SENTIMENT ANALYSIS – SARCASM DETECTION USING MACHINE LEARNING

Kavitha N[1], Dr. MN Nachappa[2]

[1]PG student, Department of Computer Application, JAIN University, Karnataka, India
[2]Head of Department, School of CS & IT, Department of Computer Application, JAIN University, Karnataka, India

---

ABSTRACT: *This survey paper reviews all the prime works on sarcasm detection. The recognition of social media such as Facebook, Twitter, Instagram etc, has resulted sarcasm analysis gaining its momentum. Sentiment Analysis and Opinion Mining are of nice interest to the researchers. It is the method of classifying the opinions or sentiments according to the polarity of the text into positive, neutral and negative. Most of the organizations and industries extremely rely upon information analytics for their planning and decision-making process.*

Having information on the sentiment about a product or a service offered by an organization is only too valuable during this era than ever before. Knowing whether or not their established client base is showing sentiment toward their product or service are often game-changing for corporations.

The goal of this analysis is to use and assess procedure intelligence techniques (such as data processing and machine learning) to find sarcasm in text. It additionally evidences the usage of various techniques utilized by different authors.

KEYWORDS: Hybrid Methodologies, Machine-learning, Natural Language Processing, Python, Sarcasm detection, Social media, Stop words, SVM, Tokenization, Translation.

BACKGROUND:

What is sarcasm?

The Oxford lexicon provides the subsequent definition: Sarcasm is that the use of language that commonly signifies the alternative so as to mock or convey contempt.

In 2013 once Justine Sacco was on the brink of board a 12-hour flight from London to Cape Town -South Africa and Tweeted: 'Going to Africa. Hope I do not get AIDS. Just Kidding. I am White!' The post had been retweeted over three thousand times and was picked up by media platforms around the world. This racist and insensitive tweet went viral. When finally let her speak, she aforementioned she was being sarcastic and joking and she didn't mean to offend anyone. However, at that time, her intention was digressive. This illustration shows how human languages might be ambiguous, and black at times.

Giving another example- "Watching the TV? Again?!" In the above statement the speaker is attempting to mock another person by asking if she is watching Tv and that too again!

The primary goal is to predict sarcasm in every auditory communication of the victimization nature of a scene. Here, rather than detection of a positive or negative sentiment, the main focus is to catch the sarcasm which implies to know whether or not a bit of text sort of a social media post is sarcastic or not (sarcastic sentiment analysis). This field has been formed with challenges, particularly since sarcasm, not like alternative human emotions, is fairly onerous to find by a computer algorithmic program.

INTRODUCTION:

Sarcasm detection in dialogues has been gaining demand since then among natural language processing (NLP) researchers with the augmented use of informal threads on social media.

The importance of sarcasm analysis became the main focus of attention because it was necessary for the social platforms to analyse sarcasm in their posts and tweets. Sarcasm detection represents a targeted analysis field in NLP. Sarcasm is an associated extensively studied linguistic development by linguistic scholars. It became an admired research field for recent durations. Automatic sarcasm detection has attracted the information processing researchers' interest as the rising social media and sentiment analysis.

Sentimental analysis is a type of text mining that uses context to extract and detect subjective data from a text or sentence. The fundamental idea here is to use machine learning techniques to extract the sentiment of the text. Sentiment analysis is commonly employed in NLP to know people's subjective opinions. However, the analysis results are also biased if individuals use sarcasm in their statements. So as to properly perceive people's true intention, having the ability to find sarcasm is essential.

Sarcasm detection may be a troublesome task, for the most part smitten by context, previous information and also the tone in during which the sentence was spoken or written. So as to properly perceive people's true intention, having the ability to find sarcasm is important. Application of sarcasm detection can profit several areas

of interest of NLP applications, together with market research, opinion mining and data categorization. However, sarcasm detection is additionally a really troublesome task, as it's for the most part about context, prior knowledge and also the tone within which the sentence was spoken or written. Sarcasm detection may be a slender analysis field in NLP, a particular case of sentiment analysis wherever rather than detection a sentiment within the whole spectrum, the main focus is on sarcasm. Therefore, the task of this field is to find if a given text is sarcastic or not. The key objective of this paper and its sarcasm detection system is to distinguishing people's opinions (sentiments) concerning product, politics, services, or peoples brings tons of advantages to the organizations. The chance of distinguishing subjective information is important. It helps generate structured knowledge that is a chunk of necessary knowledge for decision support systems and individual decision-making.

With the assistance of Machine Learning, this project aims to analyse the sentiment of the text or comments or reviews that helps in resolution the obscurity of that means and improves the overall sentiment classification of an enormous quantity of user's textual data gained from social media.

Higher accuracy of sarcasm detection. Compared to rule-based approach or Lexicon-based approach, machine learning based approach have higher precision and return mostly have relevant results as they contemplate multiple extra factors. This can be as a result of ML technologies can consider many more data points, together with the tiniest details of behavior patterns related to a specific account.

Less manual work required for extra verification. ML-driven systems separate out, roughly speaking, 99.9 percent of traditional patterns leaving only 0.1 percent of events to be verified by specialists.

Lesser false polarity: Polarity for a component defines the orientation of the expressed sentiment, i.e whether or not it's positive, negative or neutral sentiment of the user concerning the entity in thought.

Ability to spot new patterns and adapt to changes. Not like rule-based systems, ML algorithms are aligned with a perpetually ever-changing atmosphere and monetary conditions. They allow analysts to spot new suspicious patterns and build new rules to stop new forms of sarcasms.

Sarcasm detection may be a troublesome task, for the most part smitten by context, previous information and also the tone in during which the sentence was spoken or written. So as to properly perceive people's true intention, having the ability to find sarcasm is important.

Application of sarcasm detection can profit several areas of interest of NLP applications, together with marketing research, opinion mining and information categorization. This project will employ Hybrid Methodology of Classifiers to extract sarcasm using Natural Language Processing and the best plan of action. The goal of this project is to improve sentiment analysis accuracy and construct a predictive model that ensures the correctness of the overall prediction. This project uses Machine Learning to analyse the sentiment of text, comments, and reviews, which aids in the resolution of ambiguity in meaning and enhances overall sentiment categorization of a large amount of user textual data acquired from social media.

Our research aims to benchmark multiple machine learning strategies like k-nearest neighbor (KNN), random forest and support vector machines (SVM) while the lexical-based approach such as vader sentiment analysis. Accuracy score, preciseness and Recall are the three-evaluation metrics that would be used.

LITERATURE REVIEW:

[1] A study titled "Using a Hybrid-Classification Method to Analyze Twitter Data During Critical Events" Saadat M. Alhashmi, Ahmed M. Khedr, Ifra Arif and Magdi El Bannany proposed in 2021. They mentioned on using Hybrid-Classification Methodology to investigate information. The authors analysed the performance of hybrid classifier using two Twitter datasets: COVID-19 dataset and the Expo2020 dataset. The programme divided the input tweets into four stages: (i) data collection, (ii) tweet pre-processing, (iii) feature extraction, and (iv)the proposed hybrid classification strategy. This hybrid-based strategy is proposed to handle the following issues: enhancing accuracy, detecting the polarity of comparative sentences, distinguishing the intensity of opinion words, taking into account negative comments, and dealing with sarcasm.

The work presented four main contributions: a) hybrid classification techniques are well explored for sentiment analysis, b) a unique hybrid classification approach supported on BFTAN is projected for sentiment analysis, c) a brand-new Twitter dataset associated with the recent event (COVID-19) which will be used further in future research, d) it's through empirically shown that the hybrid-classification approach can do comparable performance in up. The projected approach relies on hybrid-based approach of SVM and Bayes Factor Tree Augmented Naive Bayes (SVM & BFTAN) which resulted in Accuracy of 90.84%, precision of 91.22% and Recall of 90.08%

[2] B. Venkatesh and H.N. Vishwas proposed " Real Time Sarcasm Detection on Twitter using Ensemble Methods " in a paper published in 2021 in relation to my research. This work uses 2 hybrid machine learning approaches, particularly Stacked Generalization and Boosting

ensemble ways with Support Vector Machine (SVM), Random Forest (RF) and KNN as base classifiers and Logistic Regression (LR) as Meta classifiers to find real-time sarcasm on Twitter. Boosting and stacking ensemble approaches are utilised in the proposed system, and they are compared to multi-layer perceptron. On the basis of factors such as accuracy, detection rate, and precision, an analysis of both ensemble approaches is conducted in order to determine the optimal way. In order to develop an efficient model, you must choose an appropriate combination of the base classifier and the Meta classifier. The overall design of the proposed system is split into five different elements. Data assortment, Information pre-processing (which involves Tokenization, stemming, Removal of stop words, Data Distribution), Training, Evaluation and Sarcasm Detection. We may infer that Stacked Generalization ensemble approaches outperform Boosting ensemble methods based on the results achieved, dataset used, attributes picked, and the mix of classifiers used. The model enforced using the boosting ensemble method gave an accuracy of 73%, with a detection rate of 71%. The Stacked Generalization ensemble model outperformed the Boosting ensemble method with a 97% accuracy and detection rate.

[3] Raju Kumar and Aruna Bhat proposed “An Analysis On Sarcasm Detection Over Twitter During COVID-19” in a paper published in 2021. They mentioned framework specializing on commonplace classifiers generally employed in text classification like Linear support vector classifier (libSVM), Naïve Bayes (NB), and Decision Tree. Scikit-learn python library is employed for support vector classifier. libSVM with a linear kernel of SVM is employed whereever Lagrange multipliers are used to find the optimum solution. Naive Bayes classifier is enforced with Natural Language Toolkit (NLTK) library of python. Decision Tree classifier was additionally used from “scikit-learn” library of python. The accuracy, precision, and recall of each classifier are determined using the confusion Matrix for classifier efficiency measurement.

The Decision Tree achieved up to 90% accuracy. The Naïve Bayes detected the sarcasm with high preciseness as compared to a different classifier. The Decision Tree classifier additionally achieved the best recall; So, the decision tree’s overall performance was best compared to SVM and Naïve Bayes.

[4] A paper titled "Support Vector Machine Classifier with Principal Component Analysis and K Mean for Sarcasm Detection" Jyothi Godara and Rajni Aron proposed in 2021. During this analysis, different classification strategies like Decision tree, SVM, KNN and Naïve Bayes are applied for the sarcasm detection. The performance of varied classifiers - Decision Tree, Naïve Bayes, K-nearest neighbor, Support Vector Machine, SVM with PCA and support vector machine with PCA and K-means clustering algorithms were compared and results were presented. For performance analysis, the performance of different classifiers such as SVM, KNN, PCA+SVM and PCA+K- mean+SVM was compared. The combination of PCA, K-mean, and SVM provided accuracy of 93.49 percent, precision of 61.00 percent, and recall of 93.00 percent among all the classifiers. When compared to other classifiers, the combination of PCA, K-mean, and SVM provides the best results. [5] In a very study titled " Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analysis”, Mandala Vishal Rao and Sindhu C proposed in 2021. The paper discusses the usage of classification algorithms such as Support Vector Machine (SVM), K Nearest Neighbors and Random Forest. Amazon is one among biggest websites and presently has more than 310 million active users. The reviews and comments on this website were analysed for sarcasm. The SVM Classifiers gave an accuracy rate of 67.85%, Random Forest 62.34% and K Nearest Neighbors came out with Accuracy rate 61.08%. Other methods with a higher accuracy rate performed better than the Support Vector Machine algorithm.

[6] P. Verma, N. Shukla, and A. P. Shukla explored several approaches of spotting sarcasm, including Machine Learning, Contextual, and Deep Learning Approaches, in their work "Techniques of Sarcasm Detection: A Review." Precision was 91.1 percent and accuracy was 83.1 percent using a pattern-based method. SVM had a precision rate of 74.59 percent and Voting Classifier had an accuracy rating of 83.3 percent. The ensemble-based learning algorithm attained a 95 percent accuracy rate. LSTM-CNN obtained 93% and 95% of the time, respectively. Contextual Network Approach received an F1 score of 75%. The accuracy of genetic optimization on LSTM with CNN was 93 percent to 95 percent. Balanced F-Score for Bi-directional Long Short-Term Multi-Head Attention is 77.48 percent, whereas Imbalanced F-Score is 56.79 percent. SemEval Dataset Accuracy – 97.87 using Bi-directional Long Short-Term Memory and the Soft Attention function, followed by the Convolution network model and The accuracy of the Random Forests Tweets Dataset is 93.71 percent. Among all the cutting-edge models, the Soft-Attention function based Bi-directional Long Short-Term Memory via convolution networks (sAtt-BiLSTM-ConvNet) displays excellent results.

[7] H. Nguyen, J. Moon, N. Paul, and S. S. Gokhale's research paper, "Sarcasm Detection in Politically Motivated Social Media Content," focused on extracting sarcasm in social media content by collecting positive examples of sarcasm using hashtags that explicitly convey sarcasm, such as #sarcasm, #sarcastic, #irony, and #cynicism. In three ways, the research advanced the state-of-the-art in sarcasm detection. It didn't use hashtags to collect examples of sarcastic tweets or to aid sarcastic content recognition. They devised a complete

labelling system that incorporates sarcasm terminology from four widely used English dictionaries. They measured the contribution of each group of traits rather than just detecting sarcasm.

This study discovered that non-contextual factors play a larger part in the expression of sarcasm, which can be used to create portable classifiers that can recognise sarcasm in a variety of circumstances. With an F1-score of 0.83, the CNN classifier distinguished between sarcastic and non-sarcastic tweets. When compared to the accuracy reported by modern sarcasm detection techniques, which range from roughly 60% to 85%, the performance is competitive at the higher end.

The study found that portable classifiers trained on benchmark data may be used to recognise sarcastic material in completely unrelated contexts.

[8] P. H. Lai, J. Y. Chan, and K. O. Chin published a paper titled "Ensembles for Text-Based Sarcasm Detection." In order to detect sarcasm in tweets, four different classifiers were used. Support Vector Machine, Neural Network, Nave Bayes Classifier, and k-Nearest Neighbor are all examples of machine learning algorithms. Individual algorithm classification performance was compared to an ensemble learning algorithm based on stacking in this study. Individual algorithms were trained to detect sarcasm using the ensemble idea. The accuracy and precision of individual classifications were determined. The accuracy and precision of the ensemble classifiers Nave Bayes, Neural Network, and SVM were 98.94 percent and 99.60 percent, respectively. The accuracy and precision of Nave Bayes, Neural Networks, and KNN were 98.32 percent and 98.34 percent, respectively. The accuracy and precision of neural networks, KNNs, and SVMs were 98.92 percent and 99.58 percent, respectively. The accuracy and precision of Nave Bayes, KNN, and SVM were 99.70 percent and 99.60 percent, respectively. Overall, 98.97 percent accuracy and 99.28 percent precision were achieved.

[9] The scientific work of the authors M. Zanchak, V. Vysotska, and S. Albota on "The Sarcasm Detection in News Headlines Using Machine Learning Technology." The data came from the Kaggle platform, which gathered news headlines from two American websites. Each data category was divided into two groups: sarcastic and non-sarcastic. To apply the models, the data was tokenized. The processed data was subjected to logistic regression and multinomial Bayesian classifiers. By increasing the smoothing value, the Bayesian classifier's accuracy was improved by a few hundredths. When the alpha was 0.4, the accuracy was the best.

[10] "LSTM Based Sentiment Analysis," a research paper by Dirash AR and Manju bargavi S.K. The dataset was used to train the model using the Long-Short Term Memory – LSTM technique. A large dataset was employed to avoid underfitting. Amazon owns the "Internet Movie Database" dataset. The LSTM algorithm was utilised to create a sentiment analysis classification model. The RNN (Recurrent neural network) was utilised to solve the sequence prediction problem.

The model uses Long-Short Term Memory (LSTM) to classify the sequential textual data. The strength of a negative or positive remark was depicted using probability criteria; if the probability is close to zero, the sentiment is highly negative, and if the probability is close to one, the sentiment is extremely positive. The algorithm gave an accuracy score of 86.68%

CONCLUSIONS:

Sarcasm detection is one of the interesting topics in sentiment analysis. Due to its ambiguity and complexity in nature. It is a tedious task to detect sarcasm more accurately from the dataset. This challenge is one amongst the reason that scholars are attracted to sarcasm detection study.

In our research approach, the data is collected and a deep illustration of Learning Model using various machine learning techniques is conferred for determining on-line sarcasm detection, with the benefit of attaining robust and reliable results.

Sentiment analysis or opinion mining may be a field of study that analyses people's sentiments, attitudes, or emotions towards sure entities. This paper tackles an elementary downside of sentiment analysis, sentiment polarity categorization.

RERERENCES:

[1] S. M. Alhashmi, A. M. Khedr, I. Arif and M. El Bannany, "Using a Hybrid-Classification Method to Analyze Twitter Data During Critical Events," in IEEE Access, vol. 9, pp. 141023-141035, 2021, doi: 10.1109/ACCESS.2021.3119063.

[2] B. Venkatesh and H. N. Vishwas, "Real Time Sarcasm Detection on Twitter using Ensemble Methods," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1292-1297, doi: 10.1109/ICIRCA51532.2021.9544841.

[3] R. Kumar and A. Bhat, "An Analysis On Sarcasm Detection Over Twitter During COVID-19," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-6, doi: 10.1109/INCET51464.2021.9456392.

[4] J. Godara and R. Aron, "Support Vector Machine Classifier with Principal Component Analysis and K Mean for Sarcasm Detection," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 571-576, doi: 10.1109/ICACCS51430.2021.9442033.

[5] M. V. Rao and S. C., "Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analysis," 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2021, pp. 196-199, doi: 10.1109/WiSPNET51692.2021.9419432.

[6] 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), 2021, pp. 968-972, doi: 10.1109/ICACITE51222.2021.9404585.

[7] H. Nguyen, J. Moon, N. Paul and S. S. Gokhale, "Sarcasm Detection in Politically Motivated Social Media Content," 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2021, pp. 1538-1545, doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00207.

[8] Po Hung Lai, Jia Yu Chan and Kim On Chin. “Ensembles for Text-Based Sarcasm Detection”, 2021 IEEE 19th Student Conference on Research and Development (SCOReD), 2021, pp. 284-289, doi: 10.1109/SCOReD53546.2021.9652768.

[9] M. Zanchak, V. Vysotska and S. Albota, "The Sarcasm Detection in News Headlines Based on Machine Learning Technology," 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 2021, pp. 131-137, doi: 10.1109/CSIT52700.2021.9648710.

[10] Dirash AR and Manju bargavi S.K, “LSTM Based Sentiment Analysis” International Journal of Trend in Scientific Research and Development, Vol.5 No. 4, June 2021, pp. 728-732