JCBI
Journal of Computing & Biomedical Informatics Volume 08 Issue 02
ISSN: 2710 - 1606 2025

Research Article
https://doi.org/10.56979/802/2025

A Framework for Sarcasm Detection Incorporating Roman Sindhi and Roman Urdu Scripts in Multilingual Dataset Analysis

Majdah Alvi[1,*], Muhammad Bux Alvi[1], and Noor Fatima[1]

[1]Department of Computer Systems Engineering, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[*]Corresponding Author: Author's Name. Email: majida.alvi@iub.edu.pk

Received: February 03, 2025 **Accepted:** February 28, 2025

Abstract: Sarcasm detection is imperative for successful real-time sentiment analysis in the pervasive social web. Detection of sarcastic tones expressed through text that impart bitter, satirical, or mockery expressions, remarks, or derision in Natural Language Processing (NLP) is problematic to handle for humans; making it automated is even more arduous. This work aims to propose a sarcasm detection framework tflexihat optimizes a sentiment analysis system by correctly detecting sarcastic text messages for resource-poor languages in multilingual datasets. The techniques developed to date are inadequate and require precise training data. Therefore, we propose neural networks and deep learning-based models that focus on contextual information utilizing different word embedding techniques, and we further propose a framework for multilingual sarcasm detection resources for low-resource languages such as Roman Sindhi and Roman Urdu. With this sarcasm-aware framework, individuals with limited English proficiency will be better equipped to engage on social media using sarcastic tones, emojis, and creative linguistic variations in a multilingual textual data analysis.

Keywords: Multilingual Data, Sarcasm Detection; Sentiment Analysis; Roman Sindhi; Roman Urdu

1. Introduction

In the modern era, data deluge has given birth to many new fields of knowledge in engineering, science, and technology. The field of sentiment analyses under the umbrella of natural language processing is one of them. Sentiment analysis emerged in the first decade of the twenty-first century, having roots in text categorization [1]. Text categorization has been a topic of interest for a long time, but sentiment analysis gained the attention of the research community after the exponential rise of computational power, storage capacity, and online social networks. Sentiment analysis uses human-generated text data to gain insight into their mood and behavior regarding specific events, products, policies, or services [2]. Sometimes, users may record their reactions to a given entity weirdly. Such reviews or comments usually transform the orientation of an apparently positive or negative sentence into its opposite. Such comments, if not appropriately attended, may affect the efficiency of a sentiment analyzer. This scenario has given birth to a newer field of knowledge, i.e., Sarcasm detection.

Sarcasm is a way in which the text's literal meaning is different from the intended meaning. Sarcasm detection encompasses all computational methods that may successfully determine if the given piece of text is sarcastic. For example: *'What a cricket team Pakistan is! This time, too, they maintained their tradition!'* Such statements should be correctly identified as sarcastic; otherwise, they may negatively impact the overall analysis. Sarcasm detection is an open challenge and a difficult task because of subtle ways of expressing sarcasm. A robust sarcasm detection model should be able to handle challenges that are multifold, such as:

- Difficult to extract valuable datasets from the plethora of data

- Challenges related to free text processing, mainly social media users' improvised writing styles
    - Handling exaggerated terms used in sarcastic text
- Identification of sarcasm from a given piece of text;
    - Sarcasm detection requires contextual meaning
    - It needs a domain and common-sense knowledge

Various features have been used to identify sarcasm, including lexical, punctuation, syntactic, sentiment, semantic, pattern-based, and behavioral data. In addition, deep learning and machine learning methods are used for sarcasm detection to improve accuracy and performance, showing the potential of neural networks for the task of sarcasm detection [3-5]. Figure 1 shows the adopted approaches to perform sarcasm detection in detail.

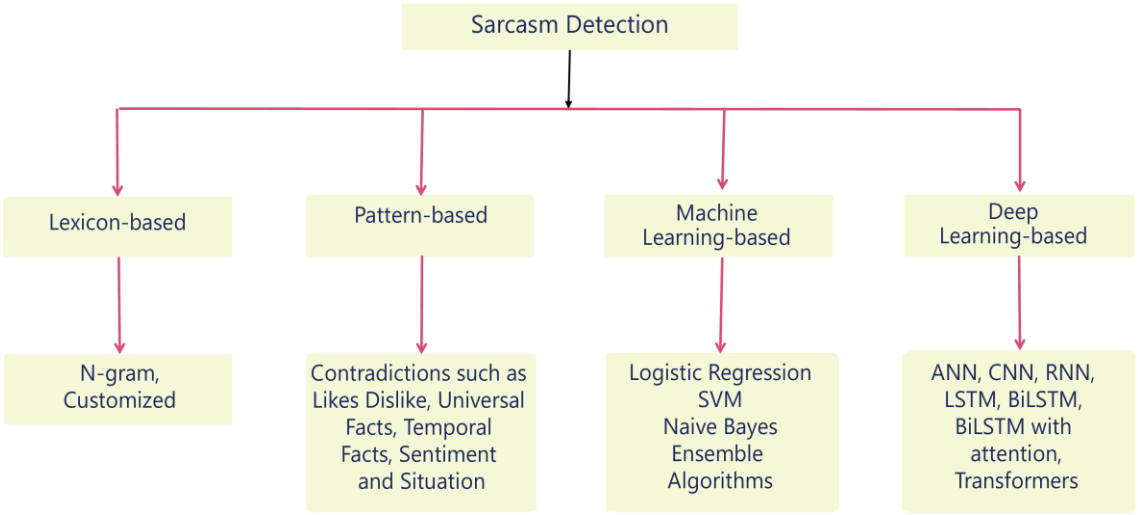

Figure 1. Approaches to Sarcasm Detection

English, a resource-rich language, is widely used on social networks to express public views. Much research has been conducted on sarcasm detection in English [6-8]. However, most people are not native English speakers; therefore, they prefer to communicate in languages other than English. A study of the languages most commonly used for exchanging information on social media showed that approximately 50% of the posts were written in languages other than English [9].

Other statistics showed that around 26% of the Pakistani population is bilingual. Such user diversity gives rise to user input code-mixing [10, 11]. Code mixing often comprises user text inputs that combine English script with other language inputs, creating a research gap. Hence, the large gap needs to be fulfilled by developing resources for poor resource languages such as Roman Urdu and Roman Sindhi. Although, Alvi et al. classified sentiment polarity (positive, negative, or neutral) in low-resource language contexts. The study employed sentiment classification techniques to analyze the Roman Sindhi text through linguistic preprocessing and machine learning; there remains a significant gap in addressing more complex challenges, such as sarcasm detection, in these languages [2]. Another similar effort was fobused on Roman Urdu sentiment analysis [12]. Again, without handling sarcastic expressions. This study proposes a framework for detecting sarcasm in a dataset that encompasses multiple languages such as Roman Sindhi, Roman Urdu, and English. Furthermore, this work will encompass neural models and deep learning-based frameworks to detect contextual sarcasm in free text used in social media.

The current research study will help an estimated 33 million Sindhi-speaking and more than 64 million Urdu-speaking people in the world [13]. This work will facilitate almost 100 million people to cast their voices on any topic using social media, which otherwise was uncounted.

Section I introduced the study. Section II describes the literature work, following by materials and methods presentation in section III. Section IV elaborates on the significance of the work, whereas section V concludes the article. At the end, a list of cited literature work is provided.

2. Literature Review

Progress in the area of sarcasm detection is underway. The literature in the field of sarcasm detection is presented in organized paragraphs, i.e., feature-based methods, contextual-based, corpora-based approaches, neural network models and deep learning-based techniques, and Bilingual and Multilingual-based Approaches.

2.1. Feature-Based Approaches

Sarcasm detection is a classification problem. Research is centered around effective feature extraction methods. Kreuz and Caucci studied lexical features and reported that punctuation and interjections are effective [14]. Carvalho et al. proved that emoticons and emojis were powerful indicators of sarcasm [15]. Both studies relied on unigram word representation. The authors in [8] employed syntactic patterns with the Amazon product reviews dataset. They proposed pattern-based features by classifying words into high-frequency words and content words. S.K. Bharti considered hyperbolic features (intensifiers and interjections) to detect sarcasm [11]. Riloff et al. identified a sarcasm type that used sentiment information of positive and negative sentences to detect sarcasm [16]. Băroiu et al. presented milestones in sarcasm detection research, such as semi-supervised pattern extraction, the use of hashtag-based supervision, and the incorporation of context beyond the target text [17]. These are other popular research studies related to feature-based sarcasm detection [18, 19].

2.2. Contextual Based

The use of sarcasm detection relies on context. Context refers to all background knowledge and common-sense information beyond the given text (target text). Contextual information has been exploited for sarcasm detection in [5, 7]. They reported the effectiveness of contextual features extracted from historical tweets for tweet sarcasm detection. Various factors have been reported to define contextual information to improve classification accuracy, such as user profiles, topics, social influence, temporal changes, domain knowledge, etc. Bamman and Smith considered both lexical and contextual features to detect sarcasm by using author and audience information and their interaction on social media platforms [6]. Akshi Kumar in [20] reported that contextual information undoubtedly improved computational models for sarcasm detection. They used content-based local and global contextual information to build predictive models for sarcasm detection. Liang et al. introduced a novel cross-modal graph architecture for multi-modal sarcasm detection to clearly illustrate the ironic relationships between textual and visual modalities [21].

2.3. Corpus-Based Approaches

Resource development is always one of the starting points in research in each field of knowledge, which initiates with manual annotation. Filatova built a corpus using crowd-sourcing for sarcastic sentences [22]. Davidov et al. studied the influence of hashtags [8]. Inspired by the work of Dividov et al, Gonzalez-Ibanez et al. used hashtags as an indicator of sarcasm and manipulating tweets without hashtags as non-sarcastic [18]. Both works were similar to Go et al.'s, who considered emoticons as gold sentiment indicators [23]. Ptáček et al. adopted the method used in [16] to create a sarcasm detection dataset for Czech, consisting of 2,450 data points [9], providing a leading path for other similar studies and developement. Similar dataset can be developed for Roman Sindhi and Roman Urdu to classifier sarcastic and non-sarcastic expression. S. Castro proposed a new dataset made up of audiovisual utterances labeled with sarcasm from famous TV shows to identify sarcasm [24].

2.4. Neural Network and Deep Learning Models

Less attention has been given to sarcasm detection using neural networks compared to their capacity, such as automatic feature induction. Neural networks can discover subtle semantic patterns automatically. However, an increasing trend has been observed for the application of neural models in sentiment analysis. The field of sarcasm detection can benefit from such induction, and several works have already attempted.

Such as authors in [3, 5] claimed to be the first to investigate the effect of neural networks on sarcasm detection. Deep learning approaches can generalize well by automatically learning from data. Sarcasm detection is topic-dependent and highly contextual; therefore, sarcasm detection is one of the problems to be better solved by deep learning techniques.

Amir et al. applied convolution operations on user embeddings and utterance embeddings for sarcasm detection. D Hazarika et al. introduced a Contextual Sarcasm Detector called CASCADE, which leveraged both content and contextual information for the classification. They used a CNN-based textual model to obtain optimal performance on a large-scale Reddit corpus [25]. Sarcasm detection using soft attention-based BiLSTM and CNN was proposed by A. Kumar et al. [26]. M. Zhang [4] proposed neural networks for tweet sarcasm detection by using a bi-directional gated recurrent neural network and a pooling neural network to capture syntactic and semantic information and contextual features, respectively. In [27], the authors proposed using Transformers to detect sarcastic text. State-of-the-art transformer architectures, such as RoBERTa and GPT, have made significant progress in sarcasm detection by successfully capturing subtle semantic nuances and long-range dependencies. A pre-trained RoBERTa model enhanced by a three-layer feed-forward neural network achieved an F1 score of 0.526 on the iSarcasm dataset [28]. Likewise, Gole et al. assessed zero-shot and fine-tuned GPT models (such as GPT-3, GPT-3.5, and GPT-4), yielding an F1 score of 0.81 on the balanced and political sections of the Self-Annotated Reddit Corpus (SARC 2.0) [29].

2.5. Bilingual and Multilingual based Approaches

In the literature, most polarity analysis studies are limited to the English language, but in opinionated user-generated textual data in social media, it is no longer sufficient to extract just English language content for analysis purposes. The authors in [30] used deep learning approaches for classifying Urdu and Roman Urdu writings for offensive language style and intention detection. In contrast, Akshita Aggarwal in [31] tried to determine sarcastic sentences and their orientation in code-mixed tweets using bilingual word embeddings. Lo, Siaw Ling [32] adopted a multilingual semi-supervised approach for polarity detection in the scarce-resource language Singlish (Singaporean English), contributing to constructing Singlish NLP resources and toolkits. Sarcasm detection in languages other than English has been under the researcher's radar [33-35]. In [36], the authors compared the performance of humans and machines on sarcasm detection using first-party labels for English and Arabic texts, which showed that both struggled with the subjectivity and context-dependent aspect of sarcasm.

3. Materials and Methods

This study aims to describe a framework to optimize a sentiment analysis system by correctly detecting sarcastic messages for resource-scarce languages in a multilingual dataset. The framework targets to develop resources for Roman Sindhi and Roman Urdu languages and select the most suitable deep learning-based methods to cope with the scaling system. The proposed framework is discussed in the following sub-sections, as illustrated in Figure 2.

3.1. Dataset Acquisition

People use X (formerly Twitter) as a de facto platform to register their perspectives. However, other online social networking platforms are also popular for exchanging views using text data [37]. The dataset for this work will be acquired partially from online repositories and partially from social platforms using public APIs. The data collected through APIs will be topical and general. The final dataset will be a resultant effort through a heterogeneous raw data collection approach. The seed dataset may be manually supervised for labeling.

3.2. Data Annotation, Augmentation, and Analysis

The acquired dataset will be annotated carefully, optionally, with the help of linguistic experts [38]. Since the aim is focused on multilingual text data, a careful annotation procedure is proposed to handle mixed-mode text data entries. Another optional step at this stage is to create a balanced-class dataset, a significant step to handle imbalanced datasets. Exploratory data analysis (EDA) often becomes very handy

by elaborating on the dataset and providing helpful insight. EDA may be performed statistically or graphically, such as with a boxplot, which uses interquartile range (IQR) to help detect dataset outliers. In addition, a boxplot may determine the data spread and any dataset skewness.

3.3. Data Preprocessing

Data preprocessing is an essential intermediate step for any text-processing project.

3.3.1. Data Cleaning

Text data is usually highly prone to errors and noise. Furthermore, text data comes from heterogeneous sources having different characteristics. Therefore, text preprocessing becomes an important step in the classification pipeline. The classification efficiency relies on the data quality, and data quality can be improved by accurately cleaning up the text data. Text data cleaning steps include cleaning special characters, removing punctuation, numbers, misspellings, acronyms, short forms, neologisms, contractions, etc.

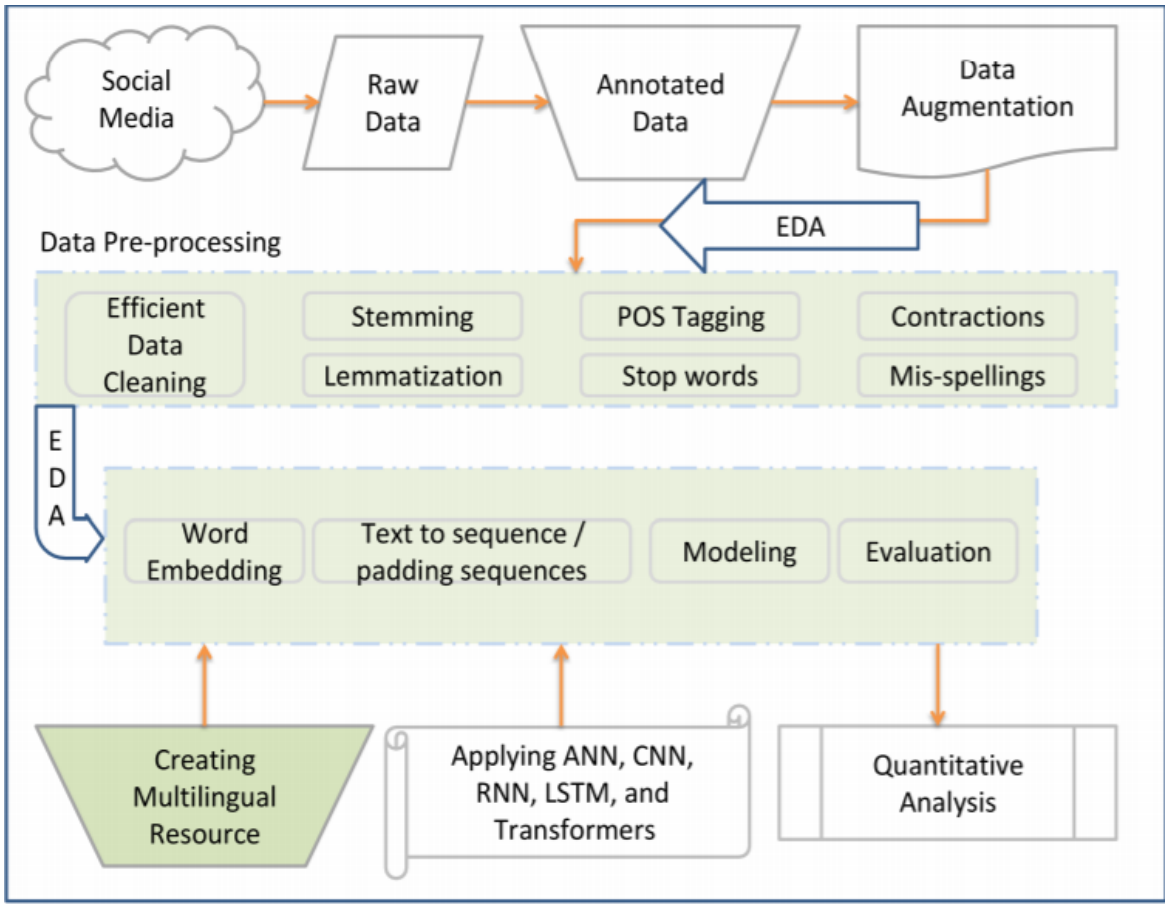

Figure 2. Proposed Sarcasm Detection Framework for incorporateing Roman Sindhi and Roman Urdu

Alvi et al. employed multiple steps for text processing, including iterative data cleaning (handling web links, missing values, and stopwords), tokenization, data normalization (spelling correction and contraction handling), and data preprocessing (stemming/lemmatization) to enhance feature representation [39]. In the case of sarcasm detection, a few punctuation marks are important. Thelwall et al. discussed the significance of punctuation. They studied the effects of single or repetitive punctuation marks [40]. They reported that punctuation was key in boosting the sentiment score orientation. Therefore, utmost attention is required while preprocessing. A two-way strategy is required: 1) to get rid of unnecessary, irrelevant, and insignificant text, and 2) to keep intact all those tokens that are significant.

3.4. Multilingual Resource Building: Bilingual Word Embedding

3.4.1. Primer on word embedding

Computers are good at representing, understanding, and manipulating numbers. Therefore, the text data tokens must be transformed into numbers before their potential usage. One way to represent text data tokens (words) is by leveraging one-hot encoding of word vectors. Other methods include binary representation, frequency-based encoding of lexical features, hash vectorization, and weighted terms. However, all of these fail to capture the semantic and contextual relationship between the words in text data and exponential growth in feature vector space. The utilization of word embeddings is the solution to these inadequacies. The word embedding method represents the document vocabulary with the capability of capturing the context of a word, semantic similarity, and relationship with other words in the document. Word embedding is an n-dimensional length vector representation of a particular word.

3.4.2. Multilingual word embedding

More than 100 languages are being used on social media. There are established resources for resource-rich languages such as English, French, Deutsche, etc. There is an immediate need for the development of word embedding for resource-poor writing styles such as Roman Sindhi and Roman Urdu to provide a seamless layer for users to raise their voices and describe their feelings. In this work, we propose the usage of Roman Sindhi and Roman Urdu word embedding to represent text data and capture context. Furthermore, this representation maintains the property of keeping similar words close together in the vector space, which is the goal of word embeddings. For example, "*sutho*" in Roman Sindhi and "*good*" in English will appear nearer in the embedding vector space because both are semantically the same after applying language transformation method(s). a list of such word score can be found in [2].

3.5. Modeling and Validation

A multi-stage model selection strategy is proposed in this work. Initial stages will filter out amongst ANN, RNN, CNN, LSTM, Bi-directional LSTM, Bi-directional LSTM with attention, and Transformers (BERT) to determine the best technique. This initial analysis will serve as a baseline model. Afterwards, the selected method will be optimized using optimization techniques to find more accurate results for sarcasm detection to enhance the efficiency of a sentiment classifier. There are two ways to evaluate and validate the results of the developed model. 1) the usage of evaluation indexes; and 2) quantitative comparison. State-of-the-art evaluation indexes will be used to evaluate the output on the validation dataset. Since this work is the first of its kind, a direct quantitative comparison may not be possible. Hence, this work will be compared with similar work performed in other resource-poor languages.

4. Research Significant

It is estimated that there are 33 million Sindhi-speaking and more than 64 million Urdu-speaking people worldwide. Most of them use a multilingual writing style to express themselves. The model developed based on the proposed framework will allow almost 100 million people to cast their perspective on any topic using online social networks. Without such a model, the multilingual public views may go uncounted. In other words, this work is an effort to make existing sentiment analysis models more productive and accurate by counting and calculating Roman Sindhi and Roman Urdu sarcastic views. Developing a multilingual resource is a milestone because such work has not been reported in the literature. Eventually, posts, tweets, and views uttered in multilingual language styles (Roman Sindhi, Roman Urdu, and English) will not go unaddressed.

5. Conclusions

This study presents a robust and comprehensive framework for sarcasm detection in multilingual textual content derived from social media platforms. The proposed methodology leverages deep learning architectures, particularly neural networks, to effectively identify sarcastic expressions within English, Roman Sindhi, and Roman Urdu texts. By incorporating diverse linguistic features, the framework enhances the performance and generalization capability of multilingual sentiment analysis models. The inclusion of under-resourced languages addresses a critical gap, enabling broader participation of over 100 million speakers in digital discourse, whose contributions might otherwise be overlooked. To further strengthen the system's accuracy and applicability, the study advocates for the integration of BERT, a state-of-the-art

multilingual pretrained language model developed by Google, as a foundational component in handling resource-scarce languages.

Funding: "This research received no external funding."

Acknowledgments: We acknowledge the support of Mr. Iftikhar Hussain Bughio and Mr. Muhammad Yousif Shaikh for their academic support for this study. Both are Assistant Profesors in the College Education Department, Government of Sindh.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. B. Liu, *Sentiment analysis and opinion mining*: Springer Nature, 2022.
2. M. B. Alvi, N. A. Mahoto, M. S. A. Reshan, M. A. Unar, M. Elmagzoub, and A. Shaikh, "Count Me Too: Sentiment Analysis of Roman Sindhi Script," *SAGE Open,* vol. 13, p. 21582440231197452, 2023.
3. S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Natural language processing and sentiment analysis: perspectives from computational intelligence," in *Computational Intelligence Applications for Text and Sentiment Data Analysis*, ed: Elsevier, 2023, pp. 17-47.
4. M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, 2016, pp. 2449-2460.
5. S. Amir, B. C. Wallace, H. Lyu, and P. C. M. J. Silva, "Modelling context with user embeddings for sarcasm detection in social media," *arXiv preprint arXiv:1607.00976,* 2016.
6. D. Bamman and N. Smith, "Contextualized sarcasm detection on twitter," in *proceedings of the international AAAI conference on web and social media*, 2015, pp. 574-577.
7. A. Joshi, P. Bhattacharyya, and M. J. Carman, *Investigations in computational sarcasm*: Springer, 2018.
8. D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in Twitter and Amazon," in *Proceedings of the fourteenth conference on computational natural language learning*, 2010, pp. 107-116.
9. T. Ptáek, I. Habernal, and J. Hong, "Sarcasm detection on czech and english twitter," in *COLING 2014, the 25th International Conference on Computational Linguistics*, 2014, pp. 213-223.
10. K. Chandu, E. Loginova, V. Gupta, J. van Genabith, G. Neumann, M. Chinnakotla, *et al.*, "Code-mixed question answering challenge: Crowd-sourcing data and techniques," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 2018, pp. 29-38.
11. S. K. Bharti, R. Naidu, and K. S. Babu, "Hyperbolic feature-based sarcasm detection in Telugu conversation sentences," *Journal of Intelligent Systems,* vol. 30, pp. 73-89, 2020.
12. K. Jawad, M. Ahmad, M. Alvi, and M.-B. Alvi, "RUSAS: Roman Urdu Sentiment Analysis System," *Computers, Materials \& Continua,* vol. 79, pp. 1463--1480, 2024.
13. Z. Mahmood, I. Safder, R. M. A. Nawab, F. Bukhari, R. Nawaz, A. S. Alfakeeh, *et al.*, "Deep sentiments in roman urdu text using recurrent convolutional neural network model," *Information Processing & Management,* vol. 57, p. 102233, 2020.
14. R. Kreuz and G. Caucci, "Lexical influences on the perception of sarcasm," in *Proceedings of the Workshop on computational approaches to Figurative Language*, 2007, pp. 1-4.
15. P. Carvalho, L. Sarmento, M. J. Silva, and E. De Oliveira, "Clues for detecting irony in user-generated contents: oh...!! it's" so easy"," in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, 2009, pp. 53-56.
16. E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 704-714.
17. A.-C. Băroiu and Ș. Trăușan-Matu, "Automatic sarcasm detection: Systematic literature review," *Information,* vol. 13, p. 399, 2022.
18. R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 581-586.
19. A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 757-762.
20. A. Kumar and V. Anand, "Transformers on sarcasm detection with context," in *Proceedings of the second workshop on figurative language processing*, 2020, pp. 88-92.
21. B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, *et al.*, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1767-1777.

22. E. Filatova, "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing," in *Lrec*, 2012, pp. 392-398.
23. A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *CS224N project report, Stanford,* vol. 1, 2009.
24. S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," *arXiv preprint arXiv:1906.01815,* 2019.
25. D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "Cascade: Contextual sarcasm detection in online discussion forums," *arXiv preprint arXiv:1805.06413,* 2018.
26. A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network," *IEEE access,* vol. 7, pp. 23319-23328, 2019.
27. R. A. Potamias, G. Siolas, and A.-G. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," *Neural Computing and Applications,* vol. 32, pp. 17309-17320, 2020.
28. M. Hercog, P. Jarański, J. Kolanowski, P. Mieczyński, D. Wiśniewski, and J. Potoniec, "Sarcastic RoBERTa: A RoBERTa-based deep neural network detecting sarcasm on Twitter," in *International Conference on Big Data Analytics and Knowledge Discovery*, 2022, pp. 46-52.
29. M. Gole, W.-P. Nwadiugwu, and A. Miranskyy, "On sarcasm detection with openai gpt-based models," in *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, 2024, pp. 1-6.
30. M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu," *IEEE Access,* vol. 8, pp. 91213-91226, 2020.
31. A. Aggarwal, A. Wadhawan, A. Chaudhary, and K. Maurya, "" Did you really mean what you said?": Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings," *arXiv preprint arXiv:2010.00310,* 2020.
32. S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection," *Knowledge-Based Systems,* vol. 105, pp. 236-247, 2016.
33. D. Jain, A. Kumar, and G. Garg, "Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN," *Applied Soft Computing,* vol. 91, p. 106198, 2020.
34. D. Al-Ghadhban, E. Alnkhilan, L. Tatwany, and M. Alrazgan, "Arabic sarcasm detection in Twitter," in *2017 international conference on engineering & MIS (ICEMIS)*, 2017, pp. 1-7.
35. Z. B. Nezhad and M. A. Deihimi, "Sarcasm detection in Persian," *Journal of Information and Communication Technology,* vol. 20, pp. 1-20, 2021.
36. I. A. Farha, S. Wilson, S. V. Oprea, and W. Magdy, "Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection," in *The 2022 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5284-5295.
37. M. B. Alvi, N. A. Mahoto, M. A. Unar, and M. A. Shaikh, "Xtractor: A Two Step Tweet Extractor for Sentiment Analysis," presented at the 5th International Conference on Advancements on Computation Sciences (ICACS), Lahore, 2024.
38. S. U. Din, S. Khusro, F. A. Khan, M. Ahmad, O. Ali, and T. M. Ghazal, "An automatic approach for the identification of offensive language in Perso-Arabic Urdu Language: Dataset Creation and Evaluation," *IEEE Access,* 2025.
39. M. B. Alvi, N. A. Mahoto, M. A. Unar, and M. A. Shaikh, "An Effective Framework for Tweet Level Sentiment Classification using Recursive Text Pre-Processing Approach," *IJACSA,* vol. 10(6), 2019.
40. M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology,* vol. 63, pp. 163-173, 2012.