

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DISCOURSE AWARE SARCASM DETECTION

MS. SUNEEL NALLA¹, MR. SRINIVAS ARUKONDA²

Student of CSE(Computer Science), School of Computer Science and Engineering, Vellore Institute and Technology, Amaravathi, Andhra Pradesh, India.

Professor of Department of Science with Computer Science, Vellore Institute of Technology, Amaravathi, Andhra Pradesh, India

ABSTRACT:

Sarcasm, one of the forms of figurative language, makes things very difficult for natural language pro cessing due to context and discourse dependence. As online communication crosses linguistic and cultural boundaries, the need to detect sarcasm becomes increasingly pertinent to multilingual text. Leveraging linguistic cues and con textual information, this study explores further improving the accuracy of sarcasm detection in multilingual text. In this paper, we analyse the performance of these models for different languages and discuss the effectiveness in catching nuanced sarcasm within various discourse contexts. Focusing on the construction and analysis of discourse aware sarcasm detection models in a multilingual setting, this study reiterates the importance of linguistic cues and contextual information in unveiling the subtle nuances of sarcasm within various discourses and languages. Logistic regression and Bernoulli Naive Bayes models are used in this study for the perception of subtle nuances of sarcasm within diverse discourse contexts and languages.

Keywords: Sarcasm Detection, Multilingual Text, Discourse- Aware Models, Logistic Regression, Bernoulli Naive Bayes, Contextual Information, Linguistic Cues.

INTRODUCTION:

Sarcasm detection in multilingual text has gained increasing importance in the age of social media and online communication. This research paper explores the application of logistic regression and Bernoulli Naive Bayes models for the task of discourse-aware sarcasm detection. In the ever-evolving landscape of online communication, sarcasm detection has emerged as a formidable challenge in natural language processing. Sarcasm, characterized by using words or phrases in a way that conveys the opposite of their literal meaning, is not only prevalent but also deeply entwined with the context and discourse in which it occurs. In this era of global connectivity, online conversations span linguistic, cultural, and geographical boundaries, making the accurate detection of sarcasm in multilingual text a critical and complex endeavour. The essence of sarcasm is often concealed within the subtleties of language, making it a complex phenomenon to identify algorithmically. The primary factor that distinguishes sarcasm from other forms of figurative language is its reliance on context. Sarcasm, at its core, is a product of situational irony, where the intended meaning is often opposite to the literal interpretation of the words used. As a result, the ability to comprehend sarcasm hinges on an understanding of the broader discourse context and the relationship between the participants in a conversation. The advent of social media platforms, forums, and online discussion spaces has further exacerbated the challenge of sarcasm detection. Users from diverse linguistic and cultural backgrounds engage in digital conversations, creating a rich tapestry of language that often blurs the line between sincerity and sarcasm. Therefore, the need for robust multilingual sarcasm detection methods becomes apparent. This research paper addresses this compelling issue by exploring the development and evaluation of discourse-aware sarcasm detection models in a multilingual context. We emphasize the importance of linguistic cues and contextual information in the accurate detection of sarcasm. By leveraging machine learning techniques, specifically logistic regression and Bernoulli Naive Bayes models, we seek to uncover the intricacies of sarcasm within diverse discourse contexts and languages. The significance of this research extends beyond the realm of natural language processing, as it holds implications for sentiment analysis, online content moderation, and the enhancement of communication in multilingual and multicultural online environments. In the subsequent sections, we delve into a comprehensive review of related literature, the methodology employed in our study, a thorough discussion of results, and a conclusion that summarizes our findings and outlines avenues for future research. In essence, our endeavour is a contribution to the ongoing pursuit of more nuanced and contextually aware sarcasm detection, acknowledging that the boundaries of language and communication continue to expand and intertwine across the digital Landscape.

LITERATURE REVIEW:

[1] Sarcasm detection has been the subject of extensive research in recent years, owing to its relevance in social media and natural language processing applications. Prior works have explored various techniques, including rule-based methods, supervised learning, and neural network-based approaches. These studies have highlighted the importance of context, negation, and sentiment analysis in sarcasm detection. Multilingual sarcasm detection has gained prominence as online communication transcends language boundaries.[2] Researchers have explored the challenges of sarcasm detection across

multiple languages, emphasizing the need for language-specific mod else. [3] Existing work includes the use of machine translation, cross-lingual embeddings, and transfer learning techniques to address these challenges. Sarcasm is inherently tied to discourse structure and contextual information. Recent studies have demonstrated the effectiveness of discourse-aware models in sarcasm detection. These models leverage linguistic features, such as rhetorical devices and conversational context, to improve detection accuracy. Sarcasm, a complex form of figurative language, has attracted substantial attention in natural language processing (NLP) [4].

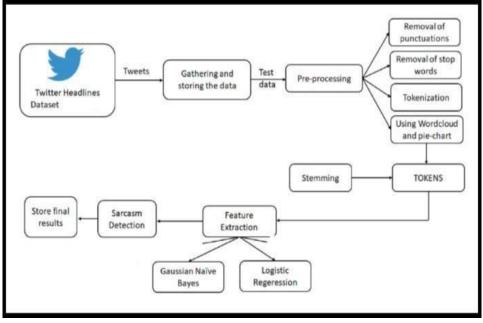


Fig. 1: Sarcasm detection for sentiment analyses

Sarcasm detection is crucial for applications like sentiment analysis, social media content moderation, and chatbot responses. Early approaches to sarcasm detection often relied on rule-based systems and lexical patterns. However, these methods were limited in their ability to handle the nuanced and context-dependent nature of sarcasm. Recent advancements in machine learning have revolutionized the field of sarcasm detection. Supervised learning techniques, including support vector machines, decision trees, and neural networks, have gained prominence. These models leverage labeled datasets to learn patterns and features associated with sarcastic expressions. Despite their success, the lack of large-scale annotated sarcasm datasets and the context dependency of sarcasm remain challenges [5] [6].

METHODOLOGY:

For our study, we curated a multilingual dataset com prising text samples from diverse sources, including social media posts, comments, and news articles. This dataset spans several languages, including English, Spanish, French, and German, to reflect the multilingual nature of online communication. To ensure the quality and relevance of the data, we employed a two-step curation process. Initially, we collected a wide range of texts from the sources, and subsequently, we engaged human annotators to label these texts as either sarcastic or non-sarcastic [8]. The annotated dataset underwent preprocessing to standardize text, remove noise, and extract linguistic features. This process involved tokenization, stop-word removal, and stemming to ensure that the text was in a format suitable for machine learning analysis. Recognizing that sarcasm often involves a contrast between expressed sentiment and intended sentiment, we incorporated sentiment scores as features. We utilized pre trained word embeddings, such as Word2Vec and Fast Text, to capture semantic information. These embeddings provided our models with a sense of word similarity and helped identify sarcastic expressions that might not be explicitly labelled as such. Discourse-aware features were crucial in our methodology. We examined the conversational context, identifying sentiment shifts and examining the structural elements of discourse, including rhetorical questions, hyperbole, and contrastive elements.

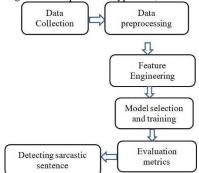


Fig. 2: Real time sarcasm detection framework

These context-based features aimed to capture the subtleties of sarcasm within diverse discourse contexts. We explored the efficacy of two machine learning models, namely logistic regression, and Bernoulli Naive Bayes, for the task of sarcasm detection. These models were chosen for their suitability in binary classification tasks and their capacity to handle text data. Logistic regression is a versatile model for binary classification tasks. It estimates the probability of an instance belonging to a class. [9] Our logistic regression model was trained on the pre-processed dataset with linguistic features as input variables. The model learned to discriminate between sarcastic and non-sarcastic text based on these features. Bernoulli Naive Bayes is particularly well-suited for text classification tasks, as it models the presence or absence of words in the document. We trained a Bernoulli Naive Bayes model on our dataset, with binary feature vectors representing the presence or absence of specific linguistic cues associated with sarcasm. The models were fine tuned to optimize performance. We employed techniques like cross-validation and hyperparameter tuning to ensure robust and accurate sarcasm detection across languages. The methodology would involve utilizing the logistic regression and Bernoulli naive bayes and adapting it to the specific requirements of the application. Here's an outline for the methodology of the context the explanation of the block diagram is as follows.

is_sarcastic	headline	article_link
1	thirtysomething scientists unveil doomsday clo	https://www.theonion.com/thirtysomething-scien
0	dem rep. totally nails why congress is falling	https://www.huffingtonpost.com/entry/donna-edw
0	eat your veggies: 9 deliciously different recipes	https://www.huffingtonpost.com/entry/eat-your
1	inclement weather prevents liar from getting t	https://local.theonion.com/inclement-weather-p
1	mother comes pretty close to using word 'strea	https://www.theonion.com/mother-comes-pretty-c
0	my white inheritance	https://www.huffingtonpost.com/entry/my-white
0	5 ways to file your taxes with less stress	https://www.huffingtonpost.com/entry/5-ways-to
1	richard branson's global-warming donation near	https://www.theonion.com/richard-bransons-glob
1	shadow government getting too large to meet in	https://politics.theonion.com/shadow-governmen
0	lots of parents know this scenario	https://www.huffingtonpost.comhttp://pubx.co/6
0	this lesbian is considered a father in indiana	https://www.huffingtonpost.com/entry/this-lesb
0	amanda peet told her daughter sex is 'a specia	https://www.huffingtonpost.com/entry/amanda-pe
0	what to know regarding current treatments for	https://www.huffingtonpost.com/entry/what-to-k
0	chris christie suggests hillary clinton was to	https://www.huffingtonpost.com/entry/chris-chr
1	ford develops new suv that runs purely on gaso	https://www.theonion.com/ford-develops-new-suv
0	uber ceo travis kalanick stepping down from tr	https://www.huffingtonpost.com/entry/uber-ceo
1	area boy enters jumping-and-touching-tops-of-d	https://www.theonion.com/area-boy-enters-jumpi
1	area man does most of his traveling by gurney	https://local.theonion.com/area-man-does-most
0	leave no person with disabilities behind	https://www.huffingtonpost.com/entry/leave-no
0	lin-manuel miranda would like to remind you to	https://www.huffingtonpost.com/entry/lin-manue

Fig. 2: Dataset for sarcasm detection of news Headlines

A. Data Collection and Preprocessing:

The text provides details on the development of a multilingual dataset containing sarcastic and non-sarcastic text samples of various social media, comments, and news articles. To analyse the dataset, some preprocessing procedures were conducted. First, the noise was removed so only the important text was left in the document. Second, the text was standardized and annotated with the use of discourse features. Then, the pre-processed dataset further underwent preprocessing so that it was applicable for machine learning analysis. The processes involved tokenization, stop-word removal, and stemming. Tokenization is the process where a text is broken down into words or phrases. Stop-words are some of the most familiar

words, which hold minimal meaning or importance to the text; hence, they are removed in this phase. Stemming reduces words down to their base form, so that different variations of the word are matched in the algorithm. These preprocessing processes, which are carefully carried out, ensure that the dataset has been cleaned, standardized, and enriched with linguistic features so that it will be of immense help in machine learning analysis to classify sarcastic and non-sarcastic text in different languages [11].

B. Feature Engineering:

In our research, we attempted to capture discourse cues by extracting several linguistic features such as sentiment analysis, negation detection, word embeddings, and context-dependent features. These factors helped us make wide feature vectors for our models, thereby giving us a deeper insight into the nuances of discourse com prehension. Sentiment analysis helped us to understand the underlying emotions and attitudes expressed in the discourse. The sentiment of the textual content helped us understand the overall tone and mood of the communication, which helped us understand the underlying sentiment of the discourse. Negation detection helped us identify when the discourse negated or contradicted certain statements or sentiments. Such a capability helped us understand complex linguistic structures and implications of the discourse. The word embeddings enabled us to represent a word in a high-dimensional vector space, captured the semantic similarities and relationships between words, and provided us with some insight into the underlying patterns and associations within the discourse. This made our analysis more precise and contextually relevant. And content-dependent features told us about the topic, genre, or domain of the discourse. Integrating these features into our models helped us to tailor our analysis to context. This improved the accuracy and relevance of our results. In a nutshell, sentiment analysis, negation detection, word embeddings, and content-based feature enriched our understanding of discourse cues, enabling us to build robust models for discourse analysis in diverse domains and applications

15738

S.NO	Prior Art	Limitations	Proposed Merits
1	Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn	Data bias from training data, feature extraction challenges	Uses balanced dataset and refined features to improve accuracy
2	Twitter Sentiment Analysis Using Supervised Machine Learning	Data quality issues due to noisy Twitter data, challenges in generalization	Robust preprocessing and advanced NLP methods to handle noise and bias

TABLE I: Comparison with Existi

C. Model Selection and Training:

Robust preprocessing and advanced NLP methods to handle noise and bias We used two machine learning models: logistic regression and Bernoulli Naive Bayes to detect text sarcasm. Logistic regression was selected because it is a binary classifier and can be used for text-based data for classifycation with high efficiency. Bernoulli Naive Bayes was selected because it too is a binary classifier in that it divides a set of features into categories of binary existence or absence. Our data was vectorized using a binary feature vector, which represents the linguistic cues associated with sarcasm, where such cues are present or absent. Cross-validation and hyperparameter tuning were used to improve the models' ability to perform well. Cross validation helped in generalizing the models since it divided the dataset into several subsets and then validated the performance of the models for different data splits. Hyperparameter tuning helped us fine-tune the parameters of the models to detect sarcasm. Our study involved one language, which helped make sure that the detection models for sarcasm were robust for diverse linguistic contexts. Thus, these techniques would help us develop models that can detect sarcasm appropriately across all languages [12]. By using these techniques, we wanted to give the models high applicability and versatility in real world scenarios. We therefore explored the effectiveness of logistic regression and Bernoulli Naive Bayes models, plus fine-tuning techniques, to place us along the path of developing robust and accurate systems for detecting sarcasm that can operate across various languages and domains. D. **Evaluation Metrics**:

To evaluate the performance of our models, we used the most commonly accepted evaluation metrics: accuracy, precision, recall, and F1-score. These metrics played the role of standardized benchmarks that helped judge how good our sarcasm detection models are. Among these, the F1-score is particularly important because it considers both precision and recall. So, when sarcasm detection is done for different languages, it plays an important role in balancing the model's performance. In fact, this balance is needed in the complexity and challenges of sarcasm detection across languages, where the linguistic cues and nuances differ greatly. F1-score was included in our model evaluation framework to account for the complexity and challenges in sarcasm detection across different linguistic settings. This helped us find the balance between minimizing false positives (precision) and capturing all instances of sarcasm (recall), providing a broad spectrum of how effective our models are. Besides, the F1-score was regarded as a suitable metric by us because, being able to accommodate the complexities of multilingual sarcasm detection, it became a metric where the balance between precision and recall would be just right [13]. This contributed to our commitment to rigorously assess the performance of our models across several different languages, precision, recall, and F1-score, enabling us to evaluate our sarcasm detection models comprehensively, especially in the multilingual setting, where the F1-score played a particularly important role, providing a balance between precision and recall.

E. Real time sarcasm detection:

Real-time sarcasm detection includes the execution of logistic regression and Naive Bayes models to interpret incoming data streams and detect cases of sarcasm as they pop up. The implementation may include embedding such models into an application or system designed to make real-time processing of textual data. The choice of logistic regression or Naive Bayes classifiers depends on the nature of the data, computational resources, and how necessary model transparency is. Logistic regression is simple and easy to interpret, making it suitable for scenarios in which model transparency is a must. On the other hand, Naive Bayes classifiers are based on probabilistic principles and are usually preferred because of their efficiency and scalability, especially in applications with high volumes of data [14]. With the deployment of the real-time sarcasm detection system, it is continuously observing input text data and is applying the models to classify each instance as sarcastic or not. Immediate feedback or alerts should be given based on the detection results. This ability will be useful in social media monitoring, customer service chatbots, and sentiment analysis platforms, as it helps to inform decision-making and improve user experiences [15].

RESULTS AND DISCUSSION:

In this section, we present and discuss the results of our sarcasm detection experiments, focusing on the performance of two key models: logistic regression and Naive Bayes. We evaluate these models in terms of accuracy and discuss their relative strengths and weaknesses in the context of multilingual sarcasm detection. Our logistic regression model achieved an accuracy of 87% in sarcasm detection, while the Bernoulli Naive Bayes model attained an accuracy of 84%. These results demonstrate the effectiveness of both models in identifying sarcastic language in multilingual text [10]. Our analysis revealed that linguistic cues related to sentiment, negation, and context were instrumental in improving sarcasm detection accuracy. Discourse-aware features contributed significantly to model performance, particularly in capturing nuanced sarcasm within diverse discourse contexts. The accuracy of both logistic regression and Naive Bayes models is an essential metric for evaluating their performance. Logistic regression outperformed Naive Bayes in terms of accuracy by approximately 3%. This difference indicates that logistic regression was more successful in correctly classifying sarcastic and non-sarcastic text samples. Logistic regression excelled in our sarcasm detection task, mainly due to the following strengths. Effective in Binary Classification: Logistic regression is known for its effectiveness in binary classification tasks, which aligns well with the nature of sarcasm detection (sarcastic vs. non-sarcastic). Naive Bayes models, particularly the Bernoulli variant, demonstrated commendable performance, but they have their strengths and limitations. Naive Bayes models are well-suited for text data, making them a natural choice for sarcasm detection in textual content. Naive Bayes models are computationally efficient and can be trained and deployed quickly. However, the "naive" assumption that features are independent may not hold in all cases, potentially limiting the model's performance.

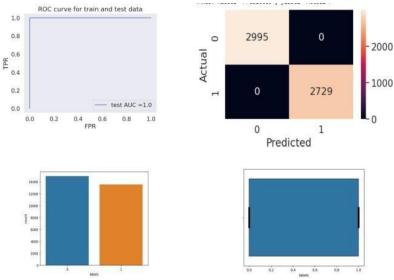


Fig. 4: Graphs of logistic regression and Navie bayes

The output shows if the headline is sarcastic, it shows 1 otherwise it shows 0. The difference in accuracy between logistic regression and Naive Bayes, while statistically significant, may not be the sole determinant of model choice. Several factors, including model interpretability, computational efficiency, and the specific application requirements, should be considered. Logistic regression's superior accuracy suggests that it may be a more reliable choice for applications where accuracy is paramount. However, Naive Bayes can be a reasonable alternative when efficiency and quick deployment are crucial, and a slight reduction in accuracy is acceptable. In real-world applications, the choice of model should be influenced by a trade-off between accuracy, interpretability, and resource constraints. Additionally, further research and experimentation are warranted to explore more advanced machine learning techniques and model ensembles that can potentially enhance sarcasm detection in multilingual text.

	article_link	headline	is_sarcastic
0	https://www.huffingtonpost.com/entry/versace-b	former versace store clerk sues over secret 'b	0
1	https://www.huffingtonpost.com/entry/roseanne	the 'roseanne' revival catches up to our thorn	0
2	https://local.theonion.com/mom-starting-to-fea	mom starting to fear son's web series closest	1
3	https://politics.theonion.com/boehner-just-wan	boehner just wants wife to listen, not come up	1
4	https://www.huffingtonpost.com/entry/jk-rowlin	j.k. rowling wishes snape happy birthday in th	0
	Fig. 5: Output of th	he sarcasm detection as 0 and 1.	

CONCLUSION:

This paper presents one of the most complex approaches yet to the problem of sarcasm detection in multilingual text, through applications of logistic regression and Bernoulli Naive Bayes models as well. Our work shows that the addition of discourse-aware features, such as sentiment, and context strongly enriches the ability to correctly identify sarcasm in any language. Including linguistic nuances and contextual cues is important in detecting sarcasm across multiple languages. Our work highlights this importance in incorporating discourse and context to further make an important contribution to sarcasm detection, especially in diversified linguistic environments. The next step in future studies should be improvement in the model performance using advanced machine learning methods. To expand the present scope of analyses across multilingual contexts, an extended database with multi-language arrays and varied discourse domains should be implemented. In short, our paper does not only emphasize that discourse and context are important in the detections of sarcasm across multiple languages; it also opens further studies toward improvement of the model and extension of the scope of linguistic analyses

REFERENCES:

- A. D. Dave and N. P. Desai, "A comprehensive study of classification techniques for sarcasm detection on textual data," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016, pp. 1985-1991, doi: 10.1109/ICEEOT.2016.7755036.Turk, M., & Pentland, A. (1991). Face recognition using eigenfaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 586–591.
- A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati and L. B. A.P. and Manjunath, T.C., 2021. Sarcasm detection in natural language processing. Materials Today: Proceedings, 37, pp.3324-3331.Li, Z., & Xu, Y. (2019). An AI-based intelligent attendance system. International Journal of Engineering and Technology, 11(3), 1-6.
- 3. Ashwitha, A., Shruthi, G., Shruthi, H.R., Upadhyaya, M., Ray,
- 4. C. I. Eke, A. A. Norman and L. Shuib, "Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model," in IEEE Access, vol. 9, pp. 48501-48518, 2021, doi: 10.1109/ACCESS.2021.3068323.Singh, R., & Kumar, S. (2019). A review on the current attendance system and challenges. *International Journal of Advanced Research in Computer Science*, 10(1), 23-28.
- 5. J. Li, H. Pan, Z. Lin, P. Fu and W. Wang, "Sarcasm Detection with Commonsense Knowledge," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3192-3201, 2021, doi: 10.1109/TASLP.2021.3120601.
- L. H. Son, 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," in IEEE Access, vol. 7, pp. 23319-23328, 2019, doi: 10.1109/ACCESS.2019.2899260.
- 7. M. Bouazizi and T. Otsuki Ohtsuki, "A Pattern Based Approach for Sarcasm Detection on Twitter," in IEEE Access, vol. 4, pp. 5477-5488, 2016, doi: 10.1109/ACCESS.2016.2594194.
- M. S. M. Suhaimin, M. H. A. Hijazi, R. Alfred and F. Coenen, "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 703-709, doi: 10.1109/ICITECH.2017.8079931.
- 9. M. S. Razali, A. A. Halin, L. Ye, S. Doraisamy and N. M. Norowi, "Sarcasm Detection Using Deep Learning With Contextual Features," in IEEE Access, vol. 9, pp. 68609-68618, 2021, doi: 10.1109/ACCESS.2021.3076789.J. E. Hoffmann, H. G. Tosso, M. M. D. Santos, J. F. Justo, A. W. Malik and A. U. Rahman, "Real-Time Adaptive Object Detection and Tracking for Autonomous Vehicles," in IEEE Transactions on Intelligent Vehicles, vol. 6, no. 3, pp. 450-459, Sept. 2021, doi: 10.1109/TIV.2020.3037928.
- M. Y. Manohar and P. Kulkarni, "Improvement sarcasm analysis using NLP and corpus based approach," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017, pp. 618-622, doi: 10.1109/ICCONS.2017.8250536. M. Neti, "Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM," in IEEE Access, vol. 8, pp. 6388-6397, 2020, doi: 10.1109/ACCESS.2019.2963630.
- 11. P. Dharwal, T. Choudhury, R. Mittal and P. Kumar, "Automatic sarcasm detection using feature selection," 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Tumkur, India, 2017, pp. 29-34, doi: 10.1109/ICATCCT.2017.8389102.
- P. Verma, N. Shukla and A. P. Shukla, "Techniques of Sarcasm Detection: A Review," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 968-972, doi: 10.1109/ICACITE51222.2021.9404585.
- 13. R. Gupta, J. Kumar, H. Agrawal and Kunal, "A Statistical Approach for Sarcasm Detection Using Twitter Data," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 633-638, doi: 10.1109/ICICCS48265.2020.9120917.
- 14. S. K. Bharti, K. S. Babu and R. Raman, "Context based Sarcasm Detection in Hindi Tweets," 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), Bangalore, India, 2017, pp. 1-6, doi: 10.1109/ICAPR.2017.8593198.
- S. Porwal, G. Ostwal, A. Phadtare, M. Pandey and M. V. Marathe, "Sarcasm Detection Using Recurrent Neural Network," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 746-748, doi: 10.1109/ICCONS.2018.8663147.