IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 December 2021, accepted 24 January 2022, date of publication 15 February 2022, date of current version 9 March 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150172

Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study

MUHAMMAD AASIM QURESHI[1], MUHAMMAD ASIF[1,2], MOHD FADZIL HASSAN[3], ADNAN ABID[4], ASAD KAMAL[1], SOHAIL SAFDAR[5], AND REHAN AKBAR[3], (Senior Member, IEEE)

[1]Department of Computer Sciences, Bahria University Lahore Campus, Lahore 54000, Pakistan
[2]Department of Law, Science and Technology, University of Bologna, 40126 Bologna, Italy
[3]Computer and Information Science Department, University Teknologi PETRONAS, Perak 32610, Malaysia
[4]School of System and Technology, University of Management and Technology, Lahore 54000, Pakistan
[5]College of Information Technology, Ahlia University, Manama 10878, Bahrain

Corresponding authors: Muhammad Aasim Qureshi (maasimq@hotmail.com), Muhammad Asif (asifhashmat255@gmail.com), and Mohd Fadzil Hassan (mfadzil_hassan@petronas.com.my)

ABSTRACT Opinion Mining from user reviews is an emerging field. Sentiment Analysis of Natural Language text helps us in finding the opinion of the customers. These reviews can be in any language e.g. English, Chinese, Arabic, Japanese, Urdu, and Hindi. This research presents a model to classify the polarity of the review(s) in Roman Urdu text (reviews). For the purpose, raw data was scraped from the reviews of 20 songs from Indo-Pak Music Industry. In this research a new dataset of 24000 reviews of Roman Urdu text is created. Nine Machine Learning algorithms— Naïve Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Artificial Neural Networks, Convolutional Neural Network, Recurrent Neural Networks, ID3 and Gradient Boost Tree, are attempted. Logistic Regression outperformed the rest, based on testing and cross validation accuracies that are 92.25% and 91.47% respectively.

INDEX TERMS Sentiment analysis, sentiment classification, Roman Urdu, supervised learning, song reviews, Roman Urdu corpus, machine learning, Naïve Bayes, decision tree, K-NN, deep learning, ANN, CNN, RNN, text classification.

I. INTRODUCTION

Though there exist some companies that collect reviews about any entity through market surveys, it‘s an outdated way to acquire feedbacks [1]. The boom of technology has squeezed the distances and digitized the world. Products are being shelved online. In order to know the response (on different aspects) of the product, the feedback mechanism is provided in the form of comments or reviews [2]. This feedback, i.e. Reviews, is valuable for both—user and the displayer [3]. Such reviews had replaced the old style of surveys that were critical to revamp quality standards. The drastic increase in the number of e-users has caused the exponential growth in these reviews [4].

Recently, the showbiz of South Asia has evolved enormously [5] in which, internet technology has played a vital role [6]. Using this paradigm people have easy access to songs, movies, plays and many more. Interestingly, the artists have direct connections with their fans through social media. Viewers not only watch the material but also give their feedbacks on the same page or channel [7]. Different websites manage such material, YouTube is one of them [8].

YouTube is one of the most popular platforms [9] for hosting videos of different kinds like entertainment, education, sports, etc. People are using YouTube, professionally, to earn money by creating their channels and posting their videos in them. YouTuber [10] is a well-known and successful profession nowadays. YouTubers make their videos and post on their pages/channels to engage the audience. These professionals, judge the popularity of the content [6] by the number of likes, dislikes and comments on the video [9]. A simple formula to check the content quality is:

$$CQ = (TL - TD > 0?\ \text{“good”} : \text{Bad”}) \tag{1}$$

where TL is Total likes and TD is Total Dislikes.

This formula provides very limited insight into the content.

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong.

Another way to check the credibility of the content of a video is by analyzing the comments on that video [11]. Like many other websites, YouTube also facilitates writing reviews in the comments section [12]. Previously, it was easy to judge the contents by manually reading these reviews as reviews were limited in quantity. But nowadays as the count of these reviews has increased tremendously it is humanly not possible to read and analyze all reviews. It raises the need for automated processes/tools for the purpose.

Sentiment Analysis is one such tool. People posts their opinion about the item—good, bad or neutral, in the given comments section [13]. It categorizes the polarity to the review [14] as positive, negative, or neutral [15].

Though educated people review the item in English, but in general, people prefer to review the item in a language they feel comfortable. Originating from English speaking countries, most of the websites support only roman script [16], for reviews [17]. That is why most of the Sentiment Analysis research work is done on the English language. People from non-English speaking countries, when they visit these websites, write comments in their native languages using Roman script like Chinese, Arabic, Japanese, Urdu, Hindi, etc. Most of the antecedent work is done in the domain of Sentiment Analysis on English and Chinese languages [18]–[20]. Urdu is one of the popular and widely used languages of the Sub-continent. There exist no scripting standards for Roman Urdu [21] to spell Urdu words. People not profound in the English language use Roman script to write their reviews.

Limited research is witnessed on Sentiment Analysis of Roman Urdu text, so there is a need to extend the research in this direction. It can help people to improve their business strategies and to tackle the consumer's needs. For this research, data of Roman Urdu text was required. Being followed by the folks, the music content of South-Asia was opted to scrap the reviews from YouTube. Reviews against the song help to understand different quality aspects and sentiments of people for the content [22].

To perform Roman Urdu Sentiment Analysis, a good dataset was required [23]. In existing research works, the datasets were not big enough. This paper contributes a benchmark dataset of Roman Urdu text extracted from songs reviews (Let's call it DRU). The entire corpus was manually annotated to perform Sentiment Analysis. To build DRU, reviews were scraped from YouTube from selected twenty songs mentioned in table 1. Total 321,504 reviews were scrapped. Urdu Reviews were extracted by applying different filters. DRU contains 24,000 manually annotated reviews.

The rest of the paper is organized into five sections. Section-II presents existing works. Section-III discusses the methodology which is adapted to meet the objectives of this research. Section-IV describes the entire corpus generation process of DRU. Experimental results are presented and discussed in section V. Section-VI concludes the research and presents possible future enhancements.

II. LITERATURE REVIEW

In [17], researchers performed Roman Urdu Sentiment Analysis on a dataset of just 1,600 hotel reviews. The maximum accuracy that they achieved was 91% using the Support Vector Machine. Reference [24] Presented Sentiment Analysis of Roman Urdu reviews on mobiles. The researchers tried Decision Tree, Naïve Bayes and KNN algorithms for this purpose. They achieved 97.50% using Naive Bayes. The researchers of [25] used the N-gram model. Maximum accuracy of 72.37% was achieved through Naïve Bayes using uni-Bigram. Authors of [26], performed Aspect Level Sentiment Analysis using different Machine Learning techniques to classify the products in three different categories (Low, Medium and High).

In [27], efforts were made to perform Sentiment Analysis using discoursed based Sentiment Analysis. Reviews were scraped from different websites providing social services. Analysis was performed using discoursed based Part of speech tagging. In [28], the "Bag of Words Meets Bag of Popcorn" dataset was used to perform Sentiment Analysis and maximum accuracy was achieved 90.90% by using word vector weighted averaging + Random Forest. In [29], the reviews were collected, on mobile, from amazon and aspect level Sentiment Analysis was performed on the dataset. Association rule mining was used for the segmentation of sentences. After the opinion orientation of the words, they simply counted the total numbers of positive and negative comments for each aspect and finally ranked the aspects based upon the numbers of positive reviews. In [30], to perform the aspect level Sentiment Analysis on movie reviews, the 5-Gram technique and the feature-based heuristic scheme is used to perform aspect level Sentiment Analysis.

In [31] Sentiment Analysis was performed using dimensionality reduction after applying different preprocessing techniques like slangs handling, stop-words, stemming, and lemmatization. Dataset—Bag of Words Bag of Popcorns, was taken from Kaggle. Total 256 experiments were conducted. The accuracies of experiments before and after preprocessing techniques were 83.417% and 84.90% respectively. In [32], researchers worked upon finding the hidden pattern from raw text data using preprocessing techniques. To avoid the outliers in data the different preprocessing techniques were implemented i.e., HTML tags, Stop words removal. Standardization of the dataset was done by implemented stemming and lemmatization. The features were extracted by applying N-gram and TF/IDF.

Development of dataset had been attempted, previously, but most of them are of small sizes. As [24], [25] and [33] used dataset of 300, 600 and 1600 Roman Urdu reviews. In English, Chinese, and affluent languages benchmark datasets are available. In [34] dataset of English reviews on Salsa Music is presented. In [35] a dataset of 1000 English reviews is presented and emotional analysis is performed. This creates a need to develop a benchmark dataset on Roman Urdu Reviews to perform Sentiment Analysis on it.

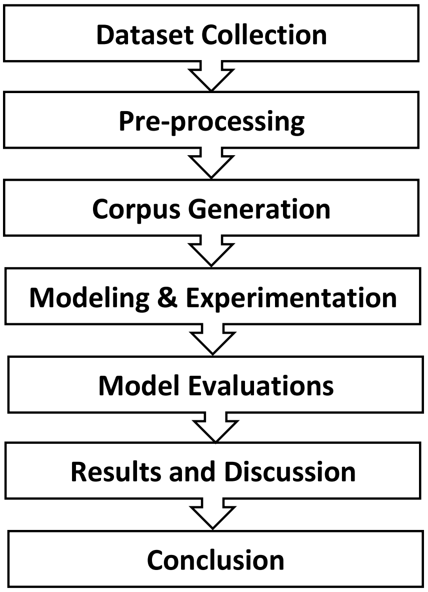


FIGURE 1. Methodology.

III. METHODOLOGY

To generate the corpus i.e. DRU and to perform the Sentiment Analysis on it, following methodology was adopted. The methodology of the paper is shown in figure 1.

A. DATASET COLLECTION

For the dataset, reviews were collected from YouTube. Different songs of Indian and Pakistani singers were selected to collect Roman Urdu reviews. Table 1 enlists the name of the songs and the number of scraped reviews.

After scraping, these reviews were saved in CSV file format. This raw dataset contained reviews in Roman Urdu as well as in other languages'. It also contained a lot of noisy data (like special characters, numbers, slang, emoji's). To clean the data and extract Roman Urdu reviews and make it ready for the analysis different pre-processing techniques were applied.

B. PRE-PROCESSING

To get good analytical results using Machine Learning techniques, data is supposed to be very refined and of high quality [38]. For the purpose different pre-processing techniques [32], [39] were applied to raw dataset to make it ready for a high standard analysis. Raw dataset had different types of issues like noise, special characters, emojis, text in different languages, varying lengths of reviews, etc. The properties of raw data were shown in table 2. To get the required quality of data, different pre-processing techniques were applied like data filtration, data integration, lowercasing, remove emoji's, and length standardization. Details can be seen in subsections.

1) DATA INTEGRATION

The raw data was in twenty different files, one for each song. To perform data analysis altogether, all the data in these files was integrated into one. The combined number of records (i.e. reviews) reached 321,504 and the file size of 107KBs.

TABLE 1. List of songs.

Sr#	Song Name	# of Reviews
1	Angel	16,941
2	Aunty ki Ghanti	3,055
3	Billi	2,534
4	Dil Chori	1,00,167
5	Eye To Eye	346
6	Romantic Medley Three	5,961
7	Kalabaaz Dil	674
8	Kaif O Suroor	185
9	Kandyaari Dhol	1,819
10	King Shah Humanity	1,124
11	Nashe Si Chadh Gayi	52,541
12	Chhalakata Hamro Jawaniya	93,963
13	Hawa Hawa	30,068
14	Saaiyaan Song	12,686
15	Sandal	43,564
16	Sajni String	293
17	Shakar Wandaan	1,008
18	Tera Mukhra Haseen	1,593
19	Trutti Frutti	3,707
20	Three Peg	49,442

TABLE 2. Properties of raw dataset.

Property	Frequency
Number of Characters	1340263
Max Length of a Review (in Characters)	5578
Min Length of a Review (in Characters)	3
Average Length of a Review (in Characters)	55.84
Number of Tokens	253724
Max Length of a Review (in Tokens)	499
Min Length of a Review (in Tokens)	1
Average Length of a Review (in Tokens)	10.57
Unique Number of Tokens	5806

2) NOISE REMOVAL

Noise is a factor that affects the analysis badly. It was observed that collected data had a lot of noise i.e. special characters, punctuations, numbers and emojis. The presence of noise badly impact the quality of classification results [40]. The focus of this research was only on text analyses, so all noise from data was removed. The changes in dataset-DRU, before and after noise removal, can be seen in table 3.

3) LOWERCASING

Raw dataset contained both types of text—uppercase and lowercase. When this type of data is used for classification, classifiers find different variations [41] of the same input class [42]. Being case sensitivity of classifiers, will recognize "mast" and "MAST" as two different inputs. To overcome this problem, the complete dataset is converted into lowercase.

TABLE 3. DRU properties before and after applying noise removal.

Property	Before	After
Number of Characters	1340263	1240168
Max Length of a Review in Characters	5578	2182
Min Length of a Review in Characters	3	3
Average Length of a Review in Characters	55.84	51.67
Number of Tokens	253724	248786
Max Length of a Review in Tokens	499	497
Min Length of a Review in Tokens	1	1
Average Length of a Review in Tokens	10.57	10.37

4) ROMAN URDU TEXT FILTRATION

The raw dataset was having reviews in multiple languages. The focus of this research was only on Roman Urdu, so the Urdu reviews in the Roman script were extracted using data filtration technique. Data Filtration was performed in Microsoft Excel by applying different filters. Filtered data was further filtered during the data annotation where annotators were asked to remove any record if it is carrying non-Urdu text. Table 4 illustrated the sample filters which were used to extract the Urdu Reviews.

5) LIMITING TEXT STRING SIZE

In the dataset some of the reviews were found too large for example there was a review of 5,578 characters. Long reviews cause issues that reduce the performance of the classifiers [43]. To avoid performance issues, and keeping mind the minimum data loss, the maximum review length was set to 150 characters. This check had an impact but was not huge. Only 6.75% of data got cropped. Before applying this check, the dataset had 1,240,168 characters and after the application of this cut, the number of characters was reduced to 1,156,473. The data traits of "DRU", at this moment, are shown in table 5, before and after applying this preprocessing function.

TABLE 4. Sample filters to extract urdu reviews.

Filters		
Gana	Bacha	Lut
Bhai	Pehly	Mujhay
Kamal	Aya	Kaam
Acha	Hoga	Mahfil
Sakta	Wajah	Huye
Nahi	Bahut	Shab
Teri	Kia	Rahe
Iske	Dekh	Mar

TABLE 5. Data traits of limiting string size function.

Property	Before	After
Number of Characters	1240168	1156473
Max Length of a Review in Characters	2182	150
Min Length of a Review in Characters	3	3
Average Length of a Review in Characters	51.67	48.18
Number of Tokens	248786	232439
Max Length of a Review in Tokens	497	38
Min Length of a Review in Tokens	1	1
Average Length of a Review in Tokens	10.37	9.68

IV. CORPUS GENERATION

This section illustrates the measures that were adopted to build a labelled dataset of Roman Urdu Reviews to perform Sentiment Analysis. Steps that were taken to construct the DRU were as follows:

Step 1: Collection of data as discussed in section III.A
Step 2: Preprocessing as discussed in section III.B
Step 3: Data Extraction as discussed in section III.B.5.
Step 4: Data Annotation as discussed below:

A. ANNOTATION GUIDELINES

Data annotation is a static component, in which a label (i.e. class) is assigned to each text according to its subjectivity expressed in the text. Annotation can be performed using three schemes i.e. Manual Annotation, Auto-annotation and Semi Auto-annotation. In this study, manual annotation was performed to label the reviews into their targeted class according to the guidelines presented in [44] and [45]. Each review was labelled with one of the two classes—Positive or negative.

A review was marked as positive if the sentiment was positive by the expression [46], [47]. In the case of a review that showed both positive and neutral expression, it was classified as positive [48], [49]. The presence of any agreement of approval made the review positive. Like a review "*kamal ka gana h*" (the song is awesome) in which the word "kamal" was the shining word that defined the polarity of the whole sentiment. Another review "*wow bhut acha sound voice hai*" (wow, the voice is too good) in which "wow" and "bhut acha'' were illocutionary words that classified this comment as positive.

A review that was negative in terms of sentiment or expression was classified as negative [50], [51]. A review that contained a negative word, classified as a negative review. Like a review ''*na hero acha na herion*'' (neither hero nor heroine is decent), was classified as negative. Another review ''*faltu gana mud khrab kr diya mera*'' (useless song, spoiled the mood) in which the word ''faltu'' 'rubbish' made this review a negative. Some sample annotated reviews are illustrated in table 6.

V. MODELING AND EXPERIMENTATION

For the best classification results, experimentation was performed using different classification algorithms. The details of the experimentation setup are as follows:

TABLE 6. Example of manually annotated reviews.

COMMENTS	CLASS
gaana top kaa hai	Positive
kya dance hai kya gaana hai kya mehnat hai	Positive
o maza aa gaya	Positive
yaar gaana sunne k baad mei ude gayii hawa mei	Positive
mast Bhai	Positive
bakbas gana	Negative
tu phir aa gaya bc	Negative
original gaane ki maa behen kr di bhai isne	Negative
sabse bekar hi actor	Negative
behaya orat	Negative

A. EXPERIMENTAL SETUP

As discussed above different classification algorithms and techniques were used to design models. The ML algorithms were applied in Python and results were generated by applying TF/IDF. For the cross-validation of the models was done using K-Fold procedure with k = 10. The experimental design is shown in figure 2.

B. DESIGN MODEL

This research targets the binary type of classification [52]. The Machine Learning algorithms used to design the model were Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), ID3 and Gradient Boost Tree (GB). To classify the reviews and to check the performance of the model, data was split into the ratio of 9:1 i.e. 90% data was used for the training of the model and 10% data was used to test the model.

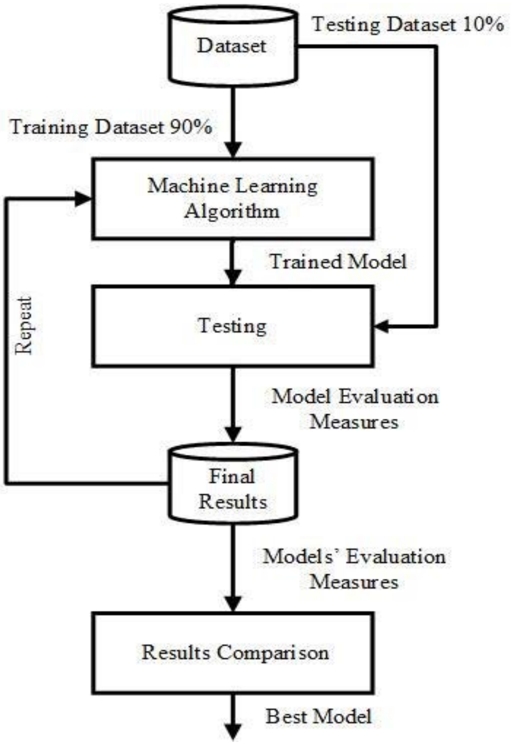


FIGURE 2. Experimental design.

1) Naïve BAYES

Naïve Bayes is a simple statistical-based probabilistic classifier that is based on the ''Bayes'' theorem. It is a fast and efficient classification technique and is famous for text data classification. It, also, can handle both continuous and discrete types of data [53]–[56]. Details of test classification results can be seen in confusion matrix shown in figure 2(a). It achieved test accuracy of 89.67% while cross validation accuracy was 85.37%.

2) SUPPORT VECTOR MACHINE

Support Vector Machine is a linear model for classification. It generalizes between two different classes if labelled dataset is provided for training to the algorithm. SVM is the representation of data as points in space. It can map the data and separate the categories divided by a clear gap that can be as far/wide as possible [57]. Details of test classification results can be seen in confusion matrix shown in figure 2(b). It achieved test accuracy of 88.46% while cross validation accuracy was 86.96%.

3) LOGISTIC REGRESSION

Logistic Regression is the most popular Machine Learning algorithm of machine to perform regression tasks. It is mostly used for forecasting and relationship between variables [58]. Details of test classification results can be seen in confusion matrix shown in figure 2(c). It achieved test accuracy of 92.25% while cross validation accuracy was 91.47%.

4) DECISION TREE

A Decision Tree is a hierarchy-based classification model that uses the divide and conquers approach. If data is discontinuous, it performs well. From the tree family, ID3 and Gradient Boost were opted in this study because these algorithms perform better on text data [59]. It captured overfitting when the data was noisy and if a small variation in data occurs prediction of the model gets unstable [60], [61]. ID3 and Gradient Boost Tree achieved 87.92% and 85.79% accuracies respectively on testing data. Details of test classification results of ID3 and Gradient Boost Tree can be seen in confusion matrix shown in figure 2(d) and 2(e) respectively. The accuracies on 10-Fold cross-validations were 86.31% and 85.52% which showed that models were good-fitted.

5) K-NEAREST NEIGHBOR

The K-nearest-neighbor algorithm is based on instance-based learning. It can simply compare the given test instances with the training set using Manhattan distance, Hamming distance, Minkowski distance and Euclidean distance. It is simple to understand and easy to implement. It is a lazy learner because it memorizes the training data and it does not perform well with missing values [62], [63]. Details of test classification results can be seen in confusion matrix shown in figure 2(f). It achieved test accuracy of 86.13% while cross validation accuracy appeared to be 86.64%. The model is good fitted.

6) ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANNs), generally called neural networks (NNs), are computing systems inspired by the biological neural networks that constitute animal brains. Neural networks learn (or get trained) by processing examples, with well-defined input(s) and result(s). It forms probabilistic weighted associations between the two. These weights are stored within the net itself. Details of test classification results can be seen in confusion matrix shown in figure 2(g). It achieved test accuracy of 90.38% while cross validation accuracy was 88.00%.

7) CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) belongs to the family of neural networks. It can simply take the input, assign the learnable weights to the objects, and classify them [64]. CNN performs well on imagery data [65], CNN did not perform well on DRU. Details of test classification results can be seen in confusion matrix shown in figure 2(h). It achieved test accuracy of 66.54% while cross validation accuracy was 67.19%.

8) RECURRENT NEURAL NETWORK

Recurrent Neural Network, known as Recursive Neural Network, also belongs to the family of Neural Networks, in which the association between nodes form a graph that is connected along a temporal sequence. Recurrent means the current output becomes the input at the next step, unlike Feed Forward Neural Network. RNN helps to make a good prediction in Sentiment Classification [66]. Details of test classification results can be seen in confusion matrix shown in figure 2(i). It achieved test accuracy of 91.71% while cross validation accuracy was 90.88%.

C. MODEL VALIDATION

To check the performance of the model, reviews are classified by using the 9:1 ratio. After getting the results of classification, the model is cross-validated using the K-Fold cross-validation technique. The value of K was defined as 10.

VI. DISCUSSION OF RESULTS

This research had two goals—Dataset and finding a better model for Roman Urdu text Sentiment Analysis. Detailed discussion is as below

TABLE 7. Statics of the DSR.

Property	Characteristics
Total Reviews	24,000
Positive Reviews	12,000
Negative Reviews	12,000
Total Num. of Tokens	2,27,858
Total Num. of Unique Tokens	24,756
Total Text Characters	11,17,303
Review max. Length (Characters)	150
Review min. Length (Characters)	03
Review avg. Length (Characters)	46.55

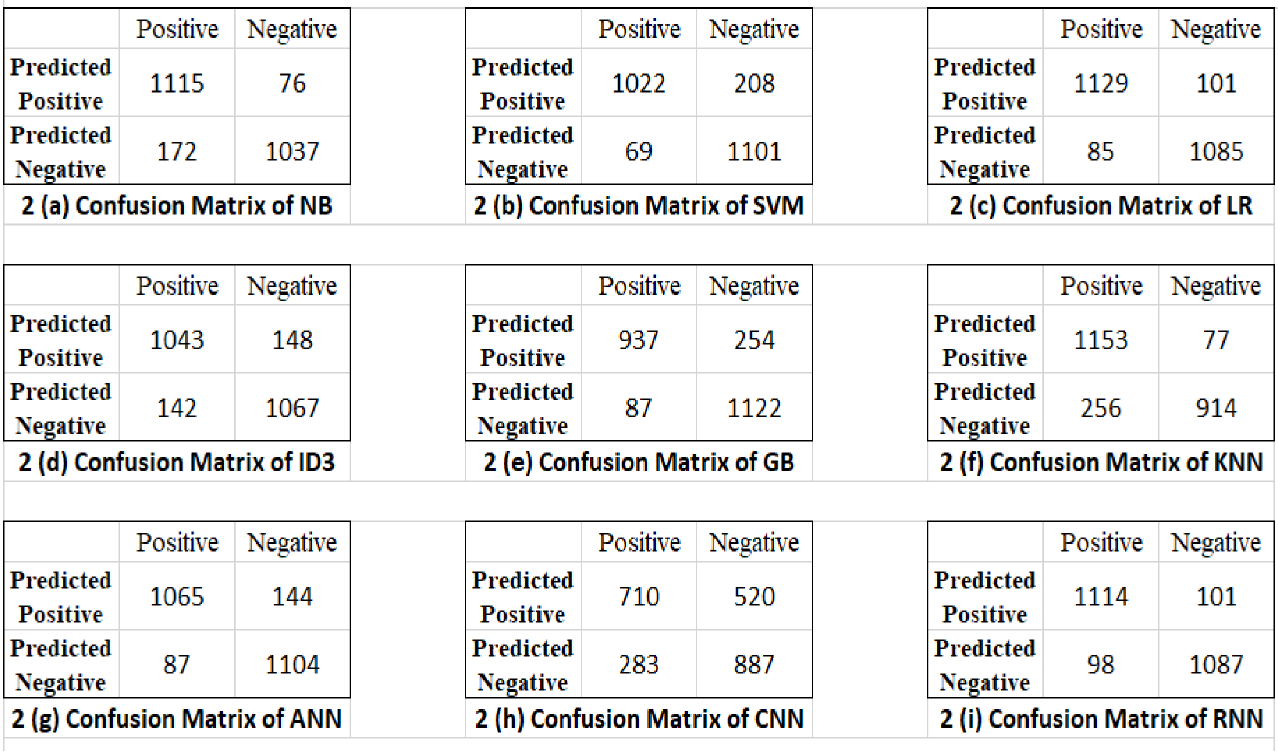
A. DATASET OF ROMAN URDU (DRU)

This dataset contains a collection of about 24,000 song reviews from YouTube. In this dataset, only polarized Urdu reviews written in roman script were being considered. The positive and negative reviews are equal in number. The entire corpus contains 11,17,303 characters. After the annotation phase, the inter-annotator agreement value was calculated with the help of the ‘‘Kappa coefficient’’. The value 0.87% showed that annotation quality is excellent. Statistics of the corpus DRU are shown in table 7.

B. FINDING A BETTER MODEL FOR ROMAN URDU TEXT SENTIMENT ANALYSIS

DRU was used for the binary classification. Four evaluation measures were calculated and considered for the purpose but keeping in view the importance and trend in literature review, highest importance was given to accuracy. Results were generated using TF/IDF method in Python. Table 8 shows comparative analysis of all nine models. Table 8 illustrates test results of all models.

As shown in Table 8, ANN and RNN showed better performance, but Logistic Regression outperformed all with 92.25% accuracy and 92.39% F-score on testing data.



	Positive	Negative
Predicted Positive	1115	76
Predicted Negative	172	1037

2 (a) Confusion Matrix of NB

	Positive	Negative
Predicted Positive	1022	208
Predicted Negative	69	1101

2 (b) Confusion Matrix of SVM

	Positive	Negative
Predicted Positive	1129	101
Predicted Negative	85	1085

2 (c) Confusion Matrix of LR

	Positive	Negative
Predicted Positive	1043	148
Predicted Negative	142	1067

2 (d) Confusion Matrix of ID3

	Positive	Negative
Predicted Positive	937	254
Predicted Negative	87	1122

2 (e) Confusion Matrix of GB

	Positive	Negative
Predicted Positive	1153	77
Predicted Negative	256	914

2 (f) Confusion Matrix of KNN

	Positive	Negative
Predicted Positive	1065	144
Predicted Negative	87	1104

2 (g) Confusion Matrix of ANN

	Positive	Negative
Predicted Positive	710	520
Predicted Negative	283	887

2 (h) Confusion Matrix of CNN

	Positive	Negative
Predicted Positive	1114	101
Predicted Negative	98	1087

2 (i) Confusion Matrix of RNN

FIGURE 3. Confusion matrixes of all classifiers.

TABLE 8. Test results of nine models.

ALGO	ACCURACY	RECALL	PRECISION	F-SCORE
NB	89.67	86.66	93.62	89.99
SVM	88.46	93.68	83.09	88.06
LR	**92.25**	92.99	91.79	92.39
ID3	87.92	88.02	87.57	87.79
GB	85.79	91.50	78.67	84.60
KNN	86.13	81.83	93.79	87.38
ANN	90.38	92.45	88.09	90.22
CNN	66.54	71.50	57.72	63.88
RNN	91.71	91.91	91.69	91.80

On the validation of the model, LR outperformed and attained 91.47% accuracy. Whereas the ANN and RNN also performed well and attained 90.38% and 91.71% accuracy on testing data, the values of F-score were 90.22% and 91.80% respectively. On Recall the SVM outperform the other classifiers with the value of 93.68%. On precision KNN outperformed all with the value of 93.79%, NB also performed very good and was close to KNN, and its value of precision is 93.62%. As stated above, the models were validated using the 10-Fold cross-validation technique to validate the results (to see if the model is over-fitted or under-fitted). It was witnessed that there was no big difference between testing accuracy and validation accuracy which confirms that model is good fitted (see Figure 4).

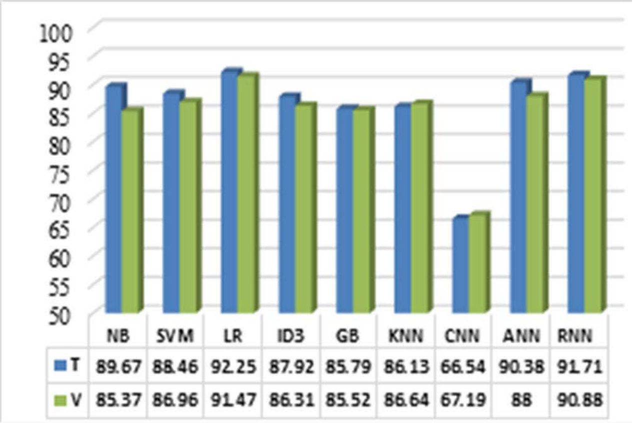


FIGURE 4. Classification and validation accuracies comparison.

Looking at the literature review it can be seen tha best accuracy of closely related work, presented in [17], on Roman Urdu Sentiment Analysis achieved 91% accuracy by using Support Vector Machine, while this research is showing 91.47 validation accuracy with 92.25% test accuracy on a a dataset of 24000 reviews (which was bigger than the dataset of [17] i.e. 1600 reviews).

VII. CONCLUSION

This research was carried out to define a mechanism to see people's sentiments about some entity through their reviews in Roman Urdu. The targeted audience of this research was the people of the sub-continent. The targeted language was Urdu written in roman script. To generate dataset of Roman Urdu text, reviews of 20 video songs were collected from YouTube. After performing preprocessing and data labelling, different Machine Learning algorithms were applied. For the results generation, implementation was performed in Python and results were generated by applying TF/IDF. The nine different classifiers NB, ID3, GB, SVM, LR, KNN, ANN, CNN and RNN were applied on labelled data. Experiments were conducted on Binomial Dataset, where Logistic Regression outperforms the other classifiers with 92.25% accuracy. For cross-validation, K-Fold was applied with k = 10. In cross-validation, LR beat all other algorithms with an accuracy of 91.47%. Based on the experimental results shown above, this study recommends LR and RNN for the classification of the Roman Urdu datasets.

This study opens a window for further research to improve classification results. This research can also be enhanced further with the standardization of Roman Urdu corpus. This study highlights the importance of Part of Speech tagging in Roman Urdu. The Dataset—DRU used for this research can be accessed by following: https://drive.google.com/file/d/1bml7fMTjJ1ZBxaDgx-AWpJjXLOK1vGsN/view?usp=sharing

ACKNOWLEDGMENT

The authors would like to thank Umar Shoukat and Muneeb Fazal for their valuable time and efforts in helping them in the data collection and annotation process.

REFERENCES

[1] R. J. Varey, ‘‘Internal marketing: A review and some interdisciplinary research challenges,’’ *Int. J. Service Ind. Manage.*, vol. 6, no. 1, pp. 40–63, Mar. 1995.

[2] W. P. M. Wong, M. C. Lo, and T. Ramayah, ‘‘The effects of technology acceptance factors on customer e-loyalty and e-satisfaction in Malaysia,’’ *Int. J. Bus. Soc.*, vol. 15, no. 3, pp. 477–502, 2014.

[3] T. Ogink and J. Q. Dong, ‘‘Stimulating innovation by user feedback on social media: The case of an online user innovation community,’’ *Technol. Forecasting Social Change*, vol. 144, pp. 295–302, Jul. 2019.

[4] H. Wang and Y. Wang, ‘‘A review of online product reviews,’’ *J. Services Sci. Manag.*, vol. 13, no. 1, pp. 88–96, 2020.

[5] Narain, A. P. Kavoori, and A. Punathambekar, ‘‘Bring back the old films, our culture is in disrepute,’’ in *Global Bollywood*. New York, NY, USA: NYU Press, 2008, pp. 164–179.

[6] S. Moon, P. K. Bergey, and D. Lacobucci, ‘‘Dynamic effects among movie ratings, movie revenues, and viewer satisfaction,’’ *J. Marketing*, vol. 74, no. 1, pp. 108–121, Jan. 2010, doi: 10.1509/jmkg.74.1.108.

[7] H. L. Vogel, *Entertainment Industry Economics: A Guide for Financial Analysis*, 9th ed. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[8] R. Raby, C. Caron, S. Théwissen-LeBlanc, J. Prioletta, and C. Mitchell, ‘‘Vlogging on YouTube: The online, political engagement of young Canadians advocating for social change,’’ *J. Youth Stud.*, vol. 21, no. 4, pp. 495–512, Apr. 2018.

[9] S. Zhang, T. Aktas, and J. Luo, ‘‘Mi Youtube Es su YouTube? Analyzing the cultures using Youtube thumbnails of popular videos,’’ 2020, *arXiv:2002.00842*.

[10] M. A. C. Jondar, ‘‘Rich Youtuber, poor Youtuber: Implementasi business intelligence dalam meningkatkan pendapatan channel Youtube YE,’’ Univ. Surabaya, Surabaya, Indonesia, Tech. Rep. 258409, 2020.

[11] S. Choi and A. Segev, ‘‘Finding informative comments for video viewing,’’ *Social Netw. Comput. Sci.*, vol. 1, no. 1, p. 47, Jan. 2020.

[12] A. Madden, I. Ruthven, and D. Mcmenemy, ‘‘A classification scheme for content analyses of Youtube video comments,’’ *J. Documentation*, vol. 69, no. 5, pp. 693–714, Sep. 2013, doi: 10.1108/JD-06-2012-0078.

[13] J. Philip, A. Baby, and A. Kannammal, ‘‘The good, the bad and the ugly: Opinion mining analysis on user tweets in Twitter,’’ *Int. J. Emerg. Technol. Innov. Res.*, vol. 6, no. 5, pp. 124–130, May 2019.

[14] P. K. Mallick, V. E. Balas, A. K. Bhoi, and A. F. Zobaa, *Preface*, vol. 768. Singapore: Springer 2019.

[15] S. Badugu, ‘‘Telugu movie review sentiment analysis using natural language processing approach,’’ in *Data Engineering and Communication Technology*. Cham, Switzerland: Springer, 2020, pp. 685–695.

[16] A. J. Dueppen, M. L. Bellon-Harn, N. Radhakrishnan, and V. Manchaiah, ‘‘Quality and readability of English-language internet information for voice disorders,’’ *J. Voice*, vol. 33, no. 3, pp. 290–296, May 2019.

[17] 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.

[18] Z. Yang, ‘‘Cognition and function research on fuzzy anaphora of English and Chinese narrative discourse in computer science area,’’ in *Proc. 4th Int. Conf. Mach., Mater. Comput.* Beijing, China: Atlantis Press, 2018, doi: 10.2991/macmc-17.2018.137.

[19] Z. Zhang, ‘‘Sentiment analysis of Chinese commodity reviews based on deep learning,’’ in *Proc. Int. Conf. Modern Educ. Technol. Innov. Entrepreneurship (ICMETIE)*, 2020, pp. 22–28.

[20] M. K. Elhadad, K. F. Li, and F. Gebali, ‘‘Sentiment analysis of Arabic and English tweets,’’ in *Proc. Workshops Int. Conf. Adv. Inf. Netw. Appl.*, 2019, pp. 334–348.

[21] M. Humayoun, H. Hammarström, and A. Ranta, *Urdu Morphology, Orthography and Lexicon Extraction*. Gothenburg, Sweden: Chalmers Tekniska Högskola, 2007.

[22] B. G. Patra, D. Das, and S. Bandyopadhyay, ‘‘Multimodal mood classification of Hindi and western songs,’’ *J. Intell. Inf. Syst.*, vol. 51, no. 3, pp. 579–596, Dec. 2018, doi: 10.1007/s10844-018-0497-4.

[23] S. Harris, A. Trippe, D. Challis, and N. Swycher, ‘‘Construction and evaluation of gold standards for patent classification—A case study on quantum computing,’’ *World Patent Inf.*, vol. 61, Jun. 2020, Art. no. 101961.

[24] M. Bilal, H. Israr, M. Shahid, and A. Khan, ‘‘Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, decision tree and KNN classification techniques,’’ *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, Jul. 2016, doi: 10.1016/j.jksuci.2015.11.003.

[25] Z. Papacharissi, ‘‘Sentiment analysis of Roman Urdu/Hindi using supervised methods,’’ *Ain Shams Eng. J.*, vol. 2, no. 3, pp. 1093–1113, 2013.

[26] D. Shubham, P. Mithil, M. Shobharani, and S. Sumathy, ‘‘Aspect level sentiment analysis using machine learning,’’ in *Proc. IOP Conf., Mater. Sci. Eng.*, 2017, vol. 263, no. 4, Art. no. 042009, doi: 10.1088/1757-899X/263/4/042009.

[27] Z. Sharf, D. Saif, and U. Rahman, ‘‘Performing natural language processing on Roman Urdu datasets,’’ *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 1, pp. 141–148, 2018.

[28] Sadeghian, ‘‘Bag of words meets bags of popcorn,’’ Stanford Univ., Stanford, CA, USA, Tech. Rep. CS224N Project, 2013, pp. 4–9.

[29] K. Sarawgi and V. Pathak, ‘‘Opinion mining: Aspect level sentiment analysis using SentiWordNet and Amazon web services,’’ *Int. J. Comput. Appl.*, vol. 158, no. 6, pp. 31–36, Jan. 2017, doi: 10.5120/ijca2017912830.

[30] P. Patil and P. Yalagi, ‘‘Sentiment analysis using aspect level classi?cation,’’ *Int. J. Sci. Res. Sci., Eng. Technol.*, vol. 4, no. 4, pp. 23–27, 2016.

[31] M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe, and K. Ghag, ‘‘Dimensionality reduction for sentiment analysis using pre-processing techniques,’’ in *Proc. Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Jul. 2017, pp. 16–21, doi: 10.1109/ICCMC.2017.8282676.

[32] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, ‘‘Preprocessing techniques for text mining–An overview,’’ *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2015.

[33] M. Asif, M. A. Qureshi, A. Abid, and A. Kamal, ‘‘A dataset for the sentiment analysis of indo-pak music industry,’’ in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2019, pp. 1–6.

[34] G. M. Sarria M., J. Diaz, and C. Arce-Lopera, ‘‘Analyzing and extending the salsa music dataset,’’ in *Proc. XXII Symp. Image, Signal Process. Artif. Vis. (STSIVA)*, Apr. 2019, pp. 1–5, doi: 10.1109/STSIVA.2019.8730229.

[35] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, ‘‘1000 songs for emotional analysis of music,’’ in *Proc. 2nd ACM Int. Workshop Crowdsourcing Multimedia (CrowdMM)*, 2013, pp. 1–6, doi: 10.1145/2506364.2506365.

[36] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2592–2602, 2009, doi: 10.1016/j.eswa.2008.02.021.

[37] J. Hendler, "Data integration for heterogenous datasets," *Big Data*, vol. 2, no. 4, pp. 205–215, Dec. 2014, doi: 10.1089/big.2014.0068.

[38] R. Kitchin, "The real-time city? Big data and smart urbanism," *Geojournal*, vol. 79, no. 1, pp. 1–14, Feb. 2014.

[39] D. Munkovă, M. Munk, and M. Vozár, "Data pre-processing evaluation for text mining: Transaction/sequence model," *Proc. Comput. Sci.*, vol. 18, pp. 1198–1207, Jan. 2013, doi: 10.1016/j.procs.2013.05.286.

[40] G. Qi, Z. Zhu, K. Erqinhu, Y. Chen, Y. Chai, and J. Sun, "Fault-diagnosis for reciprocating compressors using big data and machine learning," *Simul. Model. Pract. Theory*, vol. 80, pp. 104–127, Jan. 2018, doi: 10.1016/j.simpat.2017.10.005.

[41] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Comput. Speech Lang.*, vol. 15, no. 3, pp. 287–333, Jul. 2001.

[42] R.-M. Karampatsis, H. Babii, R. Robbes, C. Sutton, and A. Janes, "Big code!= big vocabulary: Open-vocabulary models for source code," 2020, *arXiv:2003.07914*.

[43] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Marketing*, vol. 36, no. 1, pp. 20–38, 2019.

[44] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Discriminative feature spamming technique for Roman Urdu sentiment analysis," *IEEE Access*, vol. 7, pp. 47991–48002, 2019, doi: 10.1109/ACCESS.2019.2908420.

[45] 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*, vol. 70, no. 3, pp. 4987–5004, 2022.

[46] T. I. Jain and D. Nemade, "Recognizing contextual polarity in phrase-level sentiment analysis," *Int. J. Comput. Appl.*, vol. 7, no. 5, pp. 12–21, 2010.

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

[48] M. Abdul-Mageed and M. T. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," in *Proc. 8th Int. Conf. Lang. Resour. Eval.*, vol. 515, 2012, pp. 3907–3914.

[49] 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, doi: 10.1016/j.ipm.2020.102368.

[50] S. Khedkar and S. Shinde, "Deep learning-based approach to classify praises or complaints from customer reviews," in *Proc. Int. Conf. Comput. Sci. Appl.*, vol. 2020, pp. 391–402.

[51] S. A. Mostafa and M. Z. Saringatb, "Comparative analysis for Arabic sentiment classification," in *Proc. Appl. Comput. Support Ind., Innov. Technol.: 1st Int. Conf. (ACRIT)*, Ramadi, Iraq, vol. 1174, Sep. 2020, pp. 271–285.

[52] S. Kramer and C. Helma, *Machine Learning and Data Mining*, vol. 42, no. 11. Sawston, U.K.: Woodhead, 2005.

[53] C. Bielza and P. Larrañaga, "Discrete Bayesian network classifiers: A survey," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 1–43, 2014.

[54] H. B. Barua and K. C. Mondal, "A comprehensive survey on cloud data mining (CDM) frameworks and algorithms," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–62, Sep. 2020.

[55] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics*. Cham, Switzerland: Springer, 2020, pp. 99–111.

[56] A. Mittal and S. Patidar, "Sentiment analysis on Twitter data: A survey," in *Proc. 7th Int. Conf. Comput. Commun. Manage.*, vol. 2019, pp. 91–95.

[57] D. J. Kalita, V. P. Singh, and V. Kumar, "A survey on SVM hyper-parameters optimization techniques," in *Social Networking and Computational Intelligence*. Cham, Switzerland: Springer, 2020, pp. 243–256.

[58] A. Donnelly, B. Misstear, and B. Broderick, "Real time air quality forecasting using integrated parametric and non-parametric regression techniques," *Atmos. Environ.*, vol. 103, pp. 53–65, Feb. 2015, doi: 10.1016/j.atmosenv.2014.12.011.

[59] S. Fletcher and M. Z. Islam, "Decision tree classification with differential privacy: A survey," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–33, 2019.

[60] S. Lomax and S. Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Comput. Surv.*, vol. 45, no. 2, pp. 1–35, Feb. 2013.

[61] T. Wang, Z. Li, Y. Yan, and H. Chen, "A survey of fuzzy decision tree classifier methodology," in *Fuzzy Information and Engineering* (Advances in Soft Computing), vol. 40, no. 3. Berlin, Germany: Springer, 2007, pp. 959–968.

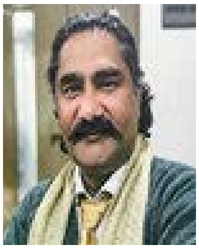
[62] C. Böhm, S. Berchtold, and D. A. Keim, "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," *ACM Comput. Surv.*, vol. 33, no. 3, pp. 322–373, 2001.

[63] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.

[64] R. Kumar, H. S. Pannu, and A. K. Malhi, "Aspect-based sentiment analysis using deep networks and stochastic optimization," *Neural Comput. Appl.*, vol. 32, no. 8, pp. 3221–3235, Apr. 2020, doi: 10.1007/s00521-019-04105-z.

[65] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, Apr. 2017, doi: 10.1016/j.eswa.2016.10.065.

[66] S. Timotheou, "The random neural network: A survey," *Comput. J.*, vol. 53, no. 3, pp. 251–267, 2010, doi: 10.1093/comjnl/bxp032.



MUHAMMAD AASIM QURESHI is a seasoned academician and researcher with 20 years of professional experience. More than 30 publications are on his credit including his Ph.D. degree in algorithms. His current areas of interest include artificial intelligence, algorithms, machine learning, and fuzzy logic. He has supervised numerous projects and theses related to virtual/augmented reality, recommender systems, sentiment analysis, robot navigation, etc.

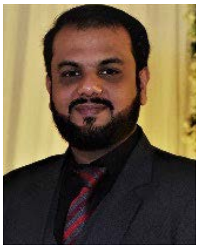
He is the session chair of several national and international conferences. He has also been an invited speaker at various research events. He is also a reviewer of various international journals and conferences.

MUHAMMAD ASIF received the M.S./M.Phil. degree in computer sciences from Bahria University Lahore Campus, Lahore, Pakistan. He is currently pursuing the Ph.D. degree with the University of Bologna, Italy. As a reviewer, he is working with multiple journals, i.e., IEEE Access, *Computers, Materials and Continua*, *Intelligent Automation and Soft Computing*, and some others. His research interests include natural language processing, machine learning, artificial intelligence, data sciences, and big data.

MOHD FADZIL HASSAN received the B.Sc. degree in information systems from Colorado State University, Fort Collins, CO, USA, in 1998, the M.Sc. degree in artificial intelligence, in 2001, and the Ph.D. degree in informatics from the University of Edinburgh, U.K., in 2007.

Since 2001, he has been an Associate Professor, the Head of Department, and the Dean of the Centre for Graduate Studies with Universiti Teknologi PETRONAS, Malaysia. His research interests include software engineering, agents, algorithms, service-oriented architectures, and information security. He has served for various academic, management, and conference committees. He had received many awards and honors throughout his career. He is a member of different societies.

ADNAN ABID received the M.S. degree in information technology from the National University of Science and Technology, Pakistan, and the Ph.D. degree in information engineering from the Politecnico di Milano. He is currently working as an Associate Professor and a Director with the Center for Advancement of Science and Technology (CAST), School of System and Technology, Lahore, Pakistan. His research interests include information retrieval, machine learning, and ranked query processing.



ASAD KAMAL received the B.S. degree in computer science from the Virtual University of Pakistan and the M.S. degree from Bahria University. From mid of 2013 to the start of 2017, he worked at multiple software houses in various positions. Since 2017, he has been working as a Faculty Member with the Computer Science Department, Bahria University Lahore Campus. His research interests include natural language processing, machine learning, and big data.

SOHAIL SAFDAR received the M.C.S. degree in computer science, in 2004, and the M.S. degree in software engineering from Bahria University, Islamabad, Pakistan, in 2008, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2014.

From 2004 to 2009, he was a Lecturer at Bahria University, Islamabad, and in 2009, he joined Universiti Teknologi PETRONAS for his Ph.D. degree studies. He was a Lecturer with Universiti Tunku Abdul Rahman, Malaysia, from 2012 to 2016. He is currently an Assistant Professor with Ahlia University, Bahrain. His research interests include software engineering and information and cyber security. He had served in various academic and conference committees.

REHAN AKBAR (Senior Member, IEEE) received the M.Sc. degree in computer science from the University of Agriculture, Faisalabad, Pakistan, in 2001, the M.S. degree in computer science with specialization in software engineering from the Government College University, Lahore, Pakistan, in 2008, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2013.

From 2001 to 2008, he was working as a Lecturer with GC University, Lahore. Later, he joined IT Industry as a Project Manager. In 2012, he joined the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia, as a Lecturer and reached to the rank of an Associate Professor and the Head of Department. Since 2022, he has been working as an Associate Professor with the Department of Computer and Information Science, Universiti Teknologi PETRONAS. His research interests include software development processes and methodologies, process tailoring, agile methodologies, big data, and cybersecurity.

Dr. Akbar was a member of the Association for Information Systems. He received the Professional Technologist (TS/P.Tech.) Status from Malaysian Board of Technologists and Teaching Excellence Award for year 2016. He is a reviewer of different journals and conferences.