SARCASM DETECTION BY MACHINE LEARNING USING TWITTER DATA

(Literature Survey)

Athira K Menon,Neethu P,Dr.G.Kiruthiga

[1] Student, Department of Computer science, IES College of Engineering Thrissur, Kerala, India

[2] Assistant Professor, Department of Computer science, IES College of Engineering Thrissur, KeralaIndia

[3] Associate Professor, Department of Computer science, IES College of EngineeringThrissurKerala, India

ABSTRACT

Complex linguistic term sarcasm is frequently used in social media and e-commerce websites. In Natural Language Processing applications like sentiment analysis and opinion mining, failure to recognise sarcastic utterances will confound classification algorithms and produce erroneous results. Numerous studies on the detection of sarcasm have used various learning methods. The majority of these learning methods, though,have always placed their primary attention solely on the expressed ideas, disregarding the context. AsAs a result, they missed the sarcastic expression's semantics and background information. Secondly,A word embedding learning technique is widely used in NLP deep learning approaches as a typical method for convolutional feature vector representation, which does not take into account the polarity of the words emotional connotations.To solve the challenges noted above, this work suggests a context-based feature technique for sarcasm identification using the deep learning model, BERT model, and traditional machine learning. For the categorization, two Twitter and Internet Argument Corpus, version two (IAC-v2) benchmark datasets were used.the three learning models are employed. The initial model makes use of deep learning and embedding-based representation.recurrent neural network with a bidirectional long short term memory (Bi-LSTM) (RNN),the development of word embedding and context using Global Vector representation (GloVe)learning. The second model is constructed using a pre-trained Bidirectional Encoder representation and is based on Transformer.both Transformer (BERT). The third model, by contrast, is based on the feature fusion of the BERT feature.a feature with sentiment-related, syntactic, and GloVe embedding

Keyword: Natural language processing, Sarcasmidentification, Bi-LSTM, GloVe embedding, BERT

1.INTRODUCTION

Sarcasm is a difficult NLP topic because of its extremely figurative nature. Sarcasm is typically defined as An ironical taunt intended to show contempt. Major signs of sarcasm have been identified as interjections, punctuation, and sentimental shifts. When these lexical hints are available, sarcasm recognition is highly accurate. The use of context—which assumes common sense and prior knowledge of an event—is another element of sarcasm.

Understanding the context of earlier remarks as well as the essential background information regarding the topic of discussion may be important in order to recognise sarcasm in a discussion forum. The algorithm is not picking up on sarcastic tweets, perhaps because they are situation-specific.. Also the sarcastic tweets written in a very polite way are undetected. Sometimes people use politeness as a way of being sarcastic, highly formal words that don't match the casual conversations. Complementing someone in a very formal way is a common way of being sarcastic

1.1 PROBLEM STATEMENT

Proposing a hybrid approach of both content and context driven modelling in online social media discussions to detect sarcasm.

1.2 OBJECTIVE

To detect sarcasm by a content driven modelling in Twitter using machine learning algorithms. As we know that sarcasm detection is a very narrow research field in Natural Language Processing, a special case of sentimental analysis where instead of detecting a sentiment in the whole spectrum, the focus is on sarcasm.

1.3 OUTLINE

The project is divided mainly into four phases as data collection, data preprocessing, feature extraction and classification.The whole report starts with a small description about the theoretical background of the proposed approaches are included in Chapter 2. Chapter 3 contains the literature survey which helped to improve the systems performance. Chapter 4 describes the methodology of the project. The implementation details, hardware and software requirements are described in Chapter 5. Chapter 6 presents the results. Performance analysis is discussed in Chapter 7. Conclusion and future work are included .

2. LITERATURE REVIEW

This section of the paper discusses some of the studies on medical diagnosis utilizing machine learning and data mining approaches. Different techniques, including lexicon-based, traditional machine learning, deep learning, or even a hybrid approach, have been used to study the task of identifying sarcasm.In addition, numerous analyses of sarcasm detection conducted. Eke, et al. [18], for instance, performed SLR on recognizing sarcasm in textual data The investigation was done by taking "dataset collecting, preparation" into consideration.engineering techniques, including feature extraction,selection of features, display of features, and classification metrics of performance and algorithms. The research showed the most often used features are those that are content-based categorization of sarcasm. The research also showed that the common evaluation criteria like precision, recall, and accuracy,The most used metrics are f-measure and Area under the curve (AUC) parameters for measuring the effectiveness of classifiers.Sarcasm identification tasks have been researched by numerous academics. Two techniques—the "Incongruent words-only" method and the "all-words: method"—were described by Joshi et al. [19] Using sentence completion for "Expect the Unexpected:Research on "Sarcasm Detection" using "Sentence for caustic analysis, "complete" Considering appraisal,There were two sets of data used, including I Twitter data.2278 tweets ('506 total) were compiled by Riloff, et al. [14]1772 of them were not ironic (sarcastic). (ii) Data from discussion forums gathered by Walker, et al. [6] that include hand-labeled 752 sarcastic and 752 non-sarcastic tweets, for balance.WordNet Similarities and Word2Vec, however, were used to gauge how similarly performers perform. For evaluation, two-fold cross-validation was employed. Thus,the total accuracy of the predictions was 54%, as measured byusing Word2Vec

The word polarity disambiguation strategy has also garnered respect by many in addition to the handmade characteristic suggested by several studies for sarcasm categorization.in recent years, scholars. Wu and Wen, for instance [20]investigated an unsupervised, knowledge-based strategy for automatic ambiguity disambiguation for dynamic sentiment using a search engine to find adjectives. Surprisingly, the author both pattern-based and character-based methods to Extrapolate the expected sentiment of nouns and locate adjectives.polarity. In a different investigation, Xia, et al.a method that resolves words using opinion-level features.ambiguity in polarity Using this strategy, the author looked at the(For example, discourse, corresponding words in both inter-opinion (e.g., indicator words, opinion) and intra-opinion target) attribute. The probability technique was used by them tothe word polarity disambiguation is resolved by using theBayesian approach.

However, they conducted research on the pinion corpus, and the findings indicated a significant impact in the disambiguation of word polarity utilising the level of opinion feature. In a current investigation, Wang, et al.proposed deep learning word sense disambiguation method. In the suggested method, a target's sense path Utilizing the domain-specific knowledge, context background information from WordNet using the Word embedding functionality taken from an outside corpus.The technique exposed the internal concealed semantic link.word sense through taking advantage of the "PageRank algorithm"while representing the text, use the WordNet structure as a sense path.targeting context through latent semantic analysis Abdalgader and Al Shibli [23] provided an alternative to the "graph-based word sense disambiguation approach" in a study that was similar to this one.utilising every instance of collected semantic information connecting graph semantics for leveraging the WordNet to facilitate determining the expected meanings of words in a given context. This suggested strategy's resemblance between the network nodes containing all associated word semantic Measured information is used for sentence-level disambiguation.The genuine meanings are then simultaneously assigned to each application of a graph centrality measure that demonstrates the critical degree between graph nodes.However, the outcomes of the experimental examination and comparison of the approach with the benchmark dataset outperformed the state-of-the-arts WSD approaches.

In their research on "Detecting Irony and Sarcasm in Microblog: the role of Expressive Grammar," Fersini, et al. [24] first presented the concept of the ensemble learning approach.Ensemble Classifiers and Signals The author thinks about the Various classifications and "Bayesian Model averaging"algorithms based on their minimal reliabilities probabilistic forecasts Nevertheless, they thought Bayesian Averaging the model and majority Voting serves as the categorization phase's primary ensemble strategy. for assessment In order to achieve his goals, the author chose a baseline model that the four configurations with the best predicted performance BoW, PoS, PP, PP, & PoS are included. Nevertheless, the prognostic Analysis shows that the majority voting proposal The ensemble model outperformed the single classifier in performance The author also demonstrated how sarcasm can be enhanced by positives and how pragmatic features can help recognise ironic speech. Another investigation by ONAN et al. [25] suggested identifying the parody in a Turkish news item. method. The authors used linguistic research and word analysis.software for feature extraction by taking into account the feature sets in psychology and linguistics. In this research,word cloud, five deep learning architectures, and group learning The embeddings scheme was taken into account. In the experimental Analysis of the suggested strategy revealed that the deep learning strategy fared better than other strategies, which demonstrated the importance of the ideas put forth.

The first attempt to investigate a multilingual strategy for sarcasm identification was conducted by Ptáek, et al. [26]. In their research, two languages were taken into consideration (English and Czech).The authors used datasets in both English and Czech to Compare how often sarcasm appears in the two languages. The dataset includes 140 thousand tweets written in Czech and 780,000 English-language tweets. the Twitter API used to broadcast the tweets. As part of the classifying process Support vector machine and maximum likelihood The models' performance was assessed using an entropy classifier. predicting capability. Each classifier underwent a 5-fold crossvalidation process throughout the testing phase. Thus, 0.947 a balanced and earned a 0.924 F-measureEnglish dataset with the RF classifier is unbalanced. However,SVM generated a superior outcome on Esuli, et al. [27] suggested a cross-lingual sentiment qualifying approach for emotional computing and sentiment classification in which the training data are missing from the target language but present in the source language necessary to carry out sentiment qualification. hence, this technique focuses on the use situations where there is a training paper for a different source is present.languages, the lack of a training manual, and the targeted language with interest The author made use of structural and distributional correspondence indexing (DCI) cross-lingual text using the systematic correspondence learning (SCL) method classification. The experimental analysis, however, employing the benchmark datasets for sentiment classification across languages provided encouraging cross-lingual prediction outcomes qualification of feelings

Yang, et al. [28] suggested a novel technique for sentiment categorization in a different study that they called "Segment-level joint topic sentiment model (STSM)". The plan's goal was to identify the paper.capturing the correlation of the topic and measuring sentiment polarity sentiment. The author used a hybrid topic-sentiment model to create using the sentiment layer to add a correlation layer between the layers for segments and topics. Nevertheless, the feeling The predictive capability of classification demonstrates that the proposed strategy can improve compound and complex performance of sentences. Additionally, the sentimental compatibility and themes also conveys the importance of the suggested method.Agrawal, et al. [29] explored emotion categories features like sadness, happiness, surprise, etc. in their study on sarcasm identification; the authors went further by taking into account the sequential encoding of information among the effective attributes say. Comparative evaluation of the proposed strategy displays the potency of method. Onan, et al. [30] recently investigated the performance of traditional machine learning for classification as well as deep learning models for product sentiment analysis review, and the predicted effectiveness shows that effectiveness of the suggested approach on feelings classification.The use of multiple task learning has recently acquired popularity and has been shown in numerous NLP Challenges like implicit discourse link recognition [32] and key phrase border detection [31] are among them.detection of implicit discourse relationships [32].

Majumder, et al. [33] suggested a "multitask learning framework employing DNN" for sentiment and sarcasm in a related study of identification. They showed in their investigation that the two activities are connected, thus we modelled them together.employing a single neural network for tasks. Though, the experimental outcomes somewhat outperformed the current technique, showing that the multi-task network gets better change in the polarity of sarcasm and its classification. Mishra, et al. [34] suggested a solution in a similar investigation.method for automatic cognitive feature extraction using a CNN version for tasks involving sentiment and sarcasm detection.The dataset's gaze data were used by the author. The author individually modelled the two tasks during the modelling phase.

Riloff, et al. [14] provided a method for identifying a particular type of sarcasm in which there is a discrepancy between pleasant feelings and unfavorable circumstances. They suggested a "algorithm bootstrapping" that uses a single seed word that recognizes and learns a phrase that indicates something automatically The sarcastic's perspective on both positive and detrimental circumstances tweets. The authors developed two benchmark methods and modelled SVM classifiers using the LIBSVM library,use 10-fold cross-validation to assess the model. However, a recall of 39% and precision of 64% were attained by using SVM to analyse bigram and unigram features. hence, this approach worked as intended, however there were many of sarcastic tweets were missed in the sarcasm classes described above.

Additionally, the strategy is ineffective when applied to newly collected twitter data because it depends on the existence of all conceivable "Negative situations" in the training data. Bala and Mukherjee [11] presented away for supplying information to a system that understands linguistic tone of the author by taking various sets of characteristics for detecting sarcasm in a microblog. When using aspects of authorial style for their research and in the classroom techniques for the classification phase, Naive Bayes, and fuzzy clustering they utilised. The results of the experimental investigation show that the use accuracy was improved by combining supervised and unsupervised learning with features that are not dependent on text. in recognising irony. However, the strategy is relatively modest. authorial style-based elements and could not be compatible with different feature sets Word-level approaches need to be trained multiple times for huge social data analysis using the current methods In order to get around the restriction, Hussain and Cambria [35] presented a new "Semi-supervised learning model" by com-Binding the vector's "random projection scaling" component Using an SVM and a space model (VSM), cognitive on an emotive common sense that is knowledge-based. How-however, the outcomes of the experimental analysis showed an important improved polarity identification and emotion recognition because both are labelled, detection prevails over the classification rule. data are used for categorization learning, including labelled and unlabelled data contrasted with the formal approach, which uses solely the data with labels. Consequently, it provided a platform for additional research on huge social networks using semi-supervised learning data evaluation.

A new semi-supervised learning technique that takes into account both training and testing sets for sentiment categorization of stock was proposed by Duan, et al. in a related study [36].sending texts. The approach was put forth to address the problem.data sparsity, which is typical in short message modelling in representation using mathematics. Additionally, the author con-built a model of generative emotions using words with categories (GEM-CW) to glean sentiment characteristics from both training as well as test sets. The extracted traits, however, were more discriminating than those produced for sentiment categorization employing a traditional strategy that just takes into account training sets. The analysis' findings show that the suggested learning The strategy and the model are crucial for sentiment classification. in brief language and can produce superior outcomes to the conventional techniques.

Rajadesingan, et al. [12] probed the psychology behind the sarcastic expression in greater detail. They offered behavioural modelling for sarcasm recognition in their study. defining the different types of sarcasm and confirming their existence on Twitter. The study highlights the value of historical information obtained in identifying from a previous tweet sarcasm. Although the strategy appears to be quite successful in such In the lack of prior experience, for example, it cannot perform well. information regarding the user. Because the majority of the features used in the study were taken from the collected data.from a previous tweet to decide Researchers have recently been drawn to the concept of the deep learning approach to integrate it with the traditional machine learning approach for sarcasm identification.. Mehndiratta, et al. [37] gave, as an example,Automatic sarcasm detection in text data method utilising a DCNN Sentiment polarity was used in their study as a skip-gram to obtain a feature vector from a feature set word2vec model approach. The authors continued to feed the convolutional neural network with a feature. Their research functioned admirably, but had a word sense limitation not being recorded individually.Recent research [40] used textual, voice, and video data known as multimodal features to identify sarcasm. The data sets' textual features were representative of BERT (Bidirectional Encoder Representation) sent a specification for sentence representation is found in Transformer) [41]. The opposite is true for voice feature extraction.was retrieved utilising Libnsa, a renowned speech library

by merely taking into account the low-level characteristic in extraction [42] to take advantage of the audio modality information for audio data.Additionally, a pool of five ImageNet [43] layers was used.a frame for the video's visual feature extraction. Nevertheless, the experimental research showed that multiple Timodal properties led to improved prediction accuracy.compared to the unimodal characteristics, with a roughly 12.9% reduction in rate of error Recently, [44] provided a framework for identifying sarcasm in social media posts by utilising deep learning and neural language models, such as such as Word2Vec, GloVe, and FastText. The authors began.on the basis of weighted word embedding, an inverse gravity moment by trigram. The empirical examination of the suggested framework-correctness of the work was 95.3 percent, indicating that it the efficiency of the suggested structure.

A unique context-based features technique is put forth in this paper to identify sarcasm in three benchmark datasets. There were two learning models created for the pro-posed approach The initial model relies on semantic features that are utilising a global vector for word embedding (GloVe). Word Learning the representations and relationships is aided by embedding amid words, ships A count-based model called the Glo demonstrates how a sentence's words are related to one another (relatedness) and creates the learned model of a real word value vector Each word in each tweet is connected to its matching vector in a word table. In order to train and evaluate the model for sarcastic and non-sarcastic categories, the generated features are used. The following model is built using transformer learning (BERT)as well as the feature fusion included in the BERT feature, hashtag feature, sentimental component, syntactic component, and Glove embedding capability. The most advanced model is BERT.expresses the circumstances in a sardonic manner.

3.CONCLUSION AND FUTURE WORK

The sarcasm detection system uses content-driven modelling and machine learning techniques to identify sarcasm in tweets. As is well known, sarcasm detection in natural language processing is a fairly specialised area of sentimental analysis where the emphasis is on sarcasm rather than detecting a sentiment over the entire spectrum. The study provided illustrations that included the techniques, datasets, and performance metrics. This demonstrates unequivocally that including variables that characterise the psychological and behavioural aspects of the user helps the process of automatically identifying sarcasm. The classifier is performing as expected and has successfully used the set of features to analyse both sarcastic and non-sarcastic tweets. The classifications using Random Forest produced the best results, as we saw.The classifications using Random Forest produced the best results, as we saw. Up to this point, several of the top classification techniques for sarcasm detection have been examined and contrasted. It is necessary to observe a better classification technique for sarcasm detection in visuals and audio.

4.REFERENCES

[1]. Reference 1 Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann and Rada Mihalcea *"CASCADE: Contextual Sarcasm Detection in Online Discussion Forums"*,2018

[**2**] Reference 2 Rajeswari K and ShanthiBala P, *"SARCASM DETECTION USING MACHINE LEARNING TECHNIQUES"*,2018

[**3**] Reference 3 Shubhodip Saha, Jainath Yadav and Prabhat Ranjan. *Proposed Approach for Sarcasm Detection in Twitter*,2017

[**4**] Reference 4 D.V.Nagarjana Devi, Dr.T.V.Rajanikanth, Dr.V.V.S.S.S. Balaram, *"Sarcasm Detection in Plain Text Using Machine Learning"*,2018

[**5**] Reference 5 Aditys Joshi,Pushpak Bhattacharya and Mark J. Carman, *"Automatic Sarcasm Detection: A Survey"*,2017

[**6**] Reference 6 Karthik Sundararajan and Anandhakumar Palanisamy, *"Multi-Rule Based En semble Feature Selection Model for Sarcasm Type Detection in Twitter"*,2019

[7] Reference 7 Mukherjee and P. K. Bala, "Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering," Technology in Society, vol. 48, pp. 19-27, 2017.

[8] Reference 8 M. Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter," IEEE Access, vol. 4, pp. 5477- 5488, 2016.

[9] Reference 9 Ravinder Ahujja, Shantanu Bansala, Shuvam Prakasha, Karthik Venkataramana and Alisha Bangab - "Comparative Study Of Different Sarcasm Detection Algorithm Based On Behavioural Approach"

.[10] Reference 10 Al Amrani, Y. Lazaar, K.E Khadiri , "Random Forest And Support Vector Machine Based Hybrid Approach To Sentiment Analysis".

[11] Reference 11 H M Keerthi Kumar , B S Harish - "Sarcasm Classification Novel Approach By Using Content Based Feature Selection Method"

[12] Reference 12 Shubhodip Saha, Jainath Yadav and Prabhat Ranjan -"Proposed Approach Detection in Twitter"[2017]

[13] Reference 13 Pooja Deshmukh and Sarika Solanke-"Review Paper:Sarcasm Detection andObserving User Behavioural'[2017].

[14] Reference 14 Dmitry Davidov ,Oren Tsur and Ari Rappoport , "Semi Supervised Recognition of Sarcastic Sentences in Twitter and Amazon"[2010]

[15] Reference 15 Mondher Bouazizi, Tomoakki Ohtsuki-"Opinion Mining in Twitter How tomake Use of Sarcasm to Enhance Sentiment Analysis"[2015].