ASIAN BULLETIN OF BIG DATA MANAGEMENT
http://abbdm.com/
ISSN (Print): 2959-0795
ISSN (online): 2959-0809

Bridging the Language Gap: Evaluating Text Data Pre-processing and Classification Techniques in Urdu Sentiment Analysis

Mohammad Bilal, Mahira Zainab, M Ramzan Shahid Khan, Ali Raza, Sonia Shehzadi, Faiz-ur-Rehman, Asif Raza,

Chronicle

Article history

Received: June 2, 2024
Received in the revised format: June 16, 2024
Accepted: June 20, 2024
Available online: June 23, 2024

Mohammad Bilal, & M. Ramzan Shahid Khan are currently affiliated with the Department of Computer Science, Namal University Mianwali, Pakistan.
Email: muhammad.bilal@namal.edu.pk
Email: ramzan.shahid@namal.edu.pk

Mahira Zainab is currently affiliated with the Department of Computer Science and Information Technology, University of Mianwali, Mianwali, Pakistan.
Email: mahiraz332@gmail.com

Sonia Shehzadi is currently affiliated with Shaanxi Normal University Xian, China.
Email: soniarana846@gmail.com

Ali Raza & Asif Raza are currently affiliated with the Department of Software Engineering, University of Mianwali, Mianwali, Pakistan.
Email: aliraza.1354@gmail.com
Email: asifraza@umw.edu.pk

Faiz-ur-Rehman is currently affiliated with the Department of Computer Science, MY University, Islamabad, Pakistan.
Email: faiz.rehman@myu.edu.pk

Corresponding Author*

Abstract

The advancement of the internet and rapid growth in social media platforms such as Facebook, X , Instagram and others which made it possible for information on review on goods, reaction on events, services from experts, and political beliefs to be easily and globally shared within no time. Due to this fast propagation of information may have both positive and negative impact on thinking and ability to take decision on a particular event or anything related. Due to the generation of such a huge number of date and while the volume of studies directed on slant investigation is quickly extending, these examinations for the most part address English language concerns. But the English the not the only language by which data is got propagated but other such as Urdu, Farsi, Hindi, Pasto, branch etc. The main goal of this research study is to critically assess the data related to Urdu and the Urdu sentiment analysis's progress and problems and present solutions. After critically reviewing the literature and related work following two research directions are identified in accordance with our study— the first one is text pre-processing and the other is sentiment classification. Here, we had the opportunity to describe the progress that has been made in this research field. The pre-processing steps include Tokenisation, stop words removal, word segmentation, text cleaning, transforming into numeric vectors and others then applying machine learning models i.e. regression and SVM. Result is compared out for each of the model use. After thorough investigation, the following results demonstrated that the LDA-TFID model achieved the highest accuracy at 0.923977 as compared to LDA-BOW and SVM-BOW and also outperforming both LDA-BOW and SVM-BOW in term of efficiency and accuracy. This indicates the effectiveness of TF-IDF in enhancing model performance for Urdu sentiment analysis.

Keywords: Urdu sentiment analysis, Pre-processing, Machine learning models, classification, Word segmentation.

INTRODUCTION

The popularity of social media platforms has made it possible and encouraging for ideas and information on products, policies, resources, and problems to be widely shared. On social platforms the sharing of data has prompted the improvement of innovative apparatuses to take right decision-making by people and firms (P.Marques, 2014). The English language possesses a wealth of resources for sentiment analysis (SA), including lexicons, parsers, part-of-speech taggers, and numerous natural language processing (NLP) tools. Although a significant percentage of existing SA research has focused on the English language, the increasing online presence of non-English languages has led to the development of SA methods for languages other than English (VK, 2015). However, despite the abundance of literature on SA strategies, challenges, and applications in English, there is a scarcity of research on SA in the Urdu language (W, X, & XL, 2006). The Majority of SA tools proposed for English cannot be applied to language Urdu due to its distinct script and morphological structure, which are unique and differ significantly from those of the English language. As Urdu online resources continue to grow in popularity, there is a need for language-specific lexical resources and SA techniques that can effectively handle the complexities of the Urdu language. This evaluation objects are to provide an summary of the present-day state of text processing, AI models, and opinion analysis techniques for the Urdu language, highlighting the current challenges and research gaps in Urdu language SA (Dashipor K, Hussain A, & AY, 2016).

Urdu, being the national language of Pakistan and broadly spoken in the Indian subcontinent, has a significant online presence (Abdalla RM, 2006). The increasing amount of Urdu text data on the web calls for the development of SA tools that can accurately analyze the sentiment and opinions expressed in Urdu text. However, the unique characteristics of the Urdu language, such as its script, grammar, and vocabulary, pose significant challenges to SA. For instance, Urdu has a complex system of suffixes and prefixes that can change the meaning of words, making it difficult to develop accurate SA models (Afraz SZ, Jan R, & Mirza). Furthermore, the lack of standardization in Urdu text data, including variations in spelling and dialects, adds to the complexity of SA in Urdu (Ali AR, 2009). To overcome these challenges, there is a need for language-specific SA resources, including annotated datasets, lexicons, and AI models that can handle the unique characteristics of the Urdu language (Hazrat Ali, 2015). Several studies have attempted to develop SA tools for Urdu, but these efforts have been limited by the availability of resources and the complexity of the language. For example, one study developed a Urdu SA dataset, but it was limited to a small size and scope.

Another study proposed a machine learning approach for Urdu SA, but it was not evaluated on a large-scale dataset. To tackle these limitations, there is a challenge to develop a comprehensive approach for Urdu SA that includes the development of large-scale annotated datasets, language-specific SA models, and evaluation metrics that can accurately measure the performance of SA tools in Urdu. Furthermore, there is a need for collaboration between researchers, industry experts, and policymakers to develop standards and guidelines for Urdu SA that can facilitate its applications in real-world scenarios (Khan W, 2018). By highlighting the challenges and research gaps in Urdu SA, this review aims to inspire future research in this topic and advance contribution to the development of accurate and effective SA tools for the Urdu language.

LITERATURE REVIEW

Urdu Sentiment Analysis stays in early phases of development as contrasted to other languages that are resource rich like English etc. (VK, 2015). Besides, confined work has been developed, which clearly impacts the quantity of studies and reviewed articles right now accessible. Anwar et al. (Hazrat Ali, 2015) shows summary of procedures concentrating on the development of Urdu corpus in their review on automated Urdu language processing. Several linguistic methods were developed, for instance part of speech tagging , parsing, and named entity recognition. However, this survey lacked the appropriate strategies needed for developing sentiment analysis in languages like Urdu, includes in major findings of that investigation. Daud et al. (Khan W, 2018) reviewed various linguistic sources and pre-processing methods in Urdu language processing, presenting effective means for several tasks, for example identification of boundary of sentence, tokenization, POS tagging, NER, and the development of WordNet lexicons. However, the overview failed to draw attention to the sentiment analysis paradigm. Consequently, a detailed analysis based on sentiment analysis is required. Singh (VK, 2015) presented subjective and sentiment analysis-based classification in survey of Urdu Sentiment analysis.

In any case, in this study, the main research focused on utilizing Urdu sentiment analysis by evaluating three aspects, specifically: (I) text pre-processing, (ii) lexical resources, and (iii) sentiment analysis, that are additionally partitioned into various types. Khan et al. (Khan W, 2018) performed an overview on Urdu sentiment analysis by summarizing fourteen plus published journal and conferences articles on sentiment analysis of the language Urdu. The Machine Learning based algorithms for the language Urdu SA were used based on lexicon, and hybrid approaches [10]. However, to cover all aspects of Ursu SA, detailed survey is required that can cover Urdu SA current challenges and their solutions. Lo et al. (Lo SL, 2017) conduct a comprehensive study on multilingual sentiment analysis with major focus languages on scarce and limited resource.

Various techniques and available tools are explored and observed for performing multilingual sentiment analysis. Besides, various difficulties are distinguished alongside suggestions for future directions. LDA-BOW combines Latent Dirichlet Allocation (LDA), a topic modeling technique, with the Bag of Words (BOW) approach (Li, 2017).

LDA-TFIDF integrates LDA with Term Frequency-Inverse Document Frequency (TF-IDF), a technique that adjusts the word frequency matrix by considering the significance of words in a text-based document comparative to the entire corpus (Li, 2017). Supervised machine learning method SVM that can be applied for either classification or regression challenges [12]. SVM generate results on classification by adjusting the hyper-plane that separates the classes visualized in n-dimensional space (Ray, 2017) (Asif Raza, 2021) (Farooq Ali, 2022).

CHALLENGES OF SENTIMENT ANALYSIS IN URDU LANGUAGE

This section highlights the challenges in Urdu Language Sentiment Analysis (ULSA). The unique characteristics of the language Urdu necessitate advanced techniques for measuring SA. For instance, Urdu language script is written from right side to left side, and the form of alphabetic characters varies depending on their position within word. Urdu features a variety of stop words, which can lead to grammatical errors when handling them. For example, کہ is sometimes mixed up with پہ, کے with پے, and other similar-sounding words that carry distinct meanings. In language Urdu writing, the use of spacing is challenging, often leading to issues with excessive letterspacing or exclusion of spaces within words. For example, the Urdu word "انکا" is actually a combination of two separate words, but it is often treated as a single word (Daud A & D, 2017). The space insertion problem is tackled by a two-stage procedure. The limited availability of datasets and corpora for sentiment analysis in Urdu is a significant challenge. Data are normally extracted from social networks, online forums, and newspapers (W. Khan, 2016) (Ali Raza R. S., 2024). Lexicon-based sentiment analysis in Urdu is hindered by the limited availability of sentiment-annotated lexicons and corpora. While online resources are available, they are limited, and most data on Urdu websites are in graphics format e.g. images etc. , making them difficult to retrieve (W. Khan, 2016) (Ali Raza S. U., 2024). Furthermore, publicly available corpora for Urdu language sentiments are limited, and few lexicons have been generated, with most not being openly available (Khan W, 2018).

Identifying target wrods and opinion words is crucial for insightful opinion mining and sentiment analysis. However, in Urdu, nouns can be used as opinion targets, and adjectives as opinion words, making it challenging to detect opinion targets and words (Khairullah Khan, 2018). Moreover, the frequent occurrence of English words in Urdu text, especially in online platforms, requires another preprocessing step to identify, clean, and transform such text into standard Urdu script (Bajwa, 2016).

Detecting Opinion Spam

The problem of opinion spam is marked in Urdu sentiment analysis. Opinion spam contains false sentiments used to mislead users, frequently employed by organizations or businesses for promotional purposes.

Feature Extraction

Urdu language text is normally unstructured, putting significant challenges for morphological analyzers and POS taggers in language Urdu processing.

Segmentation

The segmentation problem can be additionaly classified into the following a) space-inclusion and b) space-deletion issues. Such as, a single word may contain a space within it, as in "خوب صورت" (khoob surat, beautiful) (Khairullah Khan, 2018). On the other hand, a blank space between 2 distinct words may be excluded, as in "دستگیر" (dastgeer, benefactor) (Khan W, 2018). Sentiment analysis is crucial for understanding public opinion on various topics, enabling businesses and legislators to implement informed decisions based on real-time data from social media and other online platforms. However, sentiment analysis for the language Urdu presents unique challenges because of its rich morphology, diverse dialects, and lack of extensive annotated datasets and language-specific resources, making it difficult to accurately capture and interpret sentiment nuances (Liaqat, 2022).

METHODOLOGY

This section describes the methodology used for Urdu SA. The way toward planning a functional classifier for Urdu SA can be separated into following classifications:

Methodology Flow Diagram for Urdu Sentiment Analysis

The initial step involves preprocessing. In this level a progression of steps is performed e.g. removing stop words, Words Tokenization, Lemmatization. So, the linked issues must be responded in pre-processing step.

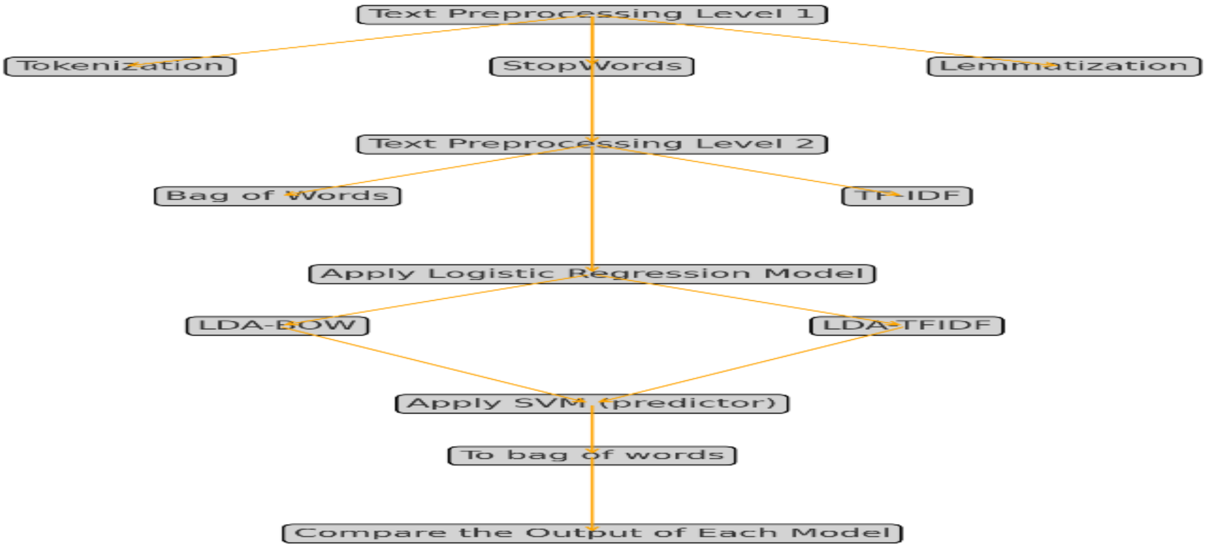
Text Preprocessing Level 1
Tokenization
StopWords
Lemmatization
Text Preprocessing Level 2
Bag of Words
TF-IDF
Apply Logistic Regression Model
LDA-BOW
LDA-TFIDF
Apply SVM (predictor)
To bag of words
Compare the Output of Each Model

Figure 1.
Methodology Flow Diagram for Urdu Sentiment Analysis

Text Preprocessing Level 1-

Tokenization
StopWords
Lemmatization

Text Preprocessing Level 2-

Bag of Words

TF-IDF

Apply Logistic Regression model

- LDA-BOW
- LDA-TFIDF

Apply SVM (predictor)

- To bag of words

COMPARE THE OUTPUT OF EACH MODEL

Text Preprocessing Level 1:

- Tokenization: It includes splitting Urdu text stream into words, symbols, and other meaningful parts identified as 'tokens.' Tokens can be split by blank white-space characters or punctuation marks. This process allows us to consider tokens as separate segments that compose a tweet.

- Stop Words: Stop words are conventional words that contain no information while utilized into text and consequently declared to be worthless. Examples include “a”, “an”, “the”, “he”, “she”, “by”, “on”, etc. It is suitable and useful sometimes to eliminate these words since they contain no extra data meanwhile they are utilized similarly in all text classes, for instance when deciding prior-sentiment-polarity of words in X tweets based on their frequency count in dissimilar classes.

- Lemmatization: The process of converting a word to its root form. For grammatical reasons, reports will utilize various type of a word, such as *play*, playing, and *played*. Similarly, there are variety of derivationally related words with reasonable implications, like *democracy*, *democratic*, and *democratization*. Much of the time, possibly it would be valued for a quest for one of these words to return archives that consist of alternative word in the set. The main purpose of stemming and lemmatization is to diminish articulation structures and occasionally derivationally related types of a word to a typical root structure.

Feature Engineering

- Since AI models don't acknowledge the crude content as input data, we need to change "Reviews" into vectors of numbers.
- There are various ways of converting text into numeric vectors.
- In this analysis, applied first Bag of Words, followed by Bag-of-n-Grams, and later moved to Tf-Idf which is a more complex representation.

Text Preprocessing Level 2

- Bag of Words: NLP based method Bag of words is used for text modeling.
- A bag of words is a representation of text that depicts the count of word with a document called word frequency. We simply monitor word checks and dismissal the linguistic subtleties and the word request.
- TF-IDF: TF-IDF (term frequency-inverse document frequency) is a statistical process that examines significance of a word of a document in collection of documents. Two Metrics Multiplication is used to compute that, word occurrence count, (TF) and the inverse document frequency of the word throughout a set of documents (IDF).

Apply Logistic Regression model:

Logistic Regression is a Supervised learning approach that predicts the probability of a categorical dependent variable. In logistic regression, the binary dependent variable possesses data coded as 1 (yes, success, etc.) (Li, 2017). in the same way, the logistic regression model using function of X predicts P(Y=1).

- LDA-BOW: A topic modeling technique, with the Bag of Words (BOW) approach is applied on data set.
- LDA-TFIDF: A technique that adjusts the word frequency matrix by considering the importance of words in a document relative to the entire corpus is applied on data set.

Apply SVM (predictor)

The "Apply SVM (predictor) to Bag of Words" step involves using a Support Vector Machine (SVM) algorithm to classify sentiments based on the Bag of Words representation of the text data.

RESULT AND DISCUSION

We have preprocessed the text data using urduhack library and applied multiple algorithms on this dataset. we evaluated the performance of three different models on our dataset: LDA-BOW, LDA-TFID, and SVM-BOW. The accuracy results for these models are as follows:

LDA-BOW: Achieved an accuracy of 0.9188. This model leverages Latent Dirichlet Allocation (LDA) for topic modeling combined with a Bag of Words (BOW) representation.

LDA-TFID: Showed a slight improvement over LDA-BOW with an accuracy of 0.9240. This variant utilizes LDA with Term Frequency-Inverse Document Frequency (TF-IDF), indicating that weighting terms by their importance enhances model performance.

SVM-BOW: Recorded an accuracy of 0.9103. This model employs a Support Vector Machine (SVM) classifier with a BOW approach, which, while slightly less accurate than the LDA-based models, still performs competitively.

Among the three, LDA-TFID emerged as the most accurate model, suggesting that the TF-IDF approach in representing textual data can capture more significant features than the traditional BOW model. The LDA-BOW model also demonstrated strong performance, validating the efficacy of topic modeling in text classification tasks. The SVM-BOW model, despite having the lowest accuracy, remains a robust choice, reflecting the versatility and effectiveness of SVM classifiers in handling high-dimensional text data. These results underline the importance of choosing appropriate text representation techniques to enhance model accuracy in natural language processing task. After preprocessing using Urdu hack, the accuracy falls down. So we conclude that Urdu hack text processing preprocess the content very well yet in addition influences the accuracy as shown in Table 1.

Table 1.
accuracy of Different Sentiment Analysis Models

	Models	Accuracy
0	LDA-BOW	0.918803
1	LDA-TFID	0.923977
2	SVM-BOW	0.910256

The graphical representation of machine learning models on Urdu datasets is shown in Figure 2 and 3:

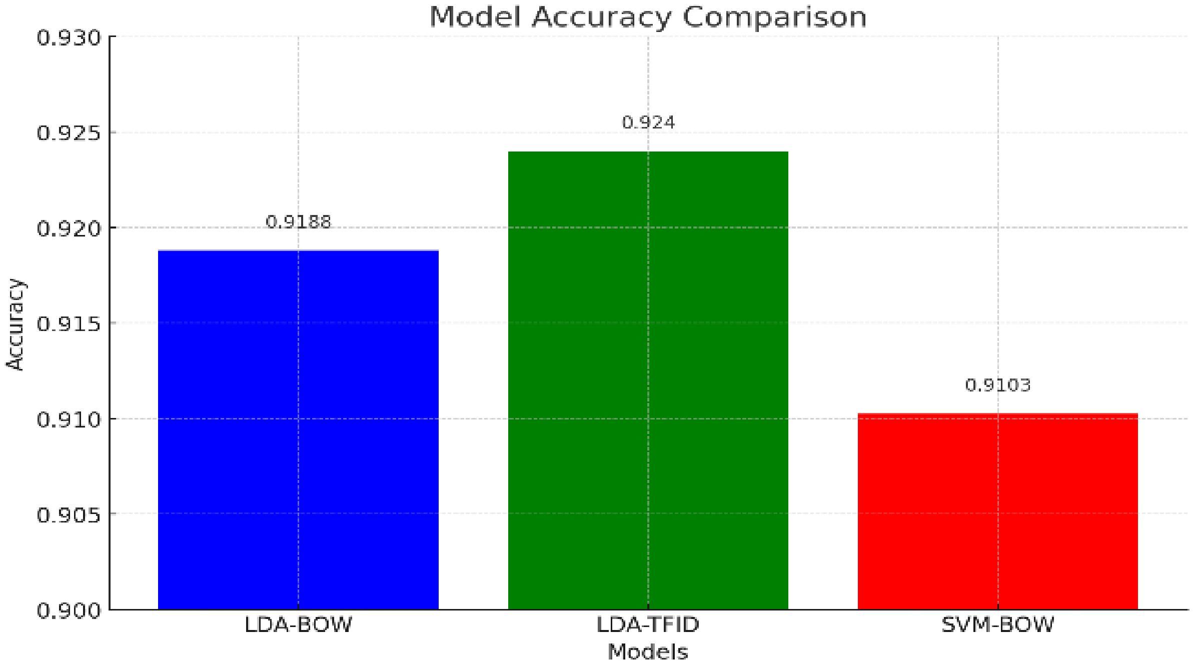

Figure 2.
Comparison of Model Accuracy

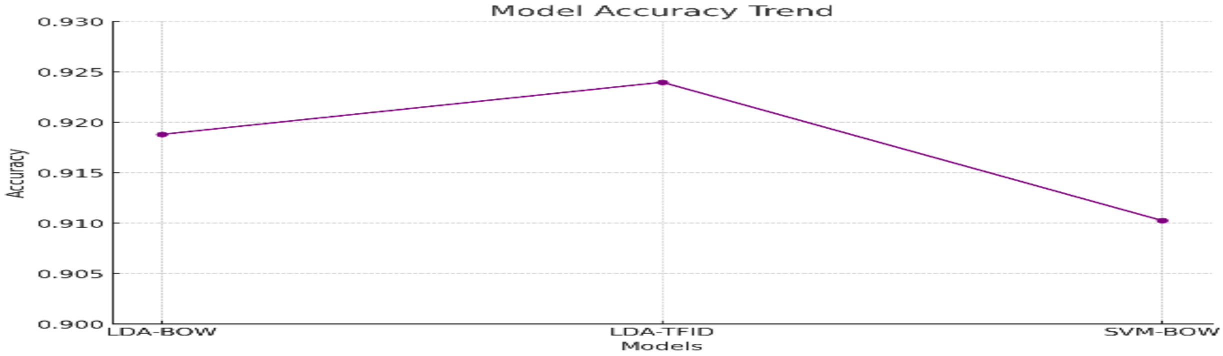

Figure 3.
Model Accuracy Trends

CONCLUSION

This research emphasizes the availability of various possible techniques and methodologies yet reveals a restricted focus on Urdu sentiment analysis. Social media appears as a significant platform where massive texts are published daily, offering valuable insights into public opinions on various subjects like services, products, and trending celebrities. The basic aim of this research is to provide an overview of current advancements in sentiment analysis and classification algorithms applied to the Urdu language. While successful in predicting sentiment on Tweets using three different approaches—Bag of Words, Support Vector Machine, and Tf-idf—we also gained substantial insights into machine learning. Future work will emphasis on collection the dataset to encompass diverse domains such as politics, healthcare, and entertainment, sourced from various social media platforms. We aim to integrate deep learning models like LSTM, GRU, and BERT to enhance sentiment analysis performance. Additionally, we'll employ data augmentation techniques to bolster training data variability and robustness and optimize model architectures through hyperparameter tuning and transfer learning. Advanced feature extraction methods, including word embeddings and domain-specific lexicons, will be explored. Furthermore, real-time sentiment analysis capabilities and user feedback integration mechanisms will be developed for continuous model improvement. In future, we will extend our dataset further and will cover more domains. Besides, we additionally want to use deep learning approaches to solve this issue.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Abdalla RM, T. S. (2006). A bootstraping approach to unsupervised detection of cue phrase variants.

Afraz SZ, A. M., Jan R, S. T., & Mirza, W. (n.d.). Sentiment Analysis of a Morphologically Rich Language. 2, 69-73.

Ali AR, I. M. (2009, DEC). Urdu text classification. In Proceedings of the 7th internationalconference on frontiers of information. 21.

Ali Raza, R. S. (2024). A Hybrid Deep Learning Based Fake News Detection System Using Temporal Features. The Asian Bulletin of Big Data Management, 4.

Ali Raza, S. U. (2024). The Impact of Recurring Events in Fake News Detection. Journal of Xi'an Shiyou University, Natural Science Edition.

Asif Raza, F.-U.-R. B. (2021). Comparative analysis of machine learning algorithms for fake review detection. International Journal of Computational Intelligence in Control, 217-2023.

Bajwa, Z. U. (2016). "Lexicon-based sentiment analysis for Urdu language" in Innovative Computing Technology (INTECH). Sixth International Conference on, 497-501.

Dashipor K, P. S., Hussain A, C. E., & AY, H. (2016). Multilingual Sentiment Analysis .

Daud A, K. W., & D, C. (2017). Urdu Language Processing: A survey. Artif Intell Rev.

Farooq Ali, A. R. (2022). Ontological automation of software essence kernel to assess progress of software project. Mehran University Research Journal of Engineering & Technology, 41, 135-145.

Hazrat Ali, N. A. (2015). Automatic speech recognition of Urdu words using linear discriminant analysis. J. Intell.

iaqat, M. I. (2022). Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. PeerJ Computer Science, 8, e1032.

Khairullah Khan, A. U. (2018). Urdu Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 9.

Khan W, D. A. (2018). Urdu part of speech tagging using conditional random fields. Language Resources and Evaluation. 1-32.

Li, S. (2017, september). Building A Logistic Regression in Python, Step by Step.

Lo SL, C. E. (2017). Multilingual sentiment analysis: from formal to informal and scarce resouce language. Artif Intell Rev, 499-527.

P.Marques, P. (2014). techHnologIcal Platform and sharing construction of communication.

Ray, S. (2017, SEPTEMBER). Understanding Support Vector Machine(SVM) algorithm from examples (along with code).

VK, S. (2015). A survey of sentiment analysis research in urdu. Ind J Sci Res Tech .

W, A., X, W., & XL, W. (2006, Aug). A Survey of Automatic Urdu language processing. In Machine Learning and Cybernetics, 2006 International Conference on, 13.

W. Khan, A. D. (2016). Urdu Named Entity Dataset for urdu Named Enity Recognition Task. in 6th International Conference on Language & Technology, 51-52.