Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2302.09

2023, 42(2) 75-87

A machine learning approach for urdu text sentiment analysis

Muhammad Akhtar, Rana Saud Shoukat, Saif Ur Rehman *

University Institute of Information Technology, PMAS Arid Agriculture University, Rawalpindi Pakistan

* Corresponding author: Saif Ur Rehman, Email: saif@uaar.edu.pk

Received: 11 November 2022, Accepted: 24 March 2023, Published: 01 April 2023

K E Y W O R D S

Machine Learning
Sentiment Analysis
Urdu Text
Roman Urdu Language
LSTM
Lexicon
Neural Networks
Feature Extractor
Sentiment Extraction
Evaluation Measures

A B S T R A C T

Product evaluations, ratings, and other sorts of online expressions have risen in popularity as a result of the emergence of social networking sites and blogs. Sentiment analysis has emerged as a new area of study for computational linguists as a result of this rapidly expanding data set. From around a decade ago, this has been a topic of discussion for English speakers. However, the scientific community completely ignores other important languages, such as Urdu. Morphologically, Urdu is one of the most complex languages in the world. For this reason, a variety of unique characteristics, such as the language's unusual morphology and unrestricted word order, make the Urdu language processing a difficult challenge to solve. This research provides a new framework for the categorization of Urdu language sentiments. The main contributions of the research are to show how important this multidimensional research problem is as well as its technical parts, such as the parsing algorithm, corpus, lexicon, etc. A new approach for Urdu text sentiment analysis including data gathering, pre-processing, feature extraction, feature vector formation, and finally, sentiment classification has been designed to deal with Urdu language sentiments. The result and discussion section provides a comprehensive comparison of the proposed work with the standard baseline method in terms of precision, recall, f-measure, and accuracy of three different types of datasets. In the overall comparison of the models, the proposed work shows an encouraging achievement in terms of accuracy and other metrics. Last but not least, this section also provides the featured trend and possible direction of the current work.

1. Introduction

There has been a significant amount of research work done in the subject of sentiment analysis during the last ten years. Sentiment analysis approaches have been used to classify algorithms for a range of purposes, including medical diagnosis and treatment. Efforts in sentiment analysis have grown significantly in recent years because of its applications in online product purchase, marketing, review analysis, and reputation management [1]. Data generated by users on social networking and e-commerce sites have become a valuable resource [2]. In order to get global input on their actions and goods, production and selling are increasingly using this resource [2-3]. Every day, millions of Roman and Urdu lines are shared on social media platforms including Twitter, Snapchat, Facebook etc. [4]. When people's opinions in Roman and Urdu are disregarded and only English is considered, a huge quantity of data is lost.

In the last several years, as the usage of cell phones and the internet has increased dramatically, so has the number of people using social media sites like Twitter and Facebook, as well as blogs, to express their feelings [23, 25-30, 36]. Currently, peoples from different domains and different cultures wish to express their views in a public forum for getting feedback from experienced community about their products, policies or any viral news. As result, communities as well as an individual-based discussions are continuously increasing, that helps businesses, institutes and peoples to get such informative analysis from social media [16]. Not only this, Intelligent technologies, such as sentiment analyzers, which can transform raw data from social media platforms into information that can be put to good use are in high demand [45]. That also recognizes and detects emotions in the sentiment analysis process in order to correctly analyze the data. Further, languages such as French, English, Spanish, and other European languages must be addressed in terms of tool accessibility. Despite this, languages like Punjabi, Urdu and Hindi are seen as lacking in [5-10].

Pakistan's native language is Urdu., the world's sixth biggest country [1-3]. There are around 66 million speakers of this language, mostly on the subcontinent of South Asia. Hindustani is also a social linguistic variety of Urdu, which is widely spoken in India [3]. Sentiment analysis of Urdu is a significant method for understanding the behavioral elements, cultural beliefs, and social habits of individuals in this region of the world. Urdu borrows a significant number of terms from other languages such as Roman, Arabic etc [4], resulting in a rich morphology and complicated grammar. As a result, the computational tools are for Urdu [4], the tools available for other languages are insufficient, and it necessitates a more specific set of tools, particularly.

Urdu is a particularly challenging language to learn because of its many dialects, such as: According to the abjad system, all long vowels and consonants have to be written, although diacritics are optional [24, 7]. The numbers are written from left to right, but the letters are written from right to left, making it a bidirectional language.

This study's focus is on developing a novel SA approach for Urdu language based review. The following are some of the top aims of the proposed research:

- The primary goal of this study is to identify the various lexical resources needed for Urdu sentiment analysis and to explore the necessity of pre-processing Urdu text.
- Evaluate the current state of Urdu sentiment categorization methods and tasks.
- The core objective of this work is to explore the role of modifiers and negations in Urdu sentiment analysis
- Discuss the constraints of current methodologies and provide a list of unresolved issues and potential solutions, as well as future prospects for Urdu sentiment analysis.

The research limitations cited above may require more refined technique to strengthen the Urdu SA process. In addition, it is necessary to increase the accuracy of Urdu SA procedures by applying an improved methodology to the narrow regions (Roman Urdu) of the sentiments [19-20]. Therefore, the following is a list of the research's key contributions.

1) a comprehensive analysis of the existing approaches on Urdu language reviews/comments sentiment analysis.

2) Proposal of a novel conceptual framework for the analysis of the Urdu language reviews from different domain.

3) Implementation of the proposed conceptual framework

4) Performance evaluation of the proposed approach using different datasets

5) A detailed comparative analysis of the proposed technique with the existing techniques to evaluate the supremacy of the proposed model.

The rest of the paper is organized as follows: Section 2 describes the analysis of the state of art techniques for sentiment analysis relevant to Urdu language comments and reviews. Section 3 demonstrates the proposed conceptual model involving data pre-processing, and classification of the sentiments. Results of the experiments, as well as comparisons with existing approaches, are presented and discussed in Section 4. Section 5 provides a conclusion as well as recommendations for future work.

2. Literature Review

In comparison to other resource-rich languages such as English, sentiment analysis of Urdu language is still in its early stages of development. Furthermore, only a small amount of research has been done, which has a direct impact on review and survey papers that are presently accessible. This section provides some of the core research that has been carried out in Urdu SA in

different. This section also highlights the research studies of the different phases of Urdu SA process. At the end of this section, the research gap is summarized to show the importance of the proposed work.

In 2017, Daud et al., looked at numerous pre-processing approaches and linguistic resources in Urdu language processing, discussing best practices like recognition, tokenization, etc [24]. They also discussed how Urdu language processing can be used for information retrieval, plagiarism detection, and categorization. Their poll, on the other hand, ignores the sentiment analysis paradigm. Finally, they came to the conclusion that a rigorous survey focusing on sentiment analysis is required.

In [42], authors discussed sentiment analysis of Urdu language by evaluating around 14 works of SA. They divided all Urdu SA techniques into different types: lexicon-based, machine learning, and hybrid techniques. Some of the key strategies based on hybrid and ensemble methods may have been overlooked. A survey on sentiment analysis of multi-language was undertaken in [38]. They also discuss various obstacles and make suggestions for future research possibilities.

Word segmentation is the process of identifying the boundaries between individual words [39]. Because a space does not indicate a boundary in Urdu, it is essential to be aware of the borders between words. The morphological analyzer is located inside the pre-processing module, which is responsible for word segmentation. In Urdu text processing, it is used to indicate the borders of individual words.

According to reports [7], the Urdu alphabets are divided into connectors and non-connectors. In a single word, a space can be introduced, for example, " (خوب صورت , beautiful). In contrast, space between two separate words might be eliminated, for example, "" (عالمگیر, universal). Word segmentation in Urdu is connected with the following two issues: There are two types of space insertion, first is space insertion and second is space omission.

During the text cleaning phase, HTML tags, URLs, and other special characters are removed from the input so that it may be processed further in the sentiment analysis module [12]. Word boundary detection and diacritic omission are now part of the text cleaning process in Urdu because of the uncertainty in word boundaries and the optional usage of diacritics. During text normalization, for example, it is usual practice to remove them. Punctuation marks and other special symbols are often removed from Urdu text before sentiment analysis is performed.

Normalization, Diacritic omission and word boundary recognition were all worked on by many researchers. In terms of the diacritic omission, they stated that the Urdu script, like Arabic, Turkish, Persian and Punjabi is made up of letters and diacritics. Because space does not always represent the word boundary, they utilized a punctuation mark as a word divider. In most cases, however, space denotes the word or morpheme boundary, which can still be utilized to identify word boundaries. Furthermore, word affix merger is still required.

The pre-processing of the given text is where sentiment analysis begins, according to [40]. Normalization, tokenization, and word segmentation are all part of this process. Urdu employs a context-sensitive script that separates tokenization and word boundary recognition. Parts of speech tags, such as verbs, nouns, adjectives, and adverbs, are frequently applied to pre-processed words. Then, using phrase chunking, these tagged words are transformed into phrases, yielding verb phrases, noun phrases and adverb phrases, among other things.

In [8], authors discussed numerous stages of Urdu lexicon development from the Urdu text data. Other issues considered included the use of optional vocalic material, Unicode variations for name recognition and spelling variance, and more. An Urdu lexicon is generated by taking into consideration the unique qualities of the POS tags, lemmas, and phonemes found in the corpus. The vocabulary built by the researchers does not cover all of the domains in the constructed corpus, which is a major flaw in their work.

A gold-standard corpus that is machine-readable of user evaluations is a required component of all SA apps. Because of the scarcity of materials in the Urdu language, there is no corpus of Urdu language [40]. According to previous research, the three most common corpus development strategies are manual and bi-lingual. This section contains the most up-to-date research on corpus building for the Urdu text SA process.

An important part of the design of sentiment analysis algorithms are the words and phrases that make up the "sentiment lexicon" [42]. A sentiment class and score are given to each word, sentence, and document to facilitate in the calculation of the score at different levels, such as word, sentence and [21]. Among the

methods for creating sentiment lexicons are "manual annotation", "corpus-based" and "bootstrapping".

The process of assigning a sentiment class and a score to each word in a review is called Sentiment orientation. By recognizing and separating sentiment information accessible in the text, [7] created a sentiment analysis method for Urdu language. The sentiment lexicon and sentiment classification are the two fundamental components of the system. In the realm of movies, 72 percent accuracy was reached, while in the category of products, 78 percent accuracy was achieved.

In their study of Urdu sentiment analysis, [43] collected 151 Urdu blogs from 14 distinct genres. NB, PART, decision tree and K closest neighbor were also used as supervised machine learning classifiers (KNN, IBK).

KNN, SVM, and Decision tree were evaluated for their ability to classify sentiment, and the results were compared in [44]. Based on the findings, it seems that KNN is superior than the other algorithms. When working with bigger data sets, the system has to be evaluated using a variety of different statistical approaches, such as the Root Mean Squared Error, amongst others. In [31], authors developed a LSTM for sentiment analysis in Urdu text for sentiment analysis. Their model solved the gradient attenuation problem and was able to acquire data across lengthy time intervals.

From the above discussion it has been concluded that, the processing of natural language in Urdu is fraught with difficulties. For starters, there is no established word bank for Urdu that provides a word's polarity and POS tag. Second, parsing the Urdu language is a difficult undertaking in and of itself, and the current state of the art is not mature. Finally, there is no publicly available sentiment corpus for the Urdu language that can be utilized to train supervised learning systems. Almost all algorithms for sentiment analysis in Urdu language use the simple BOW approach, which involves using a dictionary of adjectives, and valence shifters that contain words, their definitions, and their valence shifters. These strategies are insufficient for complex and confusing ideas, particularly in the case of Urdu, which lacks a wealth of resources [12] Discourse information improved sentiment analysis in the literature.

3. Proposed Methodology

The proposed conceptual model is divided in to two major phases: (i) pre-processing and feature vector formation; (ii) classification and sentiment analysis. The detailed description along with diagrammatical representation and algorithmic formation has been provided to clearly discuss each phase. The dataset that has been designed for the evaluation of proposed work has also discussed with detail. Fig. 1 shows the proposed Model.

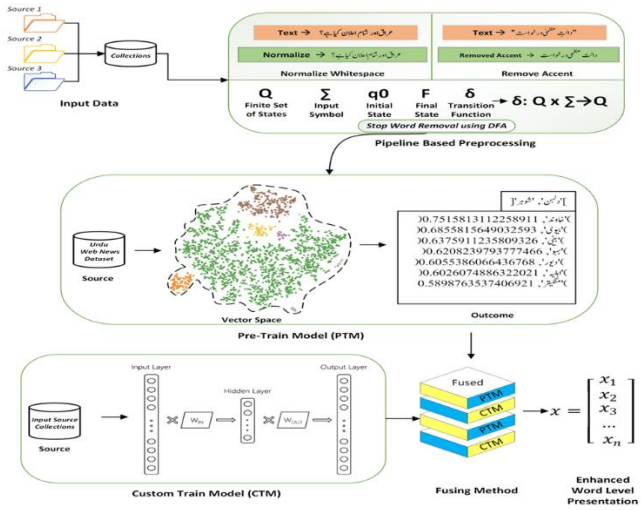


Fig. 1. Proposed conceptual model for feature set generation

3.1 Dataset Preparation

Deep learning algorithms have recently been used to study text representations and solve the challenge of sentiment categorization on huge social network datasets. Because of its superior performance in the field of sentiment analysis, improved word vectors (IWVs) datasets were also suggested for use in the process of word embedding. A limited number of research on the sentiment analysis of social network datasets to assist intelligent transportation systems have been carried out. The information was gathered from a number of different social media websites, such as Facebook.

3.2 Review Collection from Social Media Forums

User reviews are a collection of information about a variety of services, such as products, games, and politics. Because Urdu is a language with limited resources, the authors opted to compile a standard Urdu language text corpus by gathering information on various genres from widely accessible internet archives. Politics, movies, Urdu dramas, TV discussion programs, and sports are all topics covered in customer reviews. Manual data collection was carried out by four people. It took three months to acquire the raw data because they were native Urdu speakers. The data was initially collected in an Excel spreadsheet. From Table 1, Dataset Size: shows Total reviews in the dataset Review Nature: shows whether the Review was Positive or Negative. These are already labelled in the Dataset as positive or

negative. Vocabulary Size shows how many total Sentences were present in the all reviews. Tokens show total number of words in the overall Vocabulary Size. Avg. Tokens per Review shows average number of tokens in each review.

Table 1

Dataset with Descriptive Statistics

Name	Dataset Size	Review Nature	Review Description		
			Vocabulary Size	Tokens	Avg. Tokens per Review
1 [DS1]	7,623 Reviews (Product Based)	3,400 +Ve Review 4,223 –Ve Review	29,600	124,457	114.29
2 [DS2]	5,400 Reviews (Restaurant)	3,400 +Ve Review 2,000 –Ve Review	22,112	88,435	99.23
3 [DS3]	3,300 Reviews (Miscellaneous)	1,730 +Ve Review 1,570 –Ve Review	18,409	54,172	48.31

3.3 Review Formation and the Categorization

This section discusses the DS1, DS2, and DS3 datasets, as well as the annotation process used in manual corpus production for DS3. This process entails developing annotation rules or guidelines. To begin with, we create a set of rules for sentiment analysis based on current literature reviews in order to properly identify the type of review.

A statement is categorized as "positive" if it shows an overall favourable mood, both positive and neutral expressions, or agreement approval [12]. Sentences containing phrases like "congrats" and "admiration" were also considered good. A comment is considered to be negative if it either expresses a negative attitude in its whole or includes a greater number of negative phrases compared to other attitudes. A sentence is classified as "negative" if there is any disagreement in it. In terms of the manual annotation technique, three human professionals (X, Y, and Z) manually annotated the suggested DS3 Urdu dataset to establish a benchmark dataset. The user reviews were all annotated by native Urdu speakers with master's degrees in the language.

Z addressed the issue between X and Y by labelling the review. As indicated in Table 1, the DS1 contains 7,623 reviews about various products, 3,400 of which are positive and the rest are negative. Our corpus is definitely class balanced, as evidenced by the statistics in Table 1. The DS2 dataset, on the other hand, has 5,400 restaurant reviews, 34,400 of which are positive and 2,000 of which are negative. In the available literature, only a few scholars have attempted to construct such datasets for conducting experiments.

3.4 Pre-processing and Enhanced Word Level Presentation

Preprocessing Urdu text is necessary in order to make it understandable and useful for natural language processing (NLP) jobs. In order to make our model more accurate, we got rid of emojis, URLs, email addresses, phone numbers, numerical numbers, numerical digits, currency symbols, and punctuation marks. In addition to this, we carried out the text pre-treatment techniques listed below in order to increase the accuracy of our model when applied to Urdu text.

3.5 Stop Words Elimination

Using Deterministic Finite Automata, we used an algorithm to improve the stop word elimination procedure from Urdu text in this research study [12]. All stop words were separated into a stop-list and presented in a finite state machine. The DFA was then implemented in RAM as a state table with 38 columns displaying the alphabets of Fig. 2 and a variable number of rows (states). Because a content word might become a stop word in Urdu due to its frequent appearance in a document, the number of rows changes. Algorithm 3.1 contains the stop word deletion algorithm.

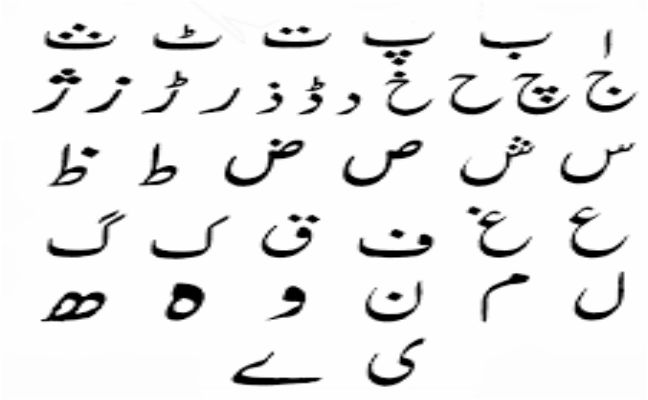


Fig. 2. Urdu alphabets

The first word of the Urdu text was selected in Algorithm 1, and it was examined for no-Urdu or special characters, which were then eliminated. The words less than or equal to 3 are likewise deleted in the next phase. Set count and DFA's starting state to 1 and check for the preset condition: if count exceeds word length and the new state is the final state in the state table, recommend the current word as a stop word.

Algorithm 1

The stop word removal process

Input: Urdu Text (UT)

Output: A list of stop word SWi of the given text.

For each Word Xi in UT

If (Xi == NUR || Xi ==SC)

Then

XW ← Xi

WL ++

If (Xi.L ≤ 3)

SW← Xi

go to the next Xi & go to previous if

Else

Cunt←1 & Qi←1

While (Cunt <WL && Qi >0)

Qi = R.Qi ∩ Cunt. C)

Cunt ++

End For

If (Cunt >WL && Qi ==QF)

Then

SW← Xi

go to the next Xi & go to previous if

END

Fig. 3 shows a created DFA after the suggested technique has been deployed to determine if it accepts a word as a stop word or not.

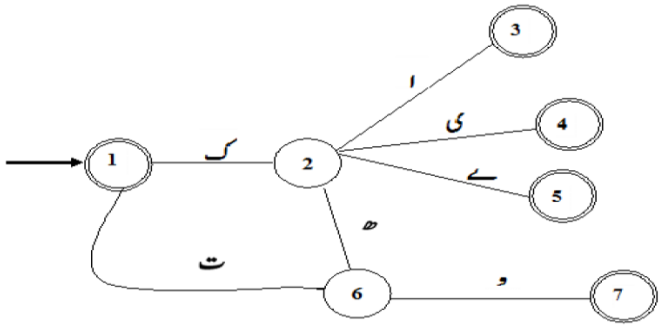


Fig. 3. Proposed DFA

In Fig. 4, columns display various Urdu alphabetical letters, while rows display state numbers. The states marked with an asterisk (*) are the final states. The value in the table represents the current word's next state in the deterministic finite automaton. Alongside is a translated state table that represents the DFA. We are now working on implementing the proposed algorithm and are awaiting results. However, because Arabic and Urdu have similar scripts, we are optimistic that our proposed algorithm will work well.

	ت	ھ	ک	و	ے	ی	ا
1	6		2				
2		6	1		5	4	3
3*							2
4*						2	
5*					2		
6	1	2		7			
7*				6			

Fig. 4. Constructed sample states

With the help of the following examples, we will now elaborate the sample of our DFA:

احمد اس کا دوست تھا

Only those stop words beginning with will be accepted when we run this phrase on this sample ک and ت and ending with ا ،ی ،ے and و .So it will accept کا and تھا and remaining words are separated as content words.

گھڑی تو اس کے ابو کی کھو گئی تھی۔

Now the accepted stop words will be کی، کھو، تو and تھی. Other words will be categorized as content words.

This DFA only accepts the stop words and it can be applied to other similar context languages for further enhancement of work.

3.6 Normalization

Text normalization refers to the process of transforming previously non-canonical text into a single form that is considered to be canonical. The ability to separate concerns is made possible by normalizing text before storing or processing it. This ensures that the input is consistent before any actions are taken on it. For text normalization to be successful, one must have a thorough comprehension of both the kind of text that has to be normalized and the subsequent use that will be made of it. In the vast majority of cases, this can be accomplished by using appropriate Urdu character encoding. For Urdu text, normalization is employed to extract all of the characters in the required Unicode range (0600–06FF). There is always a character-based Unicode assigned to each symbol. Some of the pre-defined rules for normalizing the required Urdu text are listed in this section:

- Urdu text has roots in a variety of languages, including Turkish, Arabic, and Persian, which may aid in the conversion of the required text into standard Urdu. To gain a better understanding of what the preceding explanation entails.

>>> all_fes = ['ف', 'ف', 'ﻓ', 'ﻓ',]

>>> urdu_fe = 'ف'

Even if they come from a variety of languages and each have their own unique Unicode, each and every character in each and every fe is the same. Due to the fact that computers work with numbers, the same character that appears in several locations and different languages will have a different Unicode. This may lead to confusion and makes it more difficult to understand the context of the data. This issue will be resolved after the character module is applied since it will replace all of the characters in all fes with Urdu Fes.

The normalize function based on urduhack is defined as:

>>> import normalize characters from urduhack.normalization

>>> # Characters from the Arabic Unicode block are used in the text.

>>> text = "مجھ کو جو توڑا گیا تھا"

>>> normalized_text = normalize_characters(text)

>>> # Arabic characters have been replaced with Urdu characters in the normalized text.

>>> normalized_text

مجھ کو جو توڑا گیا تھا

3.7 Attention-Based BILSTM Model

In this section, the architecture of Attention-based BiLSTM neural networks, also known as ABBiLSTM, is presented for the purpose of classifying the sentiments of texts written in Urdu. The structure of the model that has been suggested includes three layers: a word encoder, an attention layer, and a softmax layer. The specifics of the model's many components are outlined in the following paragraphs.

Word Encoder: In this work, we solely look into sentiment classification at the sentence level for each Urdu text. A sentence is supposed to have n words $[W_1, W_2, \ldots, W_n]$, with $W_k$denoting the kth word and n denoting the length of the phrase. To begin, each word is embedded in a d-dimensional vector, a process known as word embedding [37]. After all of the word vectors have been stacked, an embedding matrix M, $N\ X\ D$ is created, where d signifies the embedding size and n the sentence length. Word embedding can be regarded of as neural network parameters or as pre-trained from a corpus using a language model. The embedding matrix is then sent into the BiLSTM networks, which encode the phrase.

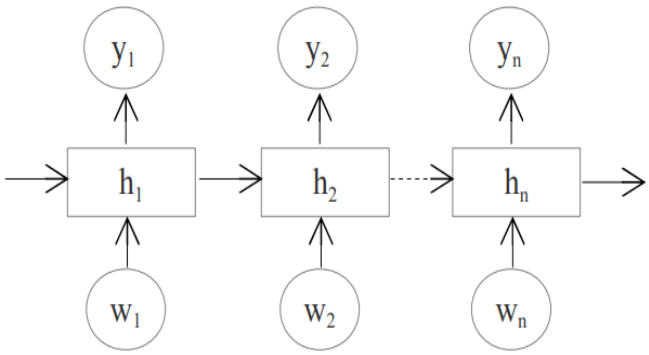


Fig. 5. The architecture of LSTM network

LSTM: The mapping of a variable-length vector of words to a fixed-length vector is one of the key roles that recurrent neural networks play in the process of modelling sequential data. However, because to difficulties with the gradients. This is because to these problems. LSTM networks, which are built on RNNs and were developed based on those networks, have been proposed and developed to solve those limitations. A LSTM has a fundamental architecture that is comprised of three gates and a cell memory state. Fig. 5 presents a detailed illustration of the specific schematic.

Model Training: The cross-entropy error of sentiment classification is applied here as a loss function:

$$L = -\sum_i y_i \log \hat{y_i}$$

Where yi is the sentiment label considered to be the gold standard and I is the label that was predicted by the model.

In order to calculate the derivative of the loss function, the back-propagation method [15] is used on the complete set of parameters. This is done so that the derivative may be determined. After this, the stochastic gradient descent technique is used so that all of the model parameters may be updated.

Evaluation Measures: In a broader sense, the success of sentiment categorization is assessed by the use of four indices. Accuracy, Precision, and Recall in addition to the F1-score are the four categories [3]. The confusion matrix, which is shown in Table 2, provides the basis for the standard method of obtaining these indexes. The specifics of each phrase are described in the following paragraphs:

Table 2

Confusion Matrix

Instance Actual Class	Predicted positives	Predicted negatives
positive	Number of True Positive instances (TP)	Number of False Negative instances (FN)
negative	Number of False Positive instances (FP)	Number of True Negative instances (TN)

Accuracy: is the ratio between the number of correct predictions and a total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision: is the ratio between TPs combined to a number of TPs and FPs.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: is defined as the ratio between TPs combined to a number of TPs and FNs.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score: It takes both false negative and false positive into consideration, and it is the harmonic mean of recall and precision.

$$\text{F1} - \text{score} = 2 * \frac{(precision * recall)}{(precision + recall)}$$

4. Experiment Description

In this section, the performance of the proposed model has been compared with the existing state-of-the-art methods. The detailed description of the baseline model has been provided below with abbreviation as Model A, and Model B.

Model A: Khan et al. 2021 which is abbreviated as Model A, is a rule based machine learning model for Urdu sentiment analysis using support vector machine, Naïve Bayesian, Adabbost, MLP, LR and RF along with deep learning model using CNN-1D, LSTM, Bi-LSTM, GRU and Bi-GRU techniques [9].

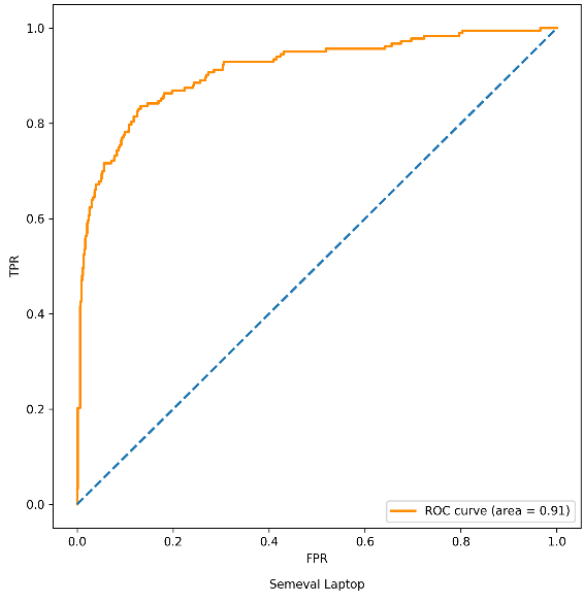
Model B: is the representation of Umair et al., [10], that classically execute different machine learning based classification models such as support vector machine and naïve Bayesian algorithm on Roman Urdu text to test its accuracy.

The performance of the suggested work was tested in the very first experiment by manually annotating the results. In this experiment, the suggested research in terms of True Positive Rate (TPR), False Positive Rate

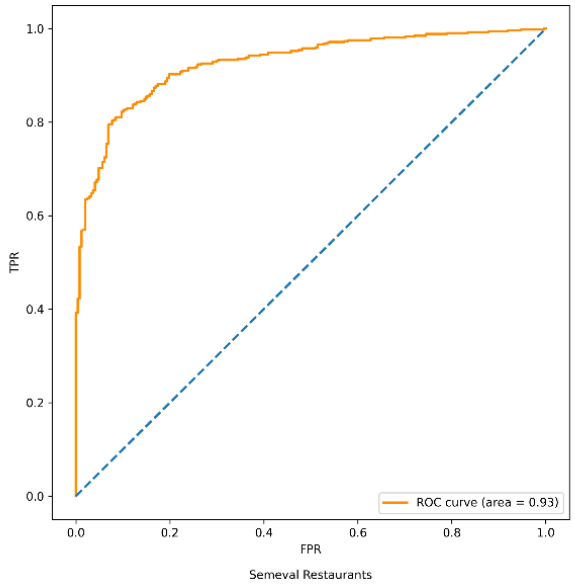
(FPR), and actual and anticipated findings has been presented. TPR stands for true positive rate, while FPR stands for false positive rate. The true positive rate (TPR) and the false positive rate (FPR) have each been shown on their own separate ROC curves, which have been generated for each dataset.

4.1 Result Outcomes

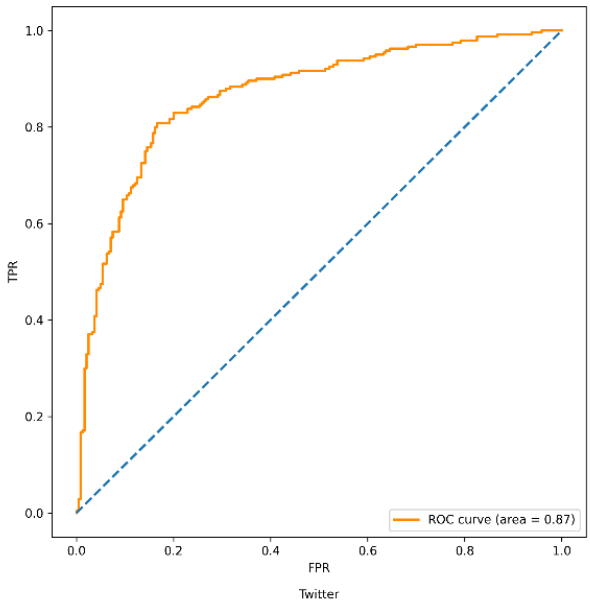
For datasets, Fig. 6 (A), ROC curve cover 0.91 area, Fig. 6 (B) covers 0.93, and Fig. 6 (C) 6c covers 0.87. From this depiction, it has been shown that out of four datasets three of the curves cover more than 90% area and one covers more than 87% area that validate more accurateness. Alternatively, a confusion matrix has also been designed to show the performance of the proposed technique in terms of TP, TN, FP, and FN.



[DS 1] (A)



[DS2] (B)



[DS3] (C)

Fig. 6. ROC of the proposed model on three different datasets

Table 3 shows the results of the experiments performed by comparing the performance of the proposed work with model A on dataset 1, 2 and 3. The results are computed on the basis of precision, recall and F-score separately for positive and negative sentiment analysis.

Further, different the computation results of different evaluation measures are presented in Table 3. Results show that the proposed model achieved maximum 97.00% accuracy. In Fig. 7, the proposed model is compared with Model A. Results clearly depicts the clear supremacy of the proposed model against the existing model. In another attempt, the performance of the proposed technique has been compared with the work of Model B. the graph, Fig. 4 portrays the overall performance of both Model A and proposed model on dataset [DS2] and [DS3].

Table 3

Comparison of the proposed work with Model A

Dataset	Precision	Recall	F-Measure	Accuracy	Orientation
			Result obtained by model A		
[DS 1]	86.00%	65.00%	75.50%	87.00%	Positive
	76.00%	66.00%	71.00%	87.00%	Negative
[DS 3]	94.00%	63.00%	78.50%	94.00%	Positive
	74.00%	66.00%	70.00%	86.00%	Negative
			Result obtained by proposed model		
[DS 1]	95.00%	65.00%	80.00%	97.00%	Positive
	95.00%	66.00%	80.50%	97.00%	Negative
[DS 3]	94.00%	63.00%	78.50%	94.00%	Positive
	94.00%	66.00%	80.00%	94.00%	Negative

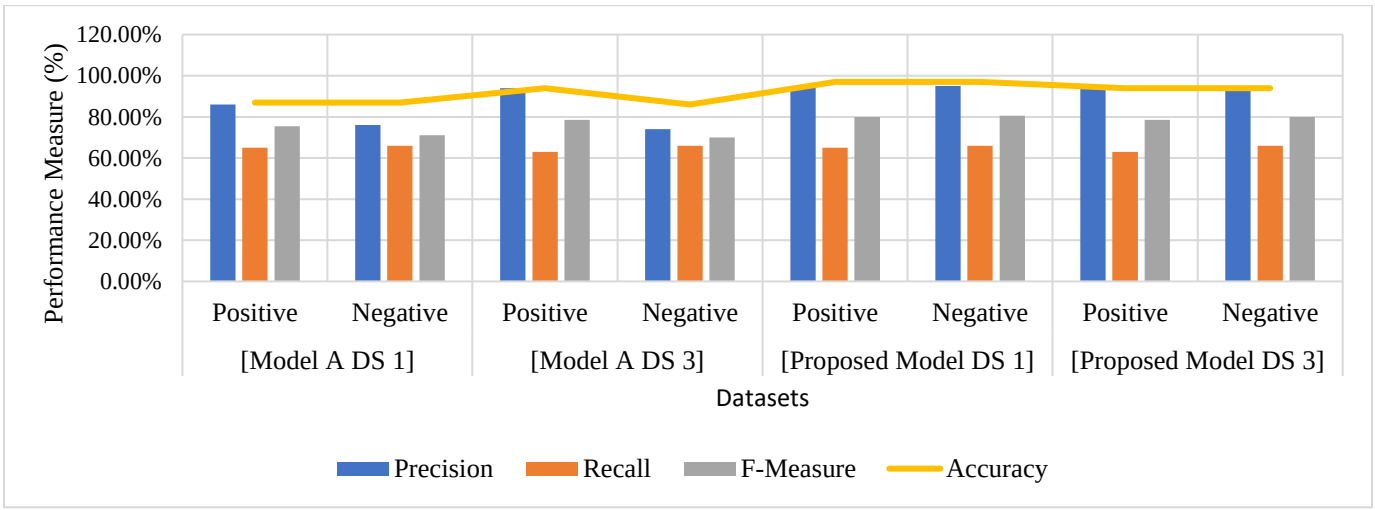


Fig. 7. Comparison of proposed model with model A

Table 4

Comparison of the proposed work with Model B

Dataset	Precision	Recall	F-Measure	Accuracy	Orientation
Result obtained by Model B					
[DS 1]	89.00%	65.00%	77.00%	79.00%	Positive
	79.00%	66.00%	72.50%	75.00%	Negative
[DS 3]	75.00%	63.00%	69.00%	79.00%	Positive
	81.00%	66.00%	73.50%	82.00%	Negative
Result obtained by Proposed Model					
[DS 1]	91.00%	65.00%	78.00%	91.00%	Positive
	71.00%	66.00%	68.50%	89.00%	Negative
[DS 3]	85.00%	63.00%	74.00%	85.00%	Positive
	85.00%	66.00%	75.50%	85.00%	Negative

4. Conclusion

Within the scope of this research, automated sentiment analysis of Urdu was examined. Urdu is a language that is both morphologically rich and resource poor. This strategy, which is inspired by grammar, is appropriate for dealing with the intricate morphology and changeable vocabulary of the language you are trying to learn. We have implemented the fundamental tasks of natural language processing, including phrase chunking, POS tagging, and word segmentation.

As a result of our preliminary research into the application of sentiment analysis to the Urdu language, we have arrived at a variety of hypotheses on the distinctive qualities of this language and the difficulties it presents for the use of computer processing. For instance, Urdu is sensitive to context, and as a result, the segmentation of its words is a significant challenge in and of itself. Word boundary identification in this language is not as straightforward as it is in the English language because of this trait.

When compared to other NLP lexicons, it turns out that a sentiment-annotated lexicon is much more complicated. This complexity may be attributed to two different factors:

- In addition to its orthographic, phonological, syntactic, and morphological properties, each entry in the lexicon provides information on the polarity of the word in question.
- This polarity information is often stated as either positive, negative, or neutral. The majority of words display several orientations depending on their usage and the domain in which they are found. Take, for instance, the phrase "This harm is forever."

5. References

[1] A. Khan, “Improved multi-lingual sentiment analysis and recognition using deep learning”, Journal of Information Science, 2023.

[2] I. U. Khan, A. Khan, Khan, W. Su’ud, M. M., Alam, F. Subhan, and M. Z. Asghar, “A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language”, Computers, 11(1), 3.

[3] L. Khan, A. Amjad, K. M. Afaq and H. T. Chang, H. T., “Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media, Applied Sciences, 12(5), 2694, 2022.

[4] R. M. Abdalla and S. Teufel. “A bootstrapping approach to unsupervised detection of cue phrase variants”, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for

Computational Linguistics, pp. 921-928, 2006.

[5] M. Abdullah and M. Hadzikadic, “Sentiment analysis on arabic tweets: Challenges to dissecting the language”, International Conference on Social Computing and Social Media, pp. 191-202. Springer, Cham, 2017.

[6] M. Abid, A. Habib, J. Ashraf, and A. Shahid, “Urdu word sense disambiguation using machine learning approach”, Cluster Computing, 2018, 21(1), 515-522.

[7] SZ. Afraz, M. Aslam, R. Jan , T. Saba, W. Mirza, “Sentiment Analysis of a Morphologically Rich Language”, vol.2 (2), pp.69-73, 2010.

[8] M. Ijaz, and S. Hussain, “Corpus based Urdu lexicon development”, Proceedings of Conference on Language Technology, University of Peshawar, Pakistan, vol. 73, 2007.

[9] L. Khan, A. Amjad, N. Ashraf, and H. T. Chang, “Multi-class sentiment analysis of urdu text using multilingual BERT”, Scientific Reports, 12(1), 5436, 2022.

[10] M. Umair, Z. Saeed, M. Ahmad, H. Amir, B. Akmal, and N. Ahmad. “Multi-class classification of Bi-lingual SMS using Naive Bayes Algorithm, IEEE 23rd International Multitopic Conference, pp. 1-5, 2022.

[11] M. Z. Ali, S. Rauf, K. Javed, and S. Hussain, “Improving hate speech detection of urdu tweets using sentiment analysis, IEEE Access, 9, 84296-84305, 2021.

[12] W. Anwar, X. Wang, XL. Li and A. Wang, A statistical based part of speech tagger for Urdu language. In Machine Learning and Cybernetics, International Conference on 2007 Aug 19 (Vol. 6, pp. 3418-3424). IEEE.

[13] U. Naqvi, A. Majid, S. A. Abbas, UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods. IEEE Access 2021, 9, 114085–114094.

[14] A. Khattak, M. Z. Asghar, A. Saeed, I. A. Hameed, S. A. Hassan and S. Ahmad. A survey on sentiment analysis in Urdu: A resource-poor language. Egypt. Inform. J. 2021, 22, 53–74.

[15] M. Z. Asghar, S. Ahmad, M. Qasim, S. R. Zahra, and F. M. Kundi. SentiHealth: creating health-related sentiment lexicon using hybrid approach. SpringerPlus, 2016, 5(1), 1-23.

[16] N. V. Babu, and E. Kanaga. Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. SN Computer Science, 2016, 3(1), 1-20.

[17] R. Batra, Z. Kastrati, A. S. Imran, S. M. Daudpota, A. Ghafoor. A large-scale tweet dataset for urdu text sentiment analysis, 2021.

[18] M. A. Qureshi, M. Asif, M. F. Hassan, G. Mustafa, and M. K. Ehsan, A novel auto-annotation technique for aspect level sentiment analysiss,'CMC-Comput., Mater. Continua, 2022, vol. 70, no. 3, pp. 4987-5004,.

[19] F. Noor, M. Bakhtyar, and J. Baber, Sentiment analysis in e-commerce using SVM on roman Urdu text,' in Proc. Int. Conf. Emerg. Technol. Comput., 2019, pp. 213-222.

[20] M. Bilal, H. Israr, M. Shahid and A. Khan. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences, 2016, 28(3), 330-344.

[21] M. Birjali, A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 2021, 226, 107134.

[22] G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica and J. Hemanth, Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. Applied Sciences, 2020, 12(3), 1030.

[23] P. Chauhan, N. Sharma and G. Sikka,, The emergence of social media data and sentiment analysis in election prediction. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(2), 2601-2627.

[24] M. Du, X. Li, and L., Luo. (2021). A training-optimization-based method for constructing domain-specific sentiment lexicon. Complexity, 1-11, 2021.

[24] M. Du, X. Li snd L. Luo. A Training-Optimization-Based Method for Constructing Domain-Specific Sentiment Lexicon. Complexity, 2021.

[25] M. Bilal, H. Israr, M. Shahid, and A. Khan, `Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, decision tree and KNN classifcation techniques, J. King Saud

Univ. Comput. Inf. Sci., 2016 vol. 28, no. 3, pp. 330-344,

[26] Z. Sharf, D. Saif, and U. Rahman, "Performing natural language processing on Roman Urdu datasets", International Journal of Computing Science Networks Secure, vol. 18, no. 1, pp. 141-148, 2018.

[27] N. Durani and S. Hussain, Urdu Word Segmentation, Human Language Technologies. In The Annual Conference of the North American Chapter of the ACL, Los Angeles, California, 2010. (pp. 528-536).

[28] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, 'Discriminative feature spamming technique for Roman Urdu sentiment analysis, IEEE Access, 2019, vol. 7, pp. 47991–48002,

[29] K. Mehmood, D. Essam, and K. Shafi, 'Sentiment analysis system for Roman Urdu, in Proc. Sci. Inf. Conf., 2019, pp. 29–42.

[30] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis, Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102368,

[31] H. Ghulam, F. Zeng, W. Li and Y. Xiao. Deep learning-based sentiment analysis for Roman Urdu text. Procedia computer science, 2019, 147, 131-135.

[32] M. S. Hossain, N. Nayla and A. Rassel,. Product Market Demand Analysis Using NLP in Banglish Text with Sentiment Analysis and Named Entity Recognition. In 2022 56th Annual Conference on Information Sciences and Systems (CISS), 2022, (pp. 166-171). IEEE.

[33] Q. Huang, R. Chen, X. Zheng and Z.. Deep Sentiment Representation Based on CNN and LSTM. In: 2017 International Conference on Green Informatics (ICGI), 2017 (pp. 30-33). IEEE.

[34] M. Ijaz, and S. Hussain, "Corpus based Urdu lexicon development", Proceedings of Conference on Language Technology, University of Peshawar, Pakistan, vol. 73, 2007.

[35] M. M. Agüero-Torales. Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 2021,107, 107373.

[36] M. S. Alharbi, "Optimize machine learning programming algorithms for sentiment analysis in social media", International Journal of Computer Applications, 2021174(25), 38-43.

[37] A. R. Ali and M. Ijaz, "Urdu text classification", Proceedings of the 7th international conference on frontiers of information technology, pp. 1-7, 2009.

[38] S. L. Lo. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artificial Intelligence Review, 48, 499-527, 2017.

[39] M. Daud, R. Khan and A. Daud, "Roman Urdu opinion mining system", arXiv preprint arXiv:1501.01386, 2015.

[40] Q. Rajput. Ontology based semantic annotation of Urdu language web documents. Procedia Computer Science, 2014, 35, 662-670.

[41] Z. Nasim, Q. Rajput and S. Haider. Sentiment analysis of student feedback using machine learning and lexicon based approaches. In 2017 international conference on research and innovation in information systems, IEEE, pp. 1-6, 2017.

[42] M. Du, X. Li and L. Luo., "A training-optimization-based method for constructing domain-specific sentiment lexicon", Complexity, 2021, 1-11.

[43] N. Mukhtar, M. A. Khan., "Urdu sentiment analysis using supervised machine learning approach", International Journal of Pattern Recognition and Artificial Intelligence, 32(02), 1851001, 2018.

[44] N. Mukhtar, M. A. Khan and N. Chiragh. "Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis", Cognitive Computation, 9, 446-456, 2017.

[45] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data", IEEE World AI IoT Congress, 2021, pp. 0285-0291, 2021.

