Rule Based Approach for Contextual Classification of Twitter Dataset

Mr. L.K. Ahire
Dept of Computer Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, India
Email:lomesh.ahire@gmail.com

Dr. S. D. Babar
Dept of Computer Engineering, Sinhgad Institute of Technology, Lonavala , Savitribai Phule Pune University, Pune, India
Email: sdbabar@sinhgad.edu

Dr. P. N. Mahalle
Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune, India
Email: parikshit.mahalle@viit.ac.in

----------ABSTRACT----------
The amount of data on social networks and the number of users has been growing quickly in recent years. Any time an event or activity occurs nearby, nearby individuals express their thoughts and reactions on social media. When a new product is introduced, users on social media platforms also comment on it. It is challenging to ascertain the genuine state of emotions because of sophisticated ways of presenting various perspectives. Sarcasm is the use of words to convey a negative emotion in a humorous way. Machines have an extremely difficult time comprehending and recognizing these caustic remarks when trying to discern sarcasm from text, it helps to know the context of the content. In this research, we propose a novel method called the Rule Based Approach for Contextual Classification (RBACC), which uses the context of tweets to identify sarcasm using a variety of already available methods. RBACC uses four features that were taken from tweets that were acquired using the tweeter API, and rule-based evaluation is done using the linguistic data of the four characteristics. The RBACC technique ensures flexibility and energy efficiency, according to experimentation. The RBACC technique is also scalable because performance and functionality are unaffected by an increase in the quantity of tweets. Results demonstrate that RBACC accurately identifies the context of text when given a variety of datasets that contain both type of data balanced as well as imbalanced.

Keyword: Automated systems, Context, Fuzzy Rules, sarcasm, Rules, social network, etc.

Date of Submission: 14/10/2023 Date of Acceptence:15/11/2023

I. Introduction

Ambient computing is a technology that performs calculations for individuals without an instant instruction. This contrasts sharply with smartphones and smart watches, which require considerable consideration before use. Many computer systems rely on active inputs from people. For instance, if one of your friends searching for a movie schedule on his smartphone, he would type the matching name of the movie or a person who is actor into the Google search bar.

We shall surpass our environment and technology by applying intelligent environments. We will gain from this in numerous ways, including cost savings for corporate executives, enhanced staff collaboration, and increased flexibility.

Since social media platforms like Twitter, Facebook, WhatsApp, and others are so popular for online idea exchange and user feedback, business, politics, entertainment, and politics have all benefited from having online allies in recent years. These answers are a collection of thoughts, feelings, or viewpoints that may point us in the direction of an occasion, a

service, a start-up, or many other things [1]. Social networking sites have recently taken on a hugely significant role in people's lives and communities. These websites are a source of news, entertainment, and information about people's daily lives. They have also generated a lot of information, and this information(data) is used for various analyses.

Sentimental analysis (SA) is the process of classifying the content emotions conveyed, for instance as neutral, positive, and so forth. Sentiment Analysis has seen a surge in research due to the data that online networking has made available. Sarcasm is a specific kind of text that is difficult to distinguish from the phrases, which makes it problematic. The term "sentiment analysis" deals with automatically identifying opinion in messages that are in the form of text. Sarcasm research is a cutting-edge subfield of sentiment analysis. Sarcastic content has been found to make up about 11% of content on social networking sites. The use of caustic language is becoming more commonplace every day. It's crucial to understand how people feel about a thing, an element, or a man, as well as to be able to spot these sarcastic remarks properly.

Machine learning and language communication processing are both necessary for sarcasm study. Features or the environment that are crucial for sarcasm detection. Sarcasm detection is inherently challenging, and the structure and nature of content on Twitter make it even more challenging. In comparison to other, more conventional sources, such as headlines in news articles and also from some books, the online social platform like Twitter is more casual in nature with a more extensive use of slang words, condensing, and includes a point of confinement of 140 characters for each tweet, providing less word-level signs and consequently including more ambiguity.

There is a lot of research being done in the area of leveraging contextual features, where the contextual features of the text are used to detect sarcasm in addition to linguistic features and pattern-based approaches. The details that surround a particular issue are known as the context. Historical context and literary context are both related to context: what was going on at the time a work was produced, and how do those events affect how we interpret it? Although occasionally a broad context incorporates all four components, the historical context, cultural context, social context, and political contexts separately at this level. We use two tweets to illustrate the aforementioned claim and demonstrate how context plays a part in identification of sarcasm.

Tweet1: " EVM machines are very hard to hack because they are having simple design".

Tweet2: “It is very good to see that 60% Indian people elect their government”.

Tweet3: “I like politics to watch on news channels only”.

Tweet4: "Oh, very sad, May his soul rest in peace".

Tweet5: “Oh dear! Get well soon!!!”

Tweet6: “Happy Birthday Dear, God Bless You”.

Out of above six tweets, first three tweets make it quite evident that the supplied tweet is political in nature, and we can also see that they are satirical. We can conclude that the fourth and fifth tweets lack sarcasm and have a depressing tone. We can infer from the sixth tweet that it is about birthday wishes and does not contain any sarcasm. Considering the two circumstances, we can say that the likelihood of sarcasm is higher in a political environment than a depressing one. The context of text data is not defined by a single quantitative number. The concept of text context lacks a distinct boundary. Therefore, we must define traits that are open to a variety of alternatives using descriptive language [2]. According to our suggested methodology, we will first extract the context from the tweets before applying existing sarcasm detection techniques to find the sarcasm. proposed strategy the four features are extracted and computed from each tweet using the Rule Based Approach for Contextual Classification (RBACC) developed in this research, in accordance with a of fuzzy inference rule set and different membership functions used in fuzzy logic system [3–4]. We can determine how relevant a message is or not to a certain context by using the membership functions. The following is a list of this paper's main contributions:

Novel RBACC approach that has been proposed for text context identification.

1. From the tweeter data used as an input for RBACC, extract features.
2. To address the issue of unbalanced data, datasets are split into three datasets.
3. Using the characteristics and fuzzy rules, results are produced.

This essay is divided into the following sections. Section II present the related study in the area of contextual sarcasm detection. The various sources of data and the process of extraction of features is presented in Section III. An introduction to the proposed approach, a Rule Based Approach for Contextual Classification, presented in Section IV. Overall result analysis is given in Section V. The difficulties in context identification are described in Section VI. Section VII gives the conclusion and further work.

II. Related Work

Many academics have focused on features such lexicon-based sarcasm detection, collective machine learning, or ensemble approaches in recent years. Very few study attempt to establish the context when sarcasm in text is being detected. We have conducted some reviews in which the context is used to identify sarcasm in the text.

To address the issues raised earlier, C. I. Eke et al.[5] combine classic machine learning methods with deep learning, BERT and feature methodology that is identification of sarcasm which is context based, and feature extraction. The basic model put out relies on embedding dependent representation using Bi-LSTM, another RNN version, to produce word embedding and context, and this author employs Global Vector representation (GloVe). BERT (Bidirectional Encoder representation and Transformer) is the foundation of a different method. The third model was suggested using feature fusion and combines classical machine learning with the BERT feature, associated feelings, syntactic, and GloVe embedding feature.

H. Gregory et al. [6] implemented "LSTM", "GRU", and "transformer models" in this study and tested fresh approaches for categorizing sarcasm in tweets. In addition, the model combined several transformer models, including ALBERT, BERT, RoBERTa, XLNet, and RoBERTa-large.

In this work, D. Ghosh et al.[7] highlighted two concerns that can aid in the identification of sarcasm, including the utilization of conversation context and the activation of sarcastic responses as a component of conversation context.Long short term memory (LSTM) networks were offered by the authors as a way to represent the context of a tweeter's dialogue and the sarcastic reaction to that tweet.'

A feature selection ensemble-based method was used by K. Sundararajan and A. Palanisamy[8] to identify the best set of features to discern sarcasm in tweets, and an algorithm was created to assess if a tweet is sarcastic or not. After identifying sarcastic sentences, writers presented a multi-rule based strategy to determining the type of sarcasm. In this paper, writers conducted a preliminary effort to identify four different varieties of sarcasm, including furious, impolite, courteous, and deadpan sarcasm, with 92.7% accuracy.

For the purpose of identifying sarcasm from tweets expressed in Hindi, S. K. Bharti et al.[9] presented a pattern that is based on context and is "sarcasm as a contradiction between a tweet and the context of its related news". The proposed method successfully predicted tweets within the same timestamp with an accuracy of 87% using a dataset of Hindi news.

A specific method for automatic twitter contextualization was put forth by R. Belkaroui and R. Faiz[10] and uses tweet content from communications between users of social networks. The information included in text on Twitter is extremely scarce compared to standard contextualization algorithms, which take into account solely text data. This is because Twitter combines a range of signals, including social, temporal, and linguistic information. The authors assert that the results of their studies can verify the benefits of their suggested strategy and guarantee that a particular tweet generates contexts that contain the necessary information.

In the suggested method, K. Pant and T. Dadu[14] employ the RoBERTalarge algorithm to detect sarcasm in both datasets. The authors assert that by utilizing three different types of data inputs—Response-only, Context-Response, and Separated Context-Response—context can be used to enhance the effectiveness of contextual word embedding based models. The authors further assert that the Reddit dataset's F1-score is increased by 5.13 percent by the addition of a token separation between the context of text and the target response. However, it should be highlighted that all of the aforementioned models are adequate for the modern computing environment. While no attempt has been made to determine the context of text data in tweets, researchers have used linguistic features, syntactic features, pattern-based approaches, and other techniques to detect sarcasm in all the work shown above. In the future study, this detected context will be related to sarcasm detection using the fuzzy approach proposed in this paper to determine the context of text data in tweets.

III. Data and Feature Extraction

1. Data Collection

The tweeter application programming interface (API) was used to get the initial data, which consists of 9 million tweets. Users of the popular social networking site Twitter can publish succinct messages of up to 140 characters. Each tweet includes a timestamp, its unique ID, coordinates, and text information. To do this, we had to examine the tweet's text data..

2. Manual Labels and Analysis

The context information is provided in many other datasets, but tweeter does not provide any context information. Therefore, we construct RBACC using labeled data. 15 volunteers are needed for our proposed project's manual labeling of 600 tweets and the randomly chosen tweets from our old tweet dataset. Every tweet is given a grade of 0 or 1, with 0 denoting irrelevance and 1 denoting relevance. Each tweet's summation score, or "Z," is calculated according to equation I.

$$Z_0 = \mathbf{O}^{0}_{\text{DOD}} C_{00} \quad \text{---------- (I)}$$

,where C_{ij} belongs to {0,1} and C_{ij} is j^{th} tweet which is i^{th} volunteer scored.

There are 15 volunteers present, and the summing score runs from 0 to 15. The four specified score intervals—Z1 [0, 4], Z2 [4, 8], Z3 [8, 11], and Z4 [11, 15]—represent the degree of relevance of each tweet in the following ways:

Z_1 = irrelevant
Z_2 = low relevance
Z_3 = moderate relevance
Z_4 = highly relevance

3. Data Preprocessing

Each of the tweeter's tweets has its own unique qualities. Data preparation is therefore necessary for our goal. Since the text data in tweets is not accurate, we cannot use them effectively. They fully incorporate jargon from the internet and some background noise, such as URL, or site address. This undesirable piece of information could cause categorization and processing speed to operate incorrectly. "All trains running late #heavyrain#jimmy http://td.com/xsdyou13" is one example. In this case, the tweet is being cluttered by the URL, which begins with the letter http. We need to remove this kind of noise from tweets throughout the data cleaning procedure. Pattern matching is an extremely successful technique for this kind of noise removal. In pattern matching, potential patterns are looked for within a collection of sequential expressions. When this type of pattern is identified, the program will automatically erase the information that follows it. As in the example above, URLs have a set format that begins with "http://. Another problem is hashtags, which can include crucial information but make it difficult to retrieve the information because not all communications use them.

The information from the messages is extracted using pattern matching. To extract meaningful information, we occasionally need to match the words that follow a hashtag with the dictionary.

Stopwords like "a," "an," "the," "on," and other similar words are another undesired element in text data. The communications contain these words repeatedly, which results in information that is meaningless. Therefore, stopwords must be cleaned.

Lower case conversion of message words is another step in the preprocessing process to prevent machine misunderstanding. For example, the terms rain and RAIN are the same, but owing to case differences, the machine interprets them differently. After being changed to lower case, the computer interprets them as the same.

4. Feature Extraction

It is crucial to realize that some words are used more frequently than others when attempting to decipher the content of the tweets. For instance, the terms "EVM", "voting", "Rally", and "Party" are widely used in political discourse. The tweets from Z_2, Z_3 and Z_4 are chosen from our training set using this suggestion to obtain the top 50 frequently used terms.

Equation II calculates the word importance, where the letter 'μ_i' represents the percentage importance of each word. [4]

$$\mu_i = \frac{O_0}{D_0} x\ 100\ \% \qquad \text{(II)}$$

where P_i represents the number of words present in the tweet which are in Z_2, Z_3 and Z_4 and T_i is the number of total words in Z_1, Z_2 , Z_3 and Z_4. Number 'μ_i' indicates the "importance of word i" in terms of percentage. The importance of word i is more if μ_i is more. After that, we arrange the words in terms of 'μ_i' from high to low order and list L was built. Three equal subsets L1, L2 and L3 having distinct weights Θ_1, Θ_2 and Θ_3 of list L are created The similarity function is computed as given in equation III is

introduced using Natural Language Toolkit. The operator '∞' is used to evaluate similarity as given in below equation:

$$S_i = \max_{DDDDO} (W_k \ X \ t_i \infty W_k) \qquad \text{(III)}$$

where i belongs to {1, N}

$$\text{where } W_k = \begin{cases} \Theta 1 \text{ if k belongs to [1,16]} \\ \Theta 2 \text{ if k belongs to [16,32]} \\ \Theta 3 \text{ otherwise} \end{cases}$$

The j^{th} tweet having n number of words, the i^{th} word in this tweet is represented by t_i. The k^{th} word in list L is W_k. Equation I gives high similarity score of t_i. We get 4 features are from tweet depend on the value S_i. The proposed model RBACC use 4 features [4]. Different scores of features extracted from the tweets are given by equation IV,V and VI. The features described as follows:

1. The highest word score in the j^{th} tweet

$$G_j = \max_{DDDDO} S_i \qquad \text{(IV)}$$

where, the highest word score is given by G_j in the j^{th} tweet.

2. The total score for j^{th} tweet

$$Y_0 = \sum_{0OD}^{0} S_0 \qquad \text{(V)}$$

where Y_0 indicates the total score.

3. The total length of j^{th} tweet.(l_j)

$$l_j = n \qquad \text{(VI)}$$

where n is the number of words in the j^{th} tweet.

4. The number of all words which occurs frequently in the j^{th} tweet.(F_j)

feature F_j indicates the count of words in the j^{th} tweet is equal to the count of words in the list L.While calculating F_j we have to use list L to compare with all tweets.

IV. Proposed Rule Based Approach for Contextual Classification (RBACC)

We will go into depth about our suggested model, which is a Rule Based Approach for Contextual Classification (RBACC), in the following section.

The proposed model's general flow is shown in Fig. 1. As previously stated, it is impossible to forecast the precise or accurate context of a tweet using the procedure of context identification of a tweet. Accordingly, the fuzzy inference system can be utilized to create a system in which inference rules are applied in order to determine the context of a tweet. As inputs for the suggested RBACC model, we employ the four attributes listed above in section 3. Fuzzification is the process of transforming true

or raw inputs into fuzzy sets with parts that have varying degrees of membership based on membership functions.

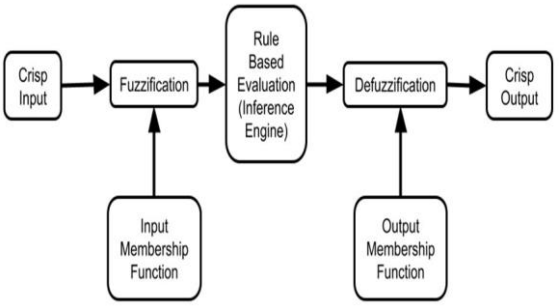

Fig. 1 RBACC Mechanism

We chose function for this article since it is quite simple and frequently used. If-THEN fuzzy rules are employed in the inference engine to create the relation in between the provided input and output, converting the supplied input which is crisp to the provided crisp output. The linguistic values contained in the fuzzy rules are organized according to human expertise information and empirical norms [3]. Numerous defuzzification strategies, which includes centroid strategy, bisector method, mean of maximum (MOM) strategy, smallest of maximum (SOM) strategy, and biggest of maximum (LOM) strategy are discussed in various literatures [16]. An aggregated set of fuzzy values that include output values is used to obtain the outcome "R" as a value through the process of defuzzification. The suggested model was defuzzied using the centroid approach. The suggested model's crisp output was created using the centroid defuzzification method.

Algorithm (1): Fuzzification and Defuzzification
Input: Feature vectors for each tweet in the preprocessed training data comprise four features..
Output: Tweet's context.

Step 1: Fuzzy rules are created using training data that has been previously processed.

Step 2: Fuzzification process

2.1 Choose the proper membership function.

2.2 To determine the degree of membership for every value in the feature vector, use the membership functions.

2.3 It is possible to map the fuzzy set to the exact input.

2.4 Create a fresh degree of membership.

Step 3: Process of Evaluation

3.1 Create a set of IF-THEN fuzzy rules

3.2 This phase will additionally include fuzzy rules in addition to the extracted fuzzy rules from step 1.

Step 4: Defuzzification procedure

1: Choose the Centroid Defuzzification function.

2: Determine the actual value of findings that are hazy.

Step 5: Display the tweet's context and the precise result value in step five.

A. Parameters

Table 1 lists the four linguistic inputs (and one output), together with the range of values for each parameter. We must observe that the changing range of parameters are distinct in this case. Taken as an example, the tweet with the greatest word score (G) is characterized as having 5 degrees: very low (G [0, 0.35], low (G [0.15, 0.45], moderate (G [0.25, 0.55], high (G [0.4, 0.7], and very high (G [0.6, 0.1]).

Table 1 Input and Output parameters

Variable	Linguistic Variable	Range	Linguistic Value	Parameter
Input	G (Word Score)	0-1	Very Low	0-0.35
			Low	0.15-0.45
			Moderate	0.25-0.55
	Y (Tweet Score)	0-20	Very Low	0-3.5
			Low	2-8
			Moderate	5-11
	l (Length)	0-20	Short	0-8
			Moderate	6-15
			Long	13-20

	F (Word Frequency)	0-8	Low	0-4
			Moderate	3-6
			High	4-8
Output	R	0-100	Irrelevant	0-35
			Lowest relevant	30-60
			Moderate relevant	45-80
			Highest relevant	75-100

B. Evaluation based on Rules

Fuzzy rules are created using IF-THEN statements and applied knowledge. The fuzzy rule is made up of IF-THEN statements and comprises an IF-THEN condition and conclusion. They are relatively simple for us to express since they resemble plain language reasoning. We develop several number of fuzzy rules to get our findings. The following is how we determine

Fuzzy Rules = Total Inputs * Linguistic variables count

We can create 64 alternative fuzzy rules in our scenario because there are 4 inputs and 16 linguistic variables. For a more detailed illustration, the ensuing rules are used:

1) R is highly relevant if G and F are both high.
2) R is of moderate consequence if G is high and l is short.
3. R has low relevance if F and Y are both low.
4) R is not relevant (irrelevant) if G is very low and l is high.

We go on to detail the straightforward guidelines mentioned above. An indication that a tweet is pertinent to the context is the quantity of commonly used words in the tweet. A user likely incorporated some crucial words in their succinct message if a tweet has a greater number of word count and a less length. There are keywords with low importance in tweets because the weight of the tweets and commonly used words is low; as a result, it is seen as a tweet with less impact. Additionally, if a tweet doesn't contain important terms, classification is meaningless for that tweet.

V. Results and Discussion

A. Test of the proposed method

From the tweets, we extract four features, which we then employ as input along with their membership functions. Each linguistic variable within a certain range is assigned a linguistic value. The four inputs and their membership functions are mapped in Figure 2. In the given range, the result contains four linguistic variables. One output with four linguistic values is available. The relation between output function of membership and their linguistic values is shown in Figure 3.

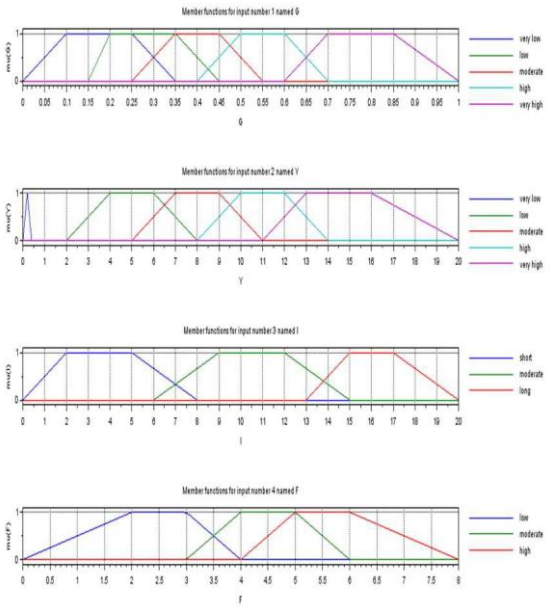

Fig .2 Input Membership Functions Mapping

Four linguistic values—irrelevant (0–35), low relevant (30–60), moderate relevant (45–80), and high relevant (75–100)—are available for the linguistic variable for output R. A highly relevant

result in the range of 75 to 80 is obtained when the linguistic variable is mapped to its linguistic value, as shown in figure 3. Figures 4 and 5 illustrate the output as a rule viewer for the rules we presented in section IV(B) and the surface viewer for the result generated by the rules using the linguistic values of the input parameters, respectively.

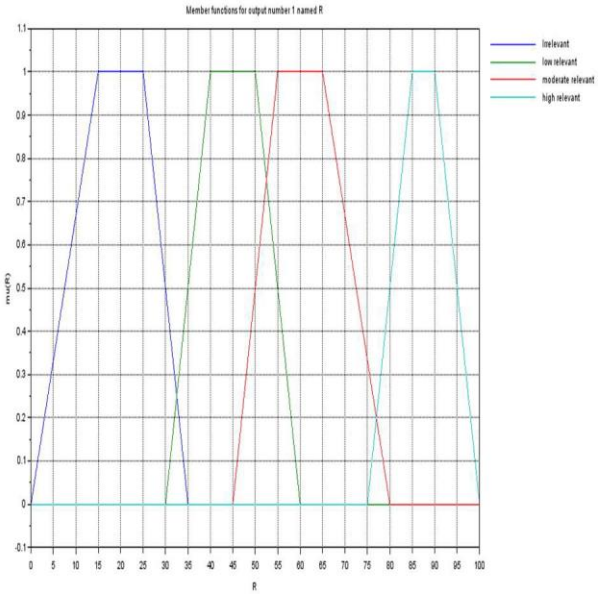

Fig. 3 Member function Mapping

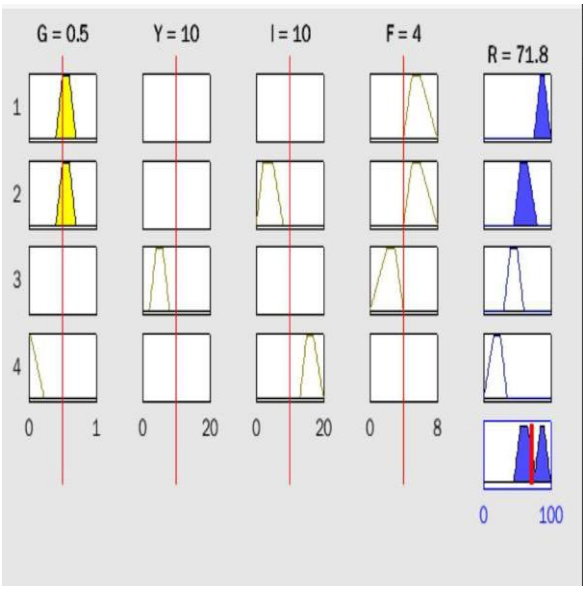

Fig. 4 Rule viewer Output

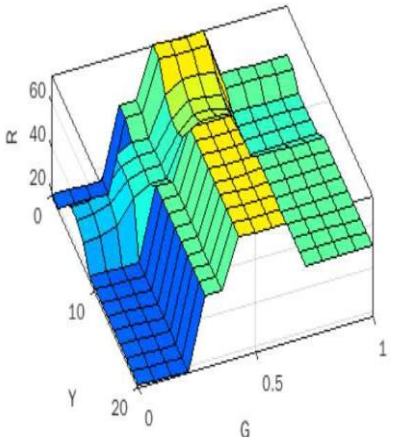

Fig. 5 Surface viewer output

One of the most difficult duties is confirming the correctness rate. People frequently draw various conclusions from the same facts and event context, leading to diverse answers. So, 600 tweets were manually labelled and compared to the outcomes determined automatically. We chose half of the tweets from the original dataset that were "relevant" and the other half that were "irrelevant" on purpose. Then, 15 volunteers are asked to grade these tweets using our RBACC on a scale of 0 to 4. After this procedure is complete, 291 tweets—including those with low, moderate, and high relevancy—are relevant, whereas 309 are irrelevant. We obtain three distinct testing datasets from them to broadly validate the model. Each dataset has 200 tweets. The sole difference between them is the relevance-to-irrelevance ratio, which is 1:1, 1:9, and 9:1 in each instance. According to a thorough examination of the first dataset, there are 100 relevant and 100 irrelevant tweets, 20 relevant and 180 irrelevant tweets, and the third dataset contradicts the first. The design contains both balance and disharmony. When there are less tweets from a positive class (positive) than a negative class (negative) in machine learning, this is known as the imbalance problem. The accuracy of a relevance problem with four degrees is shown in Table 2.

Table 2's results demonstrate that RBACC's results have higher accuracy levels for the four-degree relevance problem. The centroid approach, employed for defuzzification, produces very good results. We attempt to address the issue of imbalanced data as the dataset is split into balanced and unbalanced data, and the findings demonstrate that RBACC provides accurate results on the similar datasets.

Table 2 Four-degree accuracy difficulty for relevance.

Defuzzification Method	Relevance	First Dataset (%)	Second Dataset (%)	Third Dataset (%)
Centroid	Irrelevant	100	99.3	100
	Lowest Irrelevant	57.2	5	63
	Moderate Irrelevant	62.3	75.2	56.8
	Highest Irrelevant	100	100	98

B. Comparing the keyword search approach

The research [22–24] retrieve pertinent tweets from the original dataset using a keyword search technique. Its benefit is simple, efficient, and quite accurate for those extremely pertinent tweets. However, this method's flaw is that it can't extract enough relevant tweets due to the keyword list's limitations. We choose to compare the suggested fuzzy logic-based model with a keyword search approach, taking quantity and correctness rate into account. For this experiment, five sets of tweets from the original dataset are ready. Ten thousand tweets are chosen at random from each data collection. We employ the same keyword list as in [22] and recognize the keyword search technique. Table 3 displays the comparative studies of quantity and accuracy rate. In this experiment, we apply the polar relevance classification method, where a fuzzy logic-based model delivers tweets that are lowly, moderately, and highly relevant. Since it is impossible to determine how many irrelevant and relevant tweets were successfully recognized, a correctness rate is defined as follows:

$$\Gamma = \frac{o}{0} X\ 100 \qquad \text{VII}$$

where X represents the number of tweets in Y that were correctly identified, and Y is the total number of relevant tweets collected by each method. Keep in mind that X is determined by hand double checking; that is, we look at the number of tweets that are accurately classified in Y. Furthermore, an incremental rate λ indicates that the suggested model is able to utilize a greater amount of information compared to the keyword search approach, which is described as:

$$\lambda = \frac{oooD}{oD} X\ 100 \qquad \text{VIII}$$

When the proposed fuzzy logic-based model calculates Xf, and the keyword search approach calculates Xk.

Through a manual examination and analysis of the results generated by each approach, we discover that every tweet retrieved using a keyword search appears in the results produced by the suggested way. Specifically, the former's findings are the latter's subsets. In Table 3, The values of λ show that more tweets are successfully mined by the latter than by the former. In summary, compared to keyword search, the fuzzy logic-based methodology may extract a significantly higher number of tweets. The fuzzy logic-based approach is more powerful than the latter when only the quantity is taken into account. When accuracy rate is taken into account, the keyword search approach performs somewhat better than the previous one. Taking into account both factors, we assert that the fuzzy logic-based model is the better choice in research settings like [22–24], where it is greatly desirable to have more relevant tweets for the analysis stage. More accurate and helpful data can be ensured by having a high quantity and accuracy rate.

Table 3 Comparison Results of 2 methods

Data set	Method						λ
	Keyword Based Search			RBACC Mechanism			
	Y	X	Γ	Y	X	Γ	
1	98	96	97.9	141	135	95.7	40.6

2	103	103	100	161	157	97.7	52.4
3	86	86	100	128	126	98.4	46.5

VI. Contextual Identification Challenges

1. Handling single word polysemy is quite difficult. A word's meaning might vary depending on the context in which it is used. In the text message "I am waiting for you near the bank," the term "bank" has two different meanings: "river bank" and "financial institution."

2. Microblogging messages are typically short and illegible. It is challenging to classify them when researched independently. Using a contextual or discursive analysis, text ambiguity can be resolved. For instance, the wording "Stop yourself" is confusing. Here, advice and criticism both serve as indicators of positive and negative polarity.

3. Identifying implicit or concealed attitudes in literature can occasionally be challenging. Understanding conversation is the key to the solution. For illustration, the sentence "I made genuine attempts. Implicitly expressed is "Now I accept the outcome without any resentment."

4. It's crucial to remember that emotional polarity vary depending on the situation. Take the phrase "It will rain tomorrow" as an example. This text is favorable when read in the context of agriculture, but it turns negative when read in the context of a cricket match.

5. When the context was known, neutrality took on a positive or negative connotation. Classification is extremely difficult without context. As an illustration, the word "unpredictable" has a neutral polarity. It is regarded as having a negative orientation whenever it is applied to a person's behavior. However, it is regarded as having a positive orientation when applied to a movie plot.

6. Since sarcasm and irony don't express explicit sentiment, further details are needed to understand both language usages. As in the line "What a great host!" for example, we are unsure of the sentiments. The information of the hosting event was only used to comprehend the feeling that was kept to oneself.

VII. Conclusion and Future Work

The proposed work provides a Rule Based Approach for Contextual Classification (RBACC) that makes use of information gathered from social networking site (TWEETER) and media from other sources. Four features from each tweet are utilized to extract the input parameters for an RBACC, which is built using the manually labeled data. We compare it with the popular keyword search technique. The findings show that the suggested fuzzy logic-based method performs better at distinguishing between tweets that are relevant and those that are not. Since the centroid approach is more successful and efficient than alternative techniques, we employ it for defuzzification. The suggested Rule Based Approach for Contextual Classification (RBACC) is therefore efficiently adapted to identify weather the tweets are relevant or not relevant based on the situations. The relevance of tweet decides the different context of the tweet.

Future research will be carried out to improve the formula that determines how similar two words or two tweets are. The set of rules must be improved in order to produce a decent outcome on the 4 degree relevance problem. Additionally, by utilizing NLP techniques like stemming, lemmatization, and sentiment analysis, data pretreatment can be improved.

REFERENCES

[1] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.

[2] P. N. Mahalle, P. A. Thakre, N. R. Prasad, and R. Prasad, “A Fuzzy Approach to Trust Based Access Control in Internet of Things Vk l In _ ;: l,” pp. 2–6, 2013.

[3] T. J. Procyk and E. H. Mamdani, “A linguistic self-organizing process controller,” Automatica, vol. 15, no. 1, pp. 15–30, 1979, doi: 10.1016/0005-1098(79)90084-0.

[4] K. Wu, M. Zhou, X. S. Lu, and L. Huang, “A Fuzzy Logic-Based Text Classification Method for Social Media Data,” 2017.

[5] 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," IEEE Access, vol. 9, pp. 48501–48518, 2021, doi: 10.1109/ACCESS.2021.3068323.

[6] H. Gregory, S. Li, P. Mohammadi, N. Tarn, R. Draelos, and C. Rudin, "A Transformer Approach to Contextual Sarcasm Detection in Twitter," pp. 270–275, 2020, doi: 10.18653/v1/2020.figlang-1.37.

[7] D. Ghosh, A. R. Fabbri, and S. Muresan, "The role of conversation context for sarcasm detection in online interactions," SIGDIAL 2017 - 18th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue, Proc. Conf., no. August 2018, pp. 186–196, 2017, doi: 10.18653/v1/w17-5523.

[8] K. Sundararajan and A. Palanisamy, "Multi-rule based ensemble feature selection model for sarcasm type detection in Twitter," Comput. Intell. Neurosci., vol. 2020, 2020, doi: 10.1155/2020/2860479.

[9] S. K. Bharti, K. S. Babu, and R. Raman, "Context-based Sarcasm Detection in Hindi Tweets," 2017 9th Int. Conf. Adv. Pattern Recognition, ICAPR 2017, pp. 410–415, 2018, doi: 10.1109/ICAPR.2017.8593198.

[10] R. Belkaroui and R. Faiz, "Conversational based method for tweet contextualization," Vietnam J. Comput. Sci., vol. 4, no. 4, pp. 223–232, 2017, doi: 10.1007/s40595-016-0092-y.

[11] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter," Proc. 9th Int. Conf. Web Soc. Media, ICWSM 2015, pp. 574–577, 2015.

[12] R. Belkaroui and R. Faiz, "Towards events tweet contextualization using social influence model and users conversations," ACM Int. Conf. Proceeding Ser., vol. 13-15-July, no. April 2016, 2015, doi: 10.1145/2797115.2797134.

[13] N. Malave and S. N. Dhage, Sarcasm detection on twitter: User behavior approach, vol. 910. Springer Singapore, 2020.

[14] K. Pant and T. Dadu, "Sarcasm Detection using Context Separators in Online Discourse," arXiv, no. 2011, pp. 51–55, 2020, doi: 10.18653/v1/2020.figlang-1.6.

[15] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "Cascade: Contextual sarcasm detection in online discussion forums," arXiv, 2018.

[16] H. Hellendoorn and C. Thomax, "Defuzzification in fuzzy controllers," Journal of Intelligent & Fuzzy Systems, vol. 1, no. 2, pp.109-123, 1993.

[17] Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint arXiv:1610.08815.

[18] S. Amir, B. C. Wallace, H. Lyu, P. Carvalho, and M. J. Silva, "Modelling context with user embeddings for sarcasm detection in social media," CoNLL 2016 - 20th SIGNLL Conf. Comput. Nat. Lang. Learn. Proc., no. December, pp. 167–177, 2016, doi: 10.18653/v1/k16-1017.

[19] A. Ghosh and T. Veale, "Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., no. September, pp. 482–491, 2017, doi: 10.18653/v1/d17-1050.

[20] D. Hutchison and J. C. Mitchell, "Web Information Systems Engineering –," Wise, vol. 2, pp. 232–246, 2010, doi: 10.1007/978-3-319-26190-4.

[21] M. Bouazizi and T. Otsuki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," IEEE Access, vol. 4, pp. 5477–5488, 2016, doi: 10.1109/ACCESS.2016.2594194.

[22] X. S. Lu and M. Zhou, "Analyzing the evolution of rare events via social media data and k-means clustering algorithm," In Proc. 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC), Mexico City, Mexico, April 28-30, 2016, pp. 1-6.

[23] X. Guan and C. Chen, "Using social media data to understand and assess disasters," Natural Hazards, vol. 74, pp.837-850, 2014.

[24] H. Dong, M. Halem, and S. Zhou, "Social media data analytics applied to hurricane sandy," in Proc. 2013 IEEE International Conference on Social Computing (SocialCom), Washington, DC, USA, September 8-14, 2013, pp. 963-966.

[25] H. Bagheri and M. J. Islam, "Sentiment analysis of twitter data," arXiv, no. October 2018, 2017, doi: 10.4018/ijhisi.2019040101.

[26] Ahire, L. K., Babar, S. D., & Mahalle, P. N. (2023). Fuzzy Approach for Context Identification into Ambient Computing. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 672-681.