JCBI
Journal of Computing & Biomedical Informatics Volume 09 Issue 01
ISSN:2710-1606 2025

Research Article
https://doi.org/10.56979/901/2025

Sentiment Analysis of Urdu Language Using Machine Learning Models

Haroon Yousaf[1], M. Imran Khan Khalil[1*], Zia Ur Rahman[1*], Firdous Ayub[2], Yousaf Khan[3], Asif Nawaz[4], Zeeshan Najam[5], and Sheeraz Ahmed[6]

[1]University of Engineering & Technology Peshawar, 25000, Pakistan.
[2]Department of Computer Science, Women University, Swabi, Khyber Pakhunkhwa, Pakistan.
[3]Elementary & Secondary Education, Khyber Pakhunkhwa, Peshawar, 25000, Pakistan.
[4]Higher Colleges of Technology, UAE, Dubai Campus.
[5]Preston University, Islamabad Campus, Pakistan.
[6]Iqra National University, Peshawar, 25000, Pakistan.
[*]Corresponding Author: Muhammad Imran Khan Khalil. Email: imrankhalil@uetpeshawar.edu.pk

Received: March 30, 2025 **Accepted:** May 15, 2025

Abstract: The purpose of the study was to develop sentiment analysis tool and system for Urdu language through the use of Machine Learning models and deep learning algorithms. In modern age of social media communication and interaction, the anlysis of opinion of the users and customer's of business products, political discourses, religious and cultural crtiques and debates to explore their sentiments abouth the services, products, scenarios and stories are much needed. The methodology used by the researchers was to create dataset from tweets and posts on social media platforms in Urdu language. The researchers then used pre-processing and Lemmatization data to make it suitable for trainig and testing when using machine learning algorithms such as Support Vector Machine, Naïve Bayes, BiLSTM and other realted models.The sentiment analyzer categorized the sentiments in Urdu language text with nine categroies and thus the resulted tool or system was tested for its effectivness and efficeincy with more than 50% accuracy and high level of validaity were found. Thus, the sentiment analysis of Urdu language using machine learning models was deceloped to bbridge the gap in sentiment analysis for Urdu language in Pakistan.

Keywords: Machine Learning, Nastaleeq Script, Deep Learning, Urdu Language

1. Introduction

In the modern world of digital, online and AI-based platforms for communication in businesses, political discussions, social, cultural and religious discourses and debates, there is need of analyzing the sentiments of individuals, groups and organization for providing feedback and bringing changes, modifications, innovations and improvements in products, decisions, evaluations, judgments and plans implementation [1],[2],[3].

The software apps and tools are available in English language for sentiment analysis but there is need to develop sentiment analyzer in broad categories using Urdu language (Nastaleeq script) as gap identified by the researchers through the in depth literature review for the study [4],[5],[6],[7]. There is distinctive contextual sensitivity when using Urdu language (Nastaleeq script) in writing or text messages [8], [9]. As in the Urdu text messages the letter glyphs dynamically change their shape [1] and creating challenges for sentiment analysis. The national Language Authority (NLA) [2] has standardized the Haroof-e-Tahaji (Urdu Alphabets) as given below [10], [11].

The national language of Pakistan is Urdu and is spoken by millions of people in Pakistan and the Asians or Indians living in foreign countries as overseas in other countries of the world. So, the researchers intended to make a smart sentiment analyzer for Urdu language text using deep learning techniques and

machine learning models [12],[13]. Many researchers have tried to explore sentiment analysis but they used small categorization of emotions and sentiment using two or three categories such as happiness; sadness; sadness, happiness and neutral etc. usually the sentiment analyzers are based on three category system as positive, negative and neutral [14], [15],[16].

ا ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ

د دھ ڈ ڈھ ز ر رھ ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ

ف ق ک کھ گ گھ ل لھ م مھ ن نھ ں ںھ و ہ ۃ ء ی یھ ے

Figure 1. Urdu Alphabets Adopted from NLA (Haroof e Tahaji)

The study has tried to develop a model in nine categories of sentiment analysis with at least 50% or more accuracy. So, in the study machine learning models such as BiLSTM, CNN, RNN , Random Forest, Support Vector Machine, Logistic Regression, and Naïve Bayes were used and adopted for optimal performance of the Urdu Sentiment analysis [3] [17],[18]. The researchers adopted structured methodologies to involve the application of Machine Learning (ML) approach, tools and techniques [4], [19],[20]. For this purpose the researchers compiled dataset of Urdu text messages involving sentiments or opinions of customers and user of products or services from API of Twitter [21],[22],[23],[24].

A pre-processing steps used for cleaning and standardization, feature extraction from dataset such as tokenization; the researchers selected and used various algorithm that included BiLSTM, CNN, RNN , Random Forest, Support Vector Machine, Logistic Regression, and Naïve Bayes [3], [25],[26],[27],[28]. These algorithms were used for identifying the underlining sentiment or emotion within dataset of a given language [29]. The sentiment analysis of Urdu language using deep learning techniques made it possible to categorize a text in Urdu language in the following nine categories i.e. happiness, sadness, love, anger, hatred, fear, disappointment and peace [30],[31],[32].

After training the model with the dataset, accuracy, F1 score and recall were used for evaluation and assessment of performance as well as its effectiveness for sentiment analysis in Urdu language sentences, words and phrases in the dataset [33],[34],[35],[36]. Hence, the researchers achieved the main aim and objective of the study for sentiment analysis of Urdu language using deep learning techniques, algorithm and machine learning model system [37][,38],[39],[40].

The study is significant because it has filled the gap in research by developing a model for sentiment analysis of Urdu language using deep learning algorithm and machine learning models. It is also significant for bordering the sentiment analysis model from three category system to nine category system in Urdu language (Nastaleeq script) [41].

2. Materials and Methods

To evaluate the developed model system effectiveness, the researchers produced a Corpus of 148 tweets in each category of sentiments thus a total of 1332 tweets were used in the study for sentiment analysis in Urdu language. Figure 2 below has explained the steps used in pre-processing of dataset, Figure 3 shows the dataset distribution, Figure 4 shows dataset cleaning and lemmatization and Figure 5 shows confusion matrices of different models [42].

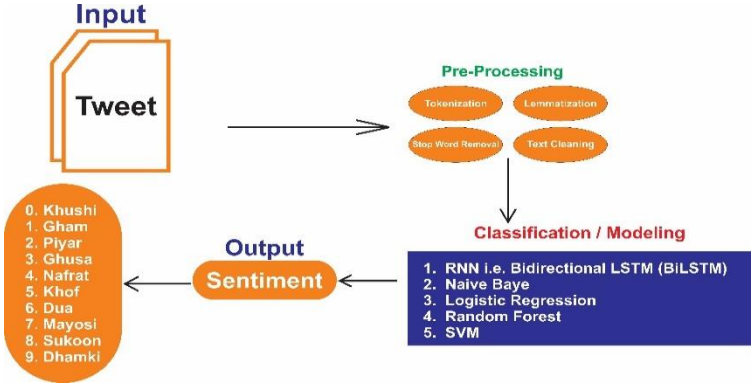

Figure 2. Flow Chart for Pre-processing of dataset

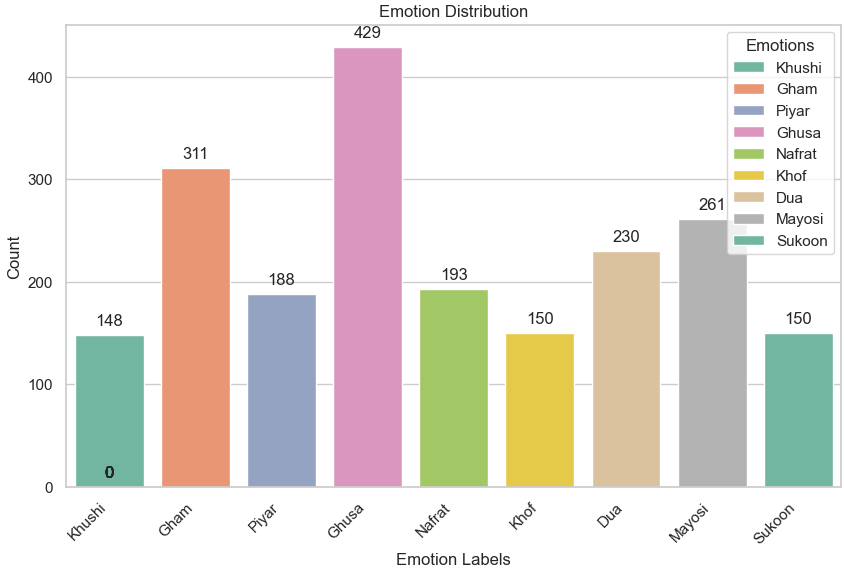

Figure 3. Bar Graph of Nine Categories of Sentiment

	Label	Tweet	Emotion	text_cleaned	Lemmatized
0	9	...گھر میں داخل ہوتے ہی جس ہستی کو دیکھ کر سکون م	Dhamki	...گھر داخل ہستی دیکھ سکون ملتا ھے ماں ھے جسے دیک	...گھر, داخل, ہستی, دیکھ, سکون, ملنا, ھے, ماں, ھ]
1	9	... سب چُھوٹ جائے مگر رَبّ سے رابطہ نہیں چُھوٹنا*	Dhamki	... چُھوٹ رَبّ رابطہ چُھوٹنا چاہیے، خود چل اُس غم	...چُھوٹ, رَبّ, رابطہ, چُھوٹنا, چاہیے, خود, چلن]
2	9	... اب تو لازمی کرنا ہوگا ... 😁.. نہیں تو میں کٹی	Dhamki	لازمی ہوگا کٹی	[لازمی, ہونا, کٹنا]
3	9	... تو کیا \n\nتو نے چکھی ہے فقط گناہوں کی لذت.....	Dhamki	چکھی فقط گناہوں لذت..... ذکر الہی سرور کتنا...	...چکھنا, فقط, گناہوں, لذت....., ذکر, الہی, سرور]
4	9	...سب سے خوبصور*\n\n*🌱مُخْتَصر پُر اَثر🌱*	Dhamki	...مُخْتَصر پُر اَثر خوبصورت حیران انجینر	... ,مُخْتَصر, پُر, اَثر, خوبصورت, حیران]

Figure 4. Cleaning and Lemmatization

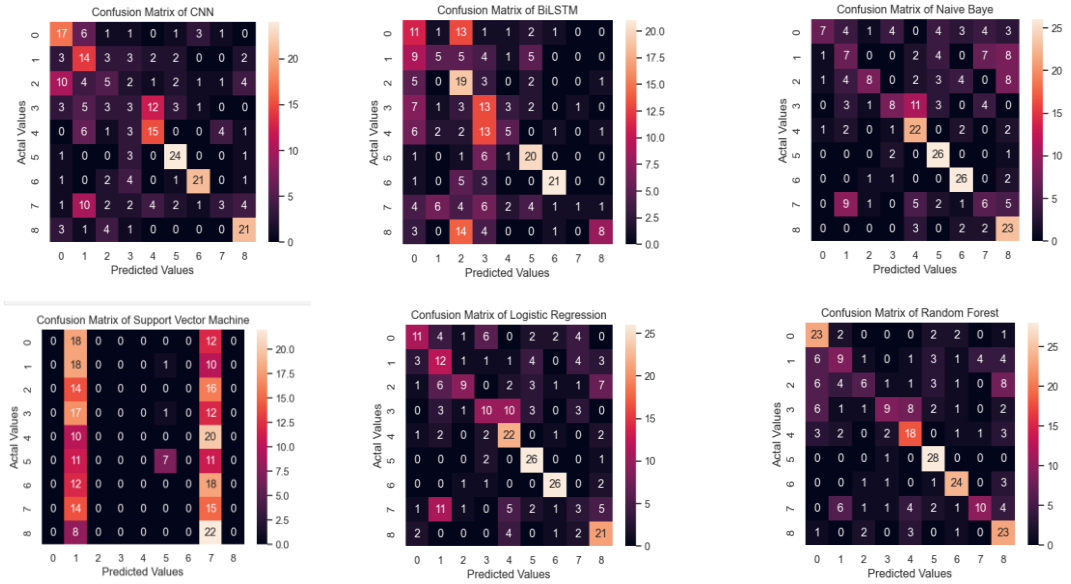

Figure 5. Confusion Matrices of different Models

The software apps used for the experiments included Python and within the Python the researchers have leveraged powerful libraries for achieving the desired tasks such as Scikit-learn, Numpy, Pandas and Seaborn.

The researchers in the study used both qualitative and quantitative research methods for data analysis. In quantitative data analysis the focus was on categorization of Urdu language text (expressed

sentiment) in nine categories as discussed in the introduction. In qualitative method it was focused to complement the analysis with more expert assessment of the systems' sentiment taxonomy [43].

The study used different algorithms for various discussions such as Bidirectional Long Short Term Memory (BiLSTM) for informative flow in both directions i.e. back and forth [5]. Support Vector Machine (SVM) was used for grouping the text and creating boundaries within the group [6]. Figure 6 has shown how SVM has drawn a line to classify two different groups of things.

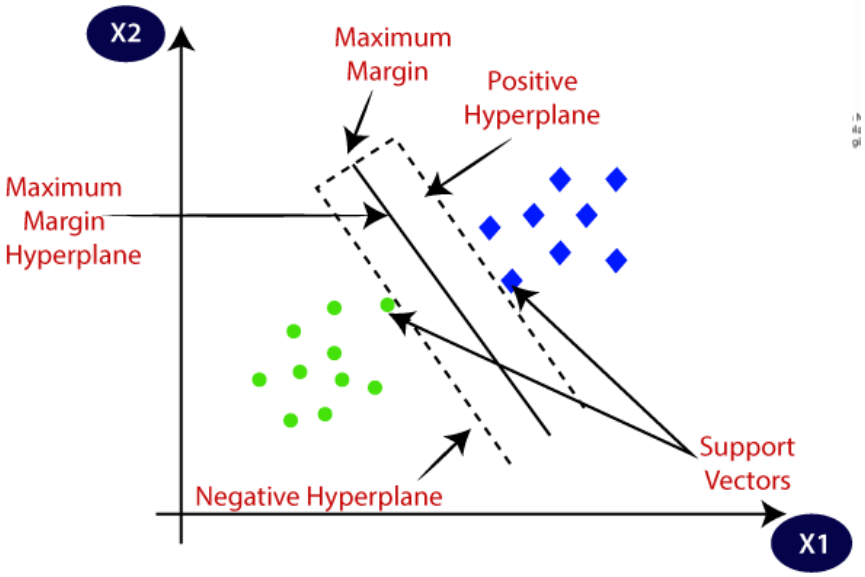

Figure 6. SVM draws a line to classify two different groups of things

Naïve Bayes (MultinomialNB) was used to allow the researcher to "inverse" conditional probabilities and to be used as generative learning algorithm [7]. Logistic Regression (LR) was used for binary classification problem as well as multiple classifications to estimate the probability such as head or tail etc. based on independent variables present in the dataset [8].

Random Forest (RF) classification algorithm was used for multiple decision trees in the output to converge into a single result, thus it helped both in classification and regression problems [9]. The performance metrics included Confusion Matrix with parameters: True and False Positive; True and False Negative (Figure 7)

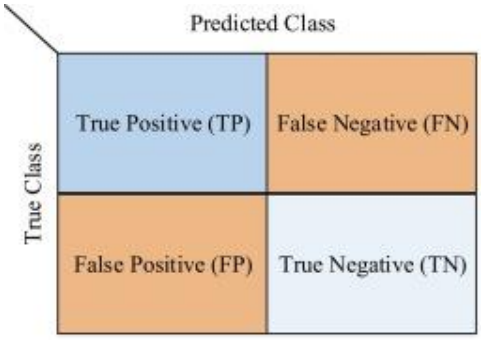

Figure 7. Confusion Matrix

It also included accuracy for presenting the proportion of correcting predicted observation and as common matrices for performance evaluation. The accuracy of the model as calculated is given by:

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

The researchers also used three additional metrics: F1 score, Precision and Recall for evaluation of the model's effectiveness [10] and evaluation.

3. Results

The main purpose of the study was to bridge the existing gap in sentiment analysis tools and techniques for Urdu language text and to develop a model to be applied in different field of life such as businesses, politics, religion or cultural discussions, social and commercial or industrial products.

The study achieved the objective by developing a model system, tool and technique for sentiment analysis in Urdu language with nine categories classification. The system involved the steps of data collection for pre-processing of the Urdu language text, extraction of relevant features and application of various machine learning algorithm with rigorous and robust evaluation of the resulting system. The sentiment analyzer developed could identify and analyze emotion in nine categories and can be extend to label more data and to improve accuracy of all the models to correctly identify emotions and sentiments expressed in Urdu language text.

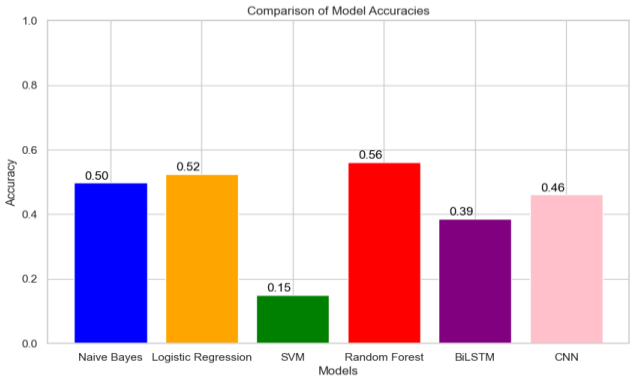

Figure 8. Different Models Accuracies

4. Discussion

This study aimed at developing an NLP tool for sentiment analysis of Urdu language using deep learning techniques and machine learning models. In the era of digital, online and social media platforms used for expression of opinion, presenting product reviews, political discourses, religious and cultural discussion or debate. There was need of developing tool and system for sentiment analysis in Urdu language. Urdu is having the challenges of orthographic variation, resource opacity and scarcity, lack of standardized text corpora, complex morphology, niche vocabulary and Perso-Arabic script (Nastaleeq script).

In the modern era deep learning has brought revolution in NLP by enabling end-to-end learning without the need of manual manipulation and data handcrafted feature [11]. While there are strategies, challenges and applications for English language sentiment analysis, Urdu is having comparatively limited strategies, challenges and applications or tools [12]. In the recent years sentiment analysis has gained popularity on different social media platforms including Twitter, Facebook and Instagram etc. [2]

Sentiment polarity and score are the two basic aspects of sentiment analysis widely used in literature [13]. In Urdu grammar there are two genders (Male/ Female), two numbers (Single / Plural) and three cases (Vocative Direct and Indirect) [14]. To detect source materials in the Urdu language text for emotion, expression or sentiment representation in subjective terms, computational linguistics and deep learning techniques are used [15].

In this study the sentiments were represented by nine categories of the text through the application of deep learning or machine learning models. While in the existing situation these are usually categorized into two or three categories [4] in NLP tools.

The study of [15] has discussed linguistic techniques such as parsing, named Entity Recognition and POS tagging for Urdu corpus development. However, it does not cover sentiment analysis in Urdu [16] which is addressed in the study there are no exclusively developed deep learning techniques for sentiment analysis in Urdu. Hence, there is lack of linguistic technique or NLP tools or system for sentiment analysis in Urdu as discussed in [17] and the present study has addressed this gap.

Another problem as discussed in literature in sentiment analysis of Urdu language is that there is need for an intelligent sentiment analysis tool to have acceptable level of accuracy and scalability [18] which are addressed in the study.

The researchers have classified the deep learning techniques and models for Urdu language sentiment analysis to be either lexicon based or hybrid approach [19]. So, there is need for a comprehensive system or tool to cover all aspects of Urdu language.

Hence, the study has tried to fill the gap and has developed sentiment analysis of Urdu language using deep learning techniques and machine learning models. The accuracy of the developed system is more than 50% with multiple categories of sentiment analysis.

5. Conclusions

The research study has filled the gap identified in sentiment analysis of Urdu language in nine categories using deep learning techniques, algorithm and machine learning models. However, there are more avenues to be explored in this area as suggested for further research.

6. Future Work

The researchers suggested the following areas for further research:

- To enhance the model performance in sentiment analysis of Urdu language and
- To integrate deep learning into sentiment analysis for Urdu language.

References

1. V., Thawakar, M., Choudhari, M., Chahande, S., Verma, S., & Pimpalkar, A. Chole, "Enhancing heart disease risk prediction with GdHO fused layered BiLSTM and HRV features: A dynamic approach," Biomedical Signal Processing and Control, vol. 95, p. 106470, 2024.
2. A. Z., Aslam, M., & Martinez-Enriquez, A. M. Syed, "Lexicon based sentiment analysis of Urdu text using SentiUnits. In Advances in Artificial Intelligence: 9th Mexican International Conference on Artificial Intelligence," in MICAI 2010, Pachuca, Mexico, 2010.
3. S., Durrani, N., & Gul, S Hussain, "Survey of language computing in Asia. Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences," , 2005.
4. L., Amjad, A., Ashraf, N., Chang, H. T., & Gelbukh, A. Khan, "Urdu sentiment analysis with deep learning methods.," IEEE access, vol. 9, pp. 97803-97812, 2021.
5. A. A., Tihami, M. N., & Islam, M. S. Sharfuddin, "A deep recurrent neural network with bilstm model for sentiment classification.," In 2018 International conference on Bangla speech and language processing (ICBSLP)) IEEE., pp. pp. 1-4, September 2018.
6. D. M Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.,"
7. Rane, N., Choudhary, S., & Rane, J. (2024). Machine learning and deep learning: A comprehensive review on methods, techniques, applications, challenges, and future directions. Techniques, Applications, Challenges, and Future Directions (May 31, 2024).
8. Liaqat, M. I., Hassan, M. A., Shoaib, M., Khurshid, S. K., & Shamseldin, M. A. (2022). Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. PeerJ Computer Science, 8, e1032.
9. B Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers Google Scholar Google Scholar Digital Library Digital Library, 2012.
10. A. Hardie, "Developing a tagset for automated part-of-speech tagging in Urdu.," in In Corpus Linguistics, 2003.
11. Y., Jin, R., & Zhou, Z. H Zhang, "Understanding bag-of-words model: a statistical framework," International journal of machine learning and cybernetics, vol. 1, pp. 43-52, 2010.
12. A., & Kirubakaran, D. E Jebaseel, "M-learning sentiment analysis with data mining techniques," International Journal of Computer Science and Telecommunications, vol. 3(8), pp. 45-48, 2012.
13. A., Asghar, M. Z., Saeed, A., Hameed, I. A., Hassan, S. A., & Ahmad, S. Khattak, "A survey on sentiment analysis in Urdu: A resource-poor language. ," Egyptian Informatics Journal, vol. 22, no. 1, pp. 53-74, 2021.
14. W., Daud, A., Nasir, J. A., Amjad, T., Arafat, S., Aljohani, N., & Alotaibi, F. S. Khan, "Urdu part of speech tagging using conditional random fields. ," Language Resources and Evaluation, vol. 53, pp. 331-362, 2019.
15. Khalil, M. I. K.; Afsheen, A.; Taj, A.; Nawaz, A.; Jan, N.; and S. Ahmad, "Enhancing Security Testing Through Evolutionary Techniques: A Novel Model" Journal of Computing & Biomedical Informatics, 2023, vol. 6, no. 1, pp. 375 – 393.
16. Khalil, M. I. K. and Taj, T. "Factors Affecting the Efficacy of Software Development: A Study of Software Houses in Peshawar, Pakistan," International Review of Basic and Applied Sciences, 2021, vol. 9, no. 3, pp. 385-393.
17. Khalil, M. I. K. Ullah, A.; Taj, A.; Khan, I.A.; Ullah, F.; Taj, F.; and Shah, S. "Analysis of Critical Risk Factors Affecting Software Quality: A Study of KPK, Pakistan Software Industry," International Review of Basic and Applied Sciences, 2022, vol. 10, no. 2, pp. 338-348.
18. Saqib, A.; Ullah, M.; Hyder, S.; Khatoon, S.; and Khalil, M. I. K. "Creative Decision Making in Leaders: A Case of Beer Game Simulation," Abasyn Journal of Social Sciences, 2020, vol. 12, no. 2, pp. 379-387.
19. Khan, I.A; Ullah, F.; Abrar, M.; Shah, S.; Taj, F. and Khalil, M.I.K. "Ransomware Early Detection Model using API-Calls at Runtime by Random Decision Forests," International Review of Basic and Applied Sciences, 2022, vol. 10, no. 2, pp. 349-359.
20. Jan, S.; Maqsood, I.; Ahmad, I.; Ashraf, M.; Khan, F. Q.; and Khalil, M. I. K. "A Systematic Feasibility Analysis of User Interfaces for Illiterate Users," Proceedings of the Pakistan Academy of Sciences, 2020, vol. 56, no. 4, pp. 2518-4253.
21. Khalil, M. I. K.; Shah, S. A. A.; Khan, I. A.; Hijji, M.; Shiraz, M.; and Shaheen, Q. "Energy cost minimization using string matching algorithm in geo-distributed data centers," Computers, Materials, and Continua, 2023, vol. 75, no. 3, pp. 6305-6322.
22. Ahmad, I.; Khalil, M. I. K.; and Shah, S. A. A. "Optimization-based workload distribution in geographically distributed data centers: A survey," International Journal of Communication Systems, 2020, vol. 33, no. 12, p. e4453.

23. Khalil, M. I. K.; Ahmad, I.; Shah, S. A. A.; Jan, S.; and Khan, F. Q. "Energy cost minimization for sustainable cloud computing using option pricing," Sustainable Cities and Society, 2020, vol. 63, p. 102440.
24. Khalil, M. I. K. "Improve quality of service and secure communication in mobile ad hoc networks (MANETS) through group key management," International Review of Basic and Applied Sciences, 2013, vol. 1, no. 3, pp. 107-115.
25. Muhammad, D.; Ahmad, I.; Khalil, M. I. K.; Khalil, W.; and Ahmad, O. A. "A generalized deep learning approach to seismic activity prediction," Applied Sciences, MDPI, 2023, vol. 13, p. 1698.
26. Khalil, M. I. K..; Umar, H.; Ullah, K.; Khamosh, S.U., Naqvi, S.F.M., Nawaz, A., and Ahmad, S. "Revolutionizing Schizophrenia diagnosis: A transfer learning approach to accurate classification," Journal of Computing & Biomedical Informatics, 2024, vol. 07, no. 2, pp. 1-12.
27. Khalil, M. I. K. "Job satisfaction and work morale among PhD's: A study of public and private sector universities of Peshawar, Pakistan," International Review of Management and Business Research, 2013, vol. 02, no. 2, p. 362.
28. Ahmad, I.; Ahmad, M. O.; Alqarni, M. A.; Almazroi, A. A.; and Khalil, M. I. K. "Using algorithmic trading to analyze short-term profitability of Bitcoin," PeerJ Computer Science, 2021, vol. 7, p. e337.
29. Khalil, M. I. K.; Shah, S. A. A.; Taj, A.; Shiraz, M.; Alamri, B.; Murawat, S.; and Hafeez, G. "Renewable aware geographical load balancing using option pricing for energy cost minimization in data centers," Processes, MDPI, 2022, vol. 10, no. 10, p. 1983.
30. Khalil, M. I. K.; Ahmad, A.; Almazroi, A.A. "Energy Efficient Workload Distribution in Geographically Distributed Data Centers," IEEE Access, 2019, vol. 7, no. 1, pp. 82672-82680.
31. Naz, S., and Khalil, M.I.K. "Determinants of job satisfaction: A case study of WAPDA, Peshawar" City University Research Journal, 2012, vol. 2, no. 2, pp. 92-107.
32. Khalil, M. I. K..; Mubeen, A.; Taj, A.; Jan, N.; Ahmad, S. "Renewable and Temperature Aware Load Balancing for Energy Cost Minimization in Data Centers: A Study of BRT, Peshawar," Journal of Computing & Biomedical Informatics, 2023, vol. 06, no. 3, pp. 183-194.
33. Anum, H.; Khalil, M. I. K..; Nawaz, A.; Jan, N.; and Ahmad, S. "Enhancing rumor detection on social media using machine learning and empath features," Journal of Computing & Biomedical Informatics, 2023, vol. 06, no. 2, pp. 272-281.
34. Khalil, M. I. K.; Khan, I.A; Nawaz, A.; Latif, S.; Ahmad, S. and Ahmad, S. "Unveiling the security Maze: A comprehensive review of challenges in Internet of things," Special Issue on Intelligent Computing of Applied Sciences and Emerging Trends (ICASET), Journal of Computing & Biomedical Informatics, 2024, pp. 10-19.
35. Khalil, M.I.K. "Improve quality of service in MANETS through 2-hop routing in cluster-based routing protocol," City University Research Journal, 2012, vol. 2, no. 2, pp. 26-35.
36. Ahmad, I.; Ahmad, M. O.; Alqarni, M. A.; Almazroi, A. A.; and Khalil, M. I. K. "Using algorithmic trading to analyze short-term profitability of Bitcoin," PeerJ Computer Science, 2021, vol. 7, p. e337.
37. Khalil, M. I. K.; Shah, S. A. A.; Taj, A.; Shiraz, M.; Alamri, B.; Murawat, S.; and Hafeez, G. "Renewable aware geographical load balancing using option pricing for energy cost minimization in data centers," Processes, MDPI, 2022, vol. 10, no. 10, p. 1983.
38. Rahman, Z.U..; Khalil, M.I.K..; Nawaz, A.; Khan, I.A..; Jan, N..; and Sheeraz, A. "Analysis and clustering of Pakistani music by lyrics: A study of CokeStudio Pakistan," Journal of Computing & Biomedical Informatics, 2024, vol. 7, no. 1, pp. 281-296.
39. Khan, I..; Khalil, M.I.K..; Nawaz, A.; Khan, I.A..; Zafar, S..; and Sheeraz, A. "Urdu language text summarization using machine learning," Journal of Computing & Biomedical Informatics, 2024, vol. 8, no. 1, pp. 1-10.
40. Akhtar, M.U.; and Khalil, M.I.K. "Link prediction techniques in complex networks," International Review of Basic and Applied Sciences, 2018, vol. 6, no. 8, pp. 60-69.
41. Taj, A., and Khalil, M.I.K. "DDOS defense mechanism and challenges," International Review of Basic and Applied Sciences, 2018, vol. 6, no. 11, pp. 86-93.
42. Khalil, M.I.K., Taj, A., Sadeeq, J. "Selection of cluster head in mobile Ad hoc networks on the basis of battery power," City University Research Journal, 2013, vol. 3, no. 1, pp. 131-139.

43. Khalid, U.; Khalil, M. I. K..; Nawaz, A.; Khan, I.Z.; Aimal, M.M.; and Ahmad, S. "Experimental Analysis of Algorithms for Community Detection," Journal of Computing & Biomedical Informatics, 2024, vol. 08, no. 1, pp. 272-281.