysis. The contribution of strategy is as follows:

1) For the Roman Urdu Sequential Modeling assignment, our work is the first to apply Deep Neural Network long short-term memory (LSTM) to capture long-term sentence relationships.

2) In comparison to baseline machine learning and vocabulary-based approaches, it gets the maximum accuracy in binary classification of Roman Urdu Sentence.

1. **Methodology:**

Using the balanced binary dataset 8K, we carried out a 10-fold cross-validation in our experimental study. In each fold, 90% of the dataset is used to create our train set, and the remaining 10% is used to create our validation and test sets. The word embedding are initialised with 300-d for our purpose (300 dimensional) Education of the mini-batch random gradient descent approach is used to run all the model parameters and delivers an adjustable learning speed. We use Dropout to regularise neural networks and prevent the overfitting problem having a 0.1 dropout rate.The loss function is the NN activation for cross-entropy loss. The trainer of the model wasuse a test set and training set, respectively, to gauge our model. The model training was permitted to run for a maximum of 100 epochs in order to prevent over-fitting. We compared machine learning baseline approaches and our suggested deep neural network mode in the findings section. In another section, we applied Deep Learning algorithms to do Sentiment Analysis while using publically accessible FastText Urdu word embedding. In order to examine and comprehend the issue, we evaluated Precision and Recall and employed conventional metrics for categorization accuracy. We employed F1, where F-score is the harmonic mean of recall and precision.Additionally, for each baseline and Deep Learning Method, Accuracy of Deep Learning Model Validation. It displays a validation accuracy of 0.95 and a validation loss of 0.0. As shown in Table 1, our deep learning model overall produced substantial results of 0.92 for Naive Bayes and 0.88 for Random Forest. The same dataset was used for both machine learning and deep learning techniques, as was previously noted. In accordance with Table 1, our model outperformed all baseline machine learning techniques, outperforming them in accuracy and F1 with a 0.95 Accuracy and 0.94 F1 score.

Table 1. CLASSIFICATION RESULTS

|  |
| --- |
| **Classifiers Precision Recall F1score Accuracy** |
| **NB 0.79 0.77 0.77 0.77240**  **RF 0.88 0.88 0.88 0.88691**  **SVM 0.93 0.92 0.92 0.92472**  **Our DL Model 0.97 0.9287 0.94 0.95180** |

1. **Conclusion**

As the LSTM network is particularly effective for sequential Data Models, we employed it for the first time in this paper to train the model for Roman Urdu Sentiment Analysis. Deep neural networks are the ideal model to execute sequential data models, according to our experimental findings, as they don't require any prior knowledge, designing, or feature engineering. The accuracy of our model has surpassed that of baseline machine learning and lexicon-based approaches. We argue that using LSTM networks with word embedding is an excellent method for performing sentiment analysis. Our Model will make sure that there is more exploration.

**Sentiment Analysis for Roman Urdu Text**

References

[1]F. Xing , E. Cambria , R. Welsch ,2018,Natural language based financial forecasting: a survey, Artif. Intell. Rev. doi:10.1007/s10462-017- 9588-9

[2]Laercio Dias,2018.” Using text analysis to quantify the similarity and evolution of scientific disciplines”, Royal Society.

[3]https://en.wikipedia.org/wiki/RomanUrdu.

[4]Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig,2013. ”Joint language and translation modelling with recurrent neural networks. In EMNLP, volume 3,page 0.

[5]Zhang, 2003,“Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain”.

[6]Sepp Hochreiter and Jurgen Schmidhuber 1997,”Long short-term memory. Neural computation”, 9(8):1735–1780.

[7]Martin Sundermeyer, Hermann Ney, and Ralf Schluter,2015, ”From feedforward to recurrent lstm neural networks for language modeling. Audio, Speech, and Language Processing”, IEEE/ACM Transactions on, 23(3):517–529.