SARCASM DETECTION

Neha Verma[1], Ayushi Gupta[2],
Sahil Bhat[3], Madhur Jain[4]
(Computer Science and Engineering, NIET Greater Noida, U.P)

Ms. Renuka Sharma(Assistant Professor ,CSE)
NIET , Greater Noida U.P

ABSTRACT - The methods and process of detecting sarcasm, as well as the comparison of results from various models and datasets, are discussed in this paper. The term "sarcasm" refers to phrases that convey a meaning that is in opposition to the intended meaning. Recently, NLP has become a subject of great interest for researchers due to its captivating nature. One aspect of NLP that has garnered particular attention is sarcasm detection. This process, similar to sentiment analysis, uses mathematical methods to classify the tone of a text or phrase, determining whether it contains sarcastic elements or not.

On social media, people use sarcasm to covertly communicate their opinions and more intense feelings. Since it necessitates a significant amount of background knowledge, the rhetoric of irony is a subfield of sentiment analysis that is indistinguishable using conventional sentiment analysis methods. The primary aim of present-day sarcasm detection methods is to scrutinize the textual material of sarcasm through a range of natural language processing approaches. Sarcasm involves using language in a derisive or satirical way to ridicule a person or thing. In sarcasm, ridicule or satire is employed harshly, often coarsely, and with disdain for unfavorable consequences. It can be challenging to decipher the true meaning of a sentence in code-mixed language because there aren't enough sarcastic cues. Therefore, the sarcasm detection research that we suggest in this paper uses a variety of techniques and their improvements. This section has examined the Glove Vector[1], Fast Text, Bag-of-Words, Term Frequency-Inverse Document Frequency, (Continuous Bag of Words/Skip Gram), and BERT feature extraction techniques. We use mockery that is past identification to illustrate significant datasets, tactics, patterns, problems, and tasks. To aid researchers in related fields to grasp the cutting-edge practices in sarcasm detection, our research presents succinct tables of sarcasm datasets, features utilized for sarcasm detection, and techniques for their extraction.We performed experiments on various datasets that are accessible to the public, and the outcomes indicate that our suggested approach can greatly improve the precision of identifying sarcasm.

Keywords - sarcasm detection, sentimental analysis, deep learning, support vector machine

1.INTRODUCTION

Sarcasm can manifest in various forms, from facial expressions and gestures to text. It is dependent on a person, the situation, the language, and even culture. It often involves both positive and negative comments. With NLP and IoMT-related fields prioritizing sentiment and sarcasm analysis, it is possible to integrate these into devices and sensors. This will lead to a cost-effective and intelligent approach for human-to-device interaction. Despite not having a clear definition, sarcasm is generally viewed as a rhetorical technique of expressing negativity through positivity. The goal of the sarcasm detection task, also known as the binary classification problem, is to recognize irony or sarcasm at the sentence or document level. Sarcasm is difficult to identify in practice and cannot be detected using traditional sentiment analysis techniques. The current approaches to sarcasm detection concentrate on the target text's content and are based on intricate linguistic knowledge. Sarcasm can be tricky to spot, especially when there is a mix of codes involved because it can cause one to misinterpret the statement's intended meaning.[2] To fully understand a sentence's meaning, it is essential to be able to spot sarcasm. In multilingual societies like India, where there are 22 official languages and people communicate in both English and their native tongues, including Hindi English, the most frequently used code-mixed language in the Indian context, code-mixing is common. Numerous studies have been conducted in recent years as a result of the prevalence of code-mixed sarcasm on social media. Sarcasm is not just about causing harm, it also adds humor to our daily language use. Implicit sarcasm is always present in our daily interactions.

With a touch of heavy sarcasm, he gazed at the blank notebook and proclaimed, "What a masterpiece of an essay you've penned." This statement is sarcastic as the surface sentiment is positive while the underlying sentiment is negative. Emojis have become a popular subject of discussion as they offer an effective means of conveying sentiment and emotion. Moreover, they simplify the process of comprehending the implicit sentiment and emotion in an utterance. It can be hypothesized that emojis could aid in determining if an utterance has any intended sarcasm as sarcasm is strongly tied to comprehending implicit sentiment and emotion.

.

2. LITERATURE SURVEY

Sarcasm

The use of machine learning, deep learning, and hybrid approaches for sarcastic post detection has recently attracted increasing interest from researchers. While other researchers used deep learning-based techniques, some used linguistic cues from the text to identify sarcasm. Castro and associates. Introduced the MUSTARD dataset, which combines text, audio, and visual modalities, and demonstrated that multimodal data outperforms unimodal data at sarcasm detection. Chauhan and other people. Due to the fact that sarcasm is inherently implicit, [1] manually annotated the MUSTARD dataset with implicit and explicit sentiment and emotion. Among others, Attardo. Data was gathered from a television comedy show and multimodal stimuli were studied to identify irony and sarcasm markers.

Sentiment

The detection of a piece of text's emotion or feeling, known as sentiment analysis, is crucial in making informed decisions in various scenarios. However, sentiment analysis can be challenging. Recently, people have turned to using video messages to convey their emotions, making sentiment recognition easier with the use of multiple modalities. Kaur et al. explored the significance of using multiple modes in sentiment analysis, and Perez et al. highlighted the benefits of utilizing visual, acoustic, and linguistic information to enhance accuracy. A study proposing a fusion model for sentiment analysis that only utilizes visual and acoustic features was also conducted. The CMU-MOSEI dataset was used in this research.

Emotion

Emotion recognition is a significant task that can provide valuable insights, such as understanding a person's behavior or generating appropriate responses for conversational AI. While text-based methods may not be sufficient for accurate emotion detection, the use of multiple modalities, such as audio and visual cues, can enhance the results. Research has shown that incorporating multiple sources of information can lead to improved emotion recognition performance. For example, Akhtar et al. proposed a framework that considers both sentiment and emotion using a multimodal dataset.

Emoji

The use of emojis has become increasingly popular, leading to the development of various related tasks such as predicting emojis in instant messages, disambiguating emoji meanings, comprehending crisis events, constructing emotion classifiers, recognizing sentiments, and generating emotional responses. For single-label emoji prediction, some researchers have investigated LSTM-based models, as in the case of Barbieri et al. Additionally, a RNN method incorporating emojis was used to investigate how people express solidarity on social media, including how emojis spread over time and space. Jin and associates

METHODOLOGY

In this section, a comprehensive overview of the models for sarcasm detection is presented. We will commence the discussion by delving into the technology used.

A. Bag of Words Model

The bag-of-words model is a commonly used representation technique for item categorization. This model aims to convert each significant aspect into a visual word, and then represent each image with a histogram of these words. Typically, clustering methods like K-means are used to generate the visual words. A heuristic clustering method has added complexity, which may be why the bag-of-words representation's properties have received little theoretical attention despite the representation's success in several experiments.[3] In contrast to clustering, our research provides a statistical framework that builds on the bag-of-words representation, creating visual words statistically while producing empirical results that are comparable to those of the clustering approach. Two algorithms that compete favorably in object categorization without using clustering are also included in the framework, along with a theoretical analysis based on statistical consistency.

B. Term Frequency- Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical method widely utilized in information retrieval and natural language processing (NLP) that gauges the relevance of a term in a document with respect to a collection of documents. This approach entails transforming the words in a text document into numeric values. Text vectorization commonly uses the TF-IDF scoring technique. The algorithm determines the weight of a word by dividing its term frequency (TF) by its inverse document frequency (IDF). The IDF tracks how many files the term appears in, allowing it to adjust the weight of each word. The TF represents the frequency of a word in a document. If a word appears in several files, it is regarded as unimportant.

C. Word2Vec (Continuous Bag of Words/Skip Gram)

Word embedding, a fascinating development in NLP, is the process of representing words with low-dimensional vectors while preserving their contextual similarities. Prediction-based and frequency-based word embedding models are both available. A well-liked prediction-based embedding model is Word2Vec, created by Mikolov.[5] Both the Skip-Gram and Continuous Bag of Words (CBOW) methods are utilized, with the former predicting the surrounding context of each word, while the latter forecasts the probability of a word based on a given context word or group of words.

D. GloVe Vector

A technique called GloVe uses unsupervised learning to produce vector representations of words. In order to reveal linear substructures in the vector representation of words, it trains on corpus-based global word-word co-occurrence statistics.[6] The Continuous Bag of Words Model and the Skip-Gram models' techniques are combined in the GloVe model, which was created by Stanford researchers. It uses word similarity as an invariant to build word vector representations. The former model is computationally expensive despite its high accuracy, while the latter has poor accuracy but is computationally efficient. GloVe combines the benefits of both models and has been shown to be more accurate and effective.[7]

CONCLUSION

The research paper on "Sarcasm Detection" highlights the importance and challenges of detecting sarcasm in natural language. The study presents various approaches and techniques used for sarcasm detection, including lexical, semantic, and contextual analysis. The results of these methods show that while there has been significant progress in detecting sarcasm, the task remains challenging due to the complex and nuanced nature of sarcasm. However, the development of advanced natural language processing techniques and the growing availability of annotated data sets have the potential to significantly improve the accuracy of sarcasm detection in the future. The findings of this research emphasize the need for continued work in this field to develop robust and reliable methods for detecting sarcasm.

explore more advanced methods of analysis, such as deep learning, to better understand the potential of geolocational data.

REFERENCES

1.Zhang, Y., Jin, R. & Zhou, ZH. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. & Cyber. 1, 43–52 (2010). https://doi.org/10.1007/s13042-010-0001-0

2.A. Hakim, A. Erwin, K. I. Eng, M. Galinium and W. Muliady, "Automated document classification and Electrical Engineering (ICITEE), 2014, pp. 1-4, doi: 10.1109/ICITEED.2014.7007894.

3.Mon, Aye Myat, and Khin Mar Soe. Clustering analogous words in myanmar language using word embedding model. Diss. MERAL Portal, 2019.

4. Y. Sharma, G. Agrawal, P. Jain and T. Kumar, "Vector representation of words for sentiment analysis using GloVe

5.https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse- document-frequency/

6.https://www.analyticsvidhya.com/blog/2017/07/word-representations- text-classification-using-fasttext-nlp-facebook/

7.Vithyatheri Govindan, Vimala Balakrishnan,‖A machine learning approach