International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 3, April 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Efficient Comment Classification through NLP and Fuzzy Classification

Shubham Derhgawen, Himaja Gogineni, Subhasish Chatterjee, Rajesh Tak

Information Technology, Dhole Patil College of Engineering, Pune, Maharashtra, India

ABSTRACT

A significant increase has been noticed in the number of people that are utilizing the internet paradigm for various purposes such as accessing various portals such as Social media and E-commerce websites. Due to their immense popularity, these platforms have seen remarkable growth and an increasing user base that is constantly interacting on the platform through the use of comments. These comments are mostly the users assisting each other on the platform in making the right decision. These comments can range from helpful to sarcastic, which can be highly difficult for a Natural Language Processing platform to determine. Supervised machine learning approach requires labels such as star ratings in reviews to understand the reviews and classify. These labels need to be reliable, whereas as they are entered by users, they could be misleading. Therefore, in this paper, an unsupervised approach towards the automatic classification of comments has been outlined in much detail. The proposed methodology utilizes an innovative combination of Term Frequency – Inverse Document Frequency (TF-IDF) in addition to the NLP paradigm along with the addition of the Entropy Estimation through Shannon Information gain. This procedure can effectively disintegrate the sentence into its basic form which can then ultimately be classified using the Fuzzy Classification technique.

KEYWORDS: *Natural Language Processing, Fuzzy Classification, Tf-IDF, Pearson Correlation*

How to cite this paper: Shubham Derhgawen | Himaja Gogineni | Subhasish Chatterjee | Rajesh Tak "Efficient Comment Classification through NLP and Fuzzy Classification" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-3, April 2020, pp.1000-1006, URL: www.ijtsrd.com/papers/ijtsrd30758.pdf

IJTSRD30758

I. INTRODUCTION

Language is one of the most innovative and novel concepts that have been originated due to the need for communication between humans. Language is a construct that is purely human and is not observed in various mammals and other animals. That is to say that there is communication between other animals, but it is not as sophisticated and as detailed as a language that is developed by humans. As there is a popular saying that the mother of all inventions is the need, similarly during the early days of human evolution, there was an extensive need to develop means of communication.

Communication was necessary to coordinate between different members of the same group while doing activities such as hunting-gathering and protecting against other predators. The language has evolved considerably from the early Hunter-gatherer days and has become extremely complex in this age and continues to evolve even further. Languages started fairly simple such as clicks signs and gestures that meant to convey intention and strategy to other humans. The language also developed as a means to convey and communicate items that were of necessity and the other items such as poisonous fruits and vegetables that should be avoided to stay alive.

Humans are social creatures. They need to have a refined vocabulary and an ever-increasing desire to convey their feelings and emotions to their fellow human beings. This is also demonstrated in the various ways that humans use language for entertainment, music, and communication. Language developed as a need to preserve valuable information and as a means to educate the younger ones, the ways of life. Most of the languages spoken across the world have a very similar structure, as most of them consist of sentences that are made up of even smaller words. These words have their individual meanings and when combined in a sentence can convey an expression or an emotion of a human being. Languages are highly Complex and can be very difficult for a computer to understand and comprehend.

Thus, for a computer to understand and comprehend language, it is a very extensive task that requires even more Complex processing which is similar to how the brain processes language. Therefore, the paradigm of artificial neural networks is one of the best applications that can be utilized to provide the computer with the ability to understand human language.

There has been an increasing amount of interest that has been emerging in this field of Natural Language Processing or NLP. The NLP paradigm is specifically designed for enabling human language comprehension for computers. This is due to the fact that the computer is a highly structured machine and due to the complex nature of language along with a set of various grammatical rules make it highly difficult for the computer to comprehend. This is where the NLP paradigm comes to the rescue. It performs various processes on the written language by making use

of paradigms such as Tokenization and application of stemming rules to segregate the written text into its quantized parts.

Stemming is a process by which the word is converted to its root form, this is because most words have different forms such as plurals, superlatives, and gender-specific words. It removes the extra clutter that the computer would have to process in order to get to the actual meaning of the sentence or the word. Stemming also reduces the processing time that is required for a computer to process a piece of text. This is highly useful as the computational power of the computer can be used for other useful work. All of these operations are highly necessary because English is a very complicated language. It has extensive rules that can only be understood by an experienced person in the language.

There is an increase in the number of researchers that have been working for the improvement in NLP or natural language processing paradigm. Due to the complicated nature of the language, this extremely large task needs to be divided into even smaller parts for effective and accurate processing by a computer. This act of dividing a large problem into smaller parts in the NLP paradigm is referred to as tokenization. Tokenization works by breaking down the complex and long words and sentences into smaller parts that can be individually understood by the computer and combined again to get the meaning of the whole sentence. Search processes along with other processes of NLP can help us to achieve nearly accurate representation of the human language by a computer.

One of the problems in understanding languages by computers is understanding sarcasm. Sarcasm is highly challenging as it is usually a very polarizing nature as well as highly contradictory. Therefore, it is almost impossible for a computer to detect sarcasm, as sometimes it is very difficult to be detected even by a human being. Sarcasm can be categorized as positive as well as negative in nature which makes it even more difficult and complex for humans and almost impossible for computers to comprehend. The analysis of sarcasm requires the system to develop an innate understanding of pop culture along with general knowledge of current events.

Therefore, sarcasm analysis is one of the most difficult as well as the useful methodology to understand the human language. It can help us understand and classify various trends through comments that are made online and other social media websites to gain an accurate understanding of the ongoing trends. The classification of sarcasm can lead to better experience online for different users as it would allow the administrators of the platform to remove objectively negative elements and users from spoiling other user's experience by their comments.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. Literature Survey

H. Yang proposes an approach to automatically identify the user requirements through application reviews and further classify them into two groups, functional and non-functional requirements, using a combination of (TF-IDF) and NLP technique with human intervention in keywords selection for requirements identification and classification [1]. They found out how the size of the sample reviews for the keywords affects the classification result of precision, recall and F-measure. By setting an appropriate size of the sample reviews a stable value for precision, recall, F-measure is received using this approach.

C. Zhou shows how CNN can be used to extract higher-level sequences of word features and LSTM to capture long-term dependencies over window feature sequences respectively [2]. CNN and LSTM have both been used independently for text classification. Authors explain how both can be used to achieve greater accuracy for both multiclass and binary classification, reaching 50% in five class classification and 87% in binary classification.

W. Jitsakul elaborates on the success of the E commerce platform which has been a highly useful addition to the internet platform as it has increased the convenience for the users and their shopping habits. But due to the inability to experience the product first hand and the limitations of the internet platform have led to the customers relying on the experience of the previous customers through reading the reviews. Therefore, the authors in this publication have proposed an efficient comment classification technique that has been simulated to ascertain its performance [3]. The major drawback in this paper is that the authors have achieved an accuracy of 80% which is quite less.

M. Andriansyah introduces the concept of comment classification on various different online platforms and social media networks. This is a highly useful and significant development in the social media platforms as there have been an increasing number of cyber bullying cases that are being reported every day. Therefore, the authors have implemented an innovative comment classification scheme that classifies the cyber bullying comments through the use of SVM or Support Vector Machine. The proposed methodology has been experimented on to achieve the performance analysis, which has claimed that the technique produces an accuracy of 79% [4]. The major drawback of this paper is the reduced accuracy of the whole system.

J. Savigny states that there has been an increase in the number of people utilizing video sharing websites such as YouTube all over the world. Most of the videos allow the people to interact with each other on the platform through the use of comments [5]. These comments can help communicate between various people of the same interest sharing therefore the author in this paper has utilized Natural Language Processing or NLP techniques to achieve emotion classification through the comments. The authors have implemented TF-IDF and utilized the convolutional neural networks to achieve their goals. The experimental results indicate that the proposed methodology is achieving accuracy of 76%.

Siswanto explains that MotoGP is one of the biggest Motorsport racing events that happen all over the world. A lot of people follow the events and competition and most of the people in this Era acquire their information about the MotoGP and the races through the internet [6]. Majority of

this information comes from social media websites and other interactive forums. These forums allow people to post their views in the form of comments and one of the most popular website is Twitter which is exactly the reason why the researchers have proposed a classification and analysis of MotoGP comments that are made on the Twitter social media through the use of naive Bayes algorithm and support vector machines and achieved satisfactory results.

N. Chandra elaborates on the social media platform and the rising popularity of various forums and communities online. Such forums and social media websites allow the users to interact and communicate with people of their same interest on the internet. But due to the increase in the affordability of the internet paradigm there have been an increasing number of people with malicious intents that gain access to this platform [7]. Such people provide racist comments and other comments that are inappropriate on the platform. Therefore, the authors in this paper have utilized a KNN algorithm to achieve efficient comment classification on social media websites and detect antisocial behavior efficiently.

A. Ikeda states that there has been an increase in the number of people that have been watching videos online every day [8]. There is also an increase in the number of people that have been joining these video sharing websites and the user base has been increasing every minute. This leads to a significant increase in the number of people who also comment on those videos. These comments are highly useful, therefore the authors in this paper have proposed an efficient technique for annotating comments on the website. These comments are highly useful for advertisement and retrieval of videos. The experimental results conclude that the proposed methodology has been achieved efficiently.

F. Prabowo introduces the concept of social media and online social platforms that have been in use by the majority of people online nowadays. Most of the social media outlets also have their own eCommerce website through which they allow people to buy and sell different kinds of products. This allows the shop owners to advertise and sell their products on the social media platform where they get a lot of customers. Majority of the customers post their reviews in the form of comments on the various products that have been posted online [9]. Such comments need to be filtered and classified using a statistical approach. Therefore, the authors in this paper have utilized CNN or convolutional neural networks to achieve their classification goals. The experimental results conclude that the proposed methodology achieves an accuracy of 84%.

M. Takeda states that there has been an increase in the number of Web Services that allow users to buy and sell different kinds of products. Such websites have various services that allow posting short reviews or comments about the product or other experiences on their platform such as Twitter etc. [10] Therefore, the authors utilized the framework of bag of words and text frequency to effectively classify the text in the comments by utilizing the hierarchy and tree kernels from tree structures the proposed methodology has been successful and used for tourism videos.

T. Peng explains that due to the increase in the number of users online there have been a significant increase in various activities performed by people with nefarious intents. Many attackers have employed a lot of different approaches to gain unauthorized access to various websites and social media accounts [11]. This has been done through performing phishing attacks which are the least defended and the most common attacks online nowadays. Therefore, the authors have tried to ameliorate this effect by utilizing machine learning paradigm in addition with natural language processing to detect phishing attacks with very high accuracy.

H. Shen Elaborates on the various different methods that have been utilized to secure the experience of an individual online. Most of the people that are being introduced to the internet recently have been the victims of various attacks that have been performed by people with malicious intentions online. This also concludes that there is a need for more storage and fast theory facilities on various log generation and Management systems [12]. Therefore, authors in this paper have utilized Natural Language Processing for the purpose of log layering and increasing the capacity of storage and easy management of the logs. The experimental results conclude that the proposed methodology increases the compression performance significantly and by a large margin.

K. Sintoris explains that there has been a significant increase in the necessity for a Natural Language Processing technique that can help understand the underlying meaning of a sentence spoken by a human being. This is especially true in the case of extraction of Business Process Models. NLP or Natural Language Processing can efficiently analyze the grammar and the syntactic structure of a sentence [13]. Therefore, the authors have utilized this feature to implement the extraction of the patterns, tasks, resources and activities from Business Process Models. The experimental results indicate that the proposed technique called BPMN has produced promising results.

III. PROPOSED METHODOLOGY

Supervised learning approach

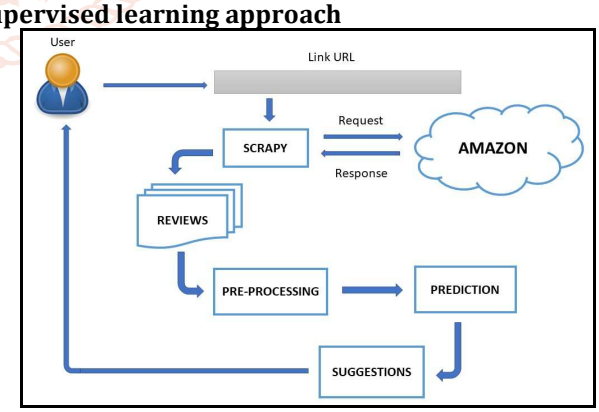

Figure 1: User-system interaction.

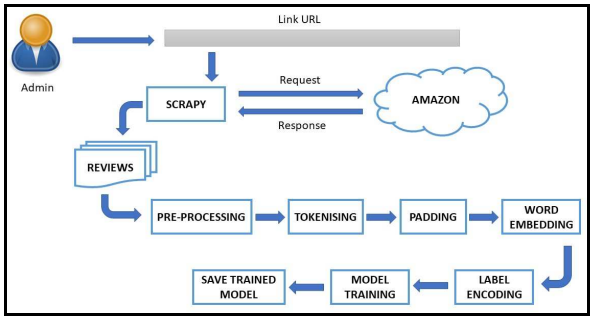

Figure 2: Admin-system interaction

The proposed model for supervised comment classification is depicted in the above figure. And the steps that involve the comment classification process is deeply narrated in the below mentioned steps.

Step 1: Data collection and Data feeding-
The proposed model for comment classification collects the live user reviews from the amazon e-commerce site using web scraping. It is done using a framework called Scrapy in python which has a construct of spiders that crawls into websites and collects required data.

Step 2: Text Tokenizing and Padding-
The text obtained from the web scraper cannot be directly fed into the neural network, however it needs to be prepared using Keras processing tools. Tokenization is a process which includes splitting a complete sentence into individual words and assigning each word a unique number.

Tokenization helps to break down complexity for the embedding layer, however the problem of non-uniformity still exists due to the difference in length of each review, this is where padding comes to the rescue. Padding is a process of appending extra bits after or before a sentence so as to make all the reviews uniform in length. The padding length is determined by the size of the largest comment.

Step 3: Word Embedding-
Word embeddings are vector representations of a word. They are used in order to find out semantic and syntactic similarities within words in a document. Words with similar meanings have similar word embeddings. These word embeddings increase the dimensionality of input data which helps in building artificial neural network models.

Step 4: Label Encoding-
For labelling the dataset, the categorical star ratings obtained from the website are in the form of text, which cannot be accepted by most keras's loss functions. This text data has to be converted into integer values via label encoding or one-hot encoding. Label encoding helps with this conversion.

Step 5: Model Architectures used-

1. LSTM:-
Long short-term memory is a modification of classic Recurrent Neural Networks (RNN). LSTM has the ability to remember inputs using memory cells for short term depending upon requirement. When LSTM was used to classify comments, the accuracy obtained was 62.4%.

2. GRU:-
Gated recurrent networks works in a similar manner like LSTM but provides better efficiency and lesser number of gates for information flow control. When GRU was used to classify comments, the accuracy obtained was 66.9%.

3. C-LSTM:-
Convolution Neural Networks mainly used for images, can here be used for converting the two dimensional input and reducing it using convolution and max pooling and then training a LSTM model on this. When C-LSTM was used to classify comments, the accuracy obtained was 66.2%.

Unsupervised learning approach

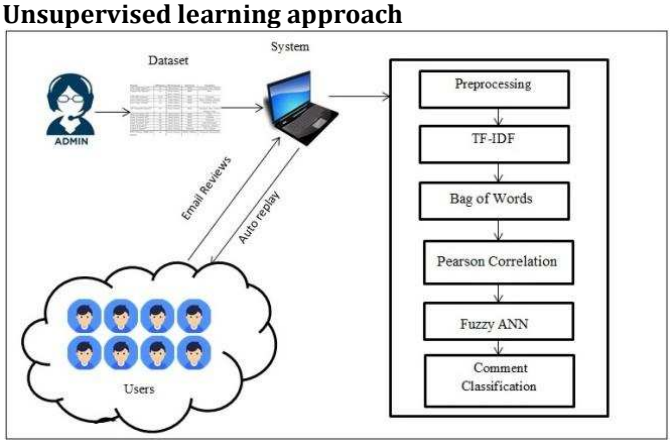

Figure 3: Comment Classification System Overview

The proposed model for unsupervised comment classification is depicted in the above figure 3. And the steps that involve the comment classification process is deeply narrated in the below mentioned steps.

Step 1: Data collection and Data feeding-
The proposed model for comment classification collects the live user reviews from the amazon e-commerce site using jsoup library present in java , used to request html and parse it to extract comments or reviews.

Step 2: Preprocessing –
This is the initial logical step of the proposed model, where the input user comments are fed to the preprocessing module. Basically preprocessing removes the burden of the unwanted textual data from the comments, so that comments become lightweight so that the time and space complexity can be maintained efficiently. This process consists of the following steps.

Special Symbol removal- The input comment text is subject to remove the presence of any special characters like!,@,,,., etc. So that the comment can be efficiently managed for the further process of classification.

String tokenization – The input comment text is divided into words using the split function of Java to store them in a list.

Stop Word Removal- All language texts are filled with the conjunction words that are mainly acting as the connecting words between the major words. On removal of these conjunction words the meaning of the phrase remains unchanged. Thereby this process makes the text more lightweight, which in turn decreases the complexity of the further process. This Stopword removal includes many words like for, and, or, to, from, is etc.

Stemming- After Stopword removal from the comment text, then many words contain the postfix phrases like ing, ion etc On replacing of these phrases with the specific required words decreases the redundancy of the given text. For example "Going " becomes " Go " , " Studied" becomes Study.

Step 3: Bag of Words: This is another entity of machine learning along with the TF-IDF that is being used in the proposed model. The bag of words model tends to evaluate the positive and negative words in the preprocessed comment. This process is conducted based on the stored

positive and negative phrases and the well narrated protocols. Each and every word of the user comment is being measured for the specific bag of the positive and negative words. And then these words are estimated by the pseudo protocol of the positive and negative words to prepare the list of positive and negative bags of words from the respective comments.

Step 4: TF-IDF: This is one of the major parts of the proposed model where sarcasm in the text is handled using the TF-IDF model. This TF-IDF model ensures the importance of the words in the text by evaluating the Term frequency and inverse document frequency. This can be represented by the following equation 1.

$$TF-IDF = TF\ X\ Log\frac{Number\ of\ Documents}{Number\ of\ Documents\ Containing\ Word\ W}$$

Initially, frequency of each of the words in a given comment is estimated and is called as Term frequency. The Obtained Term Frequency of that word is subjected to the scalar product with the logarithmic ratio of Number of Documents to the Number of documents containing specific word W.

Based on this TF-IDF the words that are really having the higher values are considered to play an important role in the sarcasm pattern forming. So these words are segregated for the future use of the proposed model.

Step 5: Pearson Correlation: The estimated positive and negative word list is optimized using the TF-IDF for each of the input comments. Then these two lists are fed to the Pearson correlation estimation.

The Pearson correlation is the technique to estimate the correlation in between the two entities. Now the proposed model uses the positive and negative word list that belongs to a comment to estimate the correlation between these two lists. The Pearson correlation can be measured using the below mentioned equation.

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(x^2 - \frac{\sum x^2}{n})}\sqrt{(y^2 - \frac{\sum y^2}{n})}} \qquad \text{_____(2)}$$

Where
x is the list of Refactored positive array
y is the list of Refactored negative array
n is the array Size
r= correlation value in between -1 to +1.

The number of positive and negative words is then optimized and refactored to the positive and negative word list to call as X list and Y list. These lists are then subjected to estimate the Correlation value using Pearson correlation equation. This Equation yields the values in between -1 and 1, which indicates the positiveness or negativeness of the comments, which is classified in the next process of Fuzzy ANN classification model.

Step 6: Comment Classification through Fuzzy ANN- This is the step where identified correlation values for each of the comments are listed in a double dimension list with two columns like comment text and comment score. This list is subjected to the Fuzzy Classification process, where maximum and minimum comment score is identified to get the difference between these two. This Difference is then divided by 5 to get the Fuzzy classification label value. As the Fuzzy Classification consists of 5 crisp values like VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH.

Based on these classification rules each and every comment scores neurons and the classified comments and then labeled as WORST, DISAPPOINTED, AVERAGE, SATISFACTORY and EXCELLENT.

The whole process is depicted in algorithm 1.

ALGORITHM 1: Comment Classification

//Input: Comment Score List C_{SL}
//Output: Comment Classified list CC_L
1: Start
2: Set min=0.5, max=0.5
3: **For** i=0 to size of C_{SL}
4: $TMP_{LST} = C_{SLi}$ [TMP_{LST} = Temporary Set]
5: $C_{SCRE} = TMP_{LST[1]}$ [Comment Score]
6: **IF** (C_{SCRE} < min)
7: min= C_{SCRE}
8: **IF** (C_{SCRE} >max)
9: max= C_{SCRE}
10: **End For**
11: RANGE1=0 , RANGE2=0 ,R_{LIST}= Ø [Rule List]
12: DF=(max-min)/5 [DF= Diffrence Distance]
13: **For** i=0 to 5
14: RANGE1=min
15: RANGE2=RANGE1+DF
16: min=RANGE2
17: $T_{LST[0]}$= RANGE1 [T_{LST}= Temporary List]
18: $T_{LST[1]}$= RANGE2
19: $R_{LIST} = R_{LIST} + T_{LST}$
20: **End For**
21: **For** i=0 to Size of R_{LIST}
22: SC_{LST} [Single Classified List]
23: $TMP_{LST} = R_{LIST}$ [TMP_{LST} = Temporary Set]
24: R1= $TMP_{LST[0]}$, R2= $TMP_{LST[1]}$
25: **For** j=0 to size of C_{SL}
26: $T_{LST} = C_{SLi}$ [T_{LST} = Temporary Set]
27: $C_{SCRE} = TMP_{LST[1]}$ [Comment Score]
28: **IF**(C_{SCRE} >= R1 AND C_{SCRE} <=R2)
29: $SC_{LST} = SC_{LST} + TMP_{LST[0]}$
30: **End IF**
31: **End For**
32: $CC_L = CC_L + SC_{LST}$
33: **END For**
34: return CCL
35: **Stop**

IV. RESULTS AND DISCUSSIONS

The comparative study between supervised and unsupervised algorithms for the purpose of comment classification suggests that, when the data is extracted from a webpage, it is purely live and random data, that has been put there by random people. The star ratings given to the product may sometimes not match the comment posted by the same user, for eg. If someone posts a comment saying 'I loved the product' and gives a rating of 3 or 2 stars, this would label the comment accordingly but loving a product would be considered somewhat as an average rating, which ultimately is not true.

Also, a supervised approach fails to perform when the comments posted are sarcastic.

The unsupervised methodology on the other hand is seen to give a better performance when working with sarcastic comments and unfair star ratings.

The proposed methodology of this paper for efficient comment classification has been developed on NetBeans IDE in the Java programming language. For the development process, a computing machine is required with a configuration consisting of at least Intel i5 processor handling the processing requirements with 4GB of primary memory and 500 GB of Storage. MySQL database server fulfilled the database responsibilities.

For determining the performance metrics of the proposed methodology, extensive experimentation was executed and analyzed through the use of Precision and Recall.

Performance Evaluation based on Precision and Recall

Precision and Recall are highly insightful and appropriate parameters that are utilized to analyze the performance of the methodology. Precision calculates the respective accuracy of the process by evaluating the precise values of the extent of the accuracy of the system.

Precision in this experimental evaluation is being outlined as the ratio of the number of accurate classifications performed and the combined sum of all the comments that have been extracted. Therefore, the parameters in precision allow for an in-depth extraction of the effectiveness of the technique in its entirety.

The Recall parameters are complementary to the precision parameters and instead evaluate the absolute accuracy of the proposed technique.

The recall is therefore evaluated by the evaluation of the ratio of the number of accurate classifications for the given comments extracted to the total number of inaccurate classifications for the given comments extracted. This confirms that the recall evaluates the absolute accuracy of the methodology.

Precision and recall are mathematically elaborated in the equations detailed below.

Precision can be concisely explained as below

- A = True positives
- B= False positives
- C = False negative

So, precision can be defined as

Precision = (A / (A+ B)) *100
Recall = (A / (A+ C)) *100

The above equations are further used for conducting extensive experimentation on the proposed methodology through the analysis of the fuzzy classification results. The analysis results are detailed in table 1, given below.

No of Comments Extracted	Accurate Classifications (A)	Inaccurate Classifications (B)	Accurate classifications not done (C)	Precision	Recall
10	8	1	1	88.88888889	88.88888889
20	17	1	2	94.44444444	89.47368421
30	28	2	1	93.33333333	96.55172414
40	35	2	3	94.59459459	92.10526316
50	42	3	5	93.33333333	89.36170213
60	50	6	4	89.28571429	92.59259259
70	57	5	8	91.93548387	87.69230769
80	79	7	6	91.86046512	92.94117647
90	74	9	7	89.15662651	91.35802469
100	81	10	9	89.01098901	90

Table 1: Precision and Recall Measurement Table for the Fuzzy classification analysis

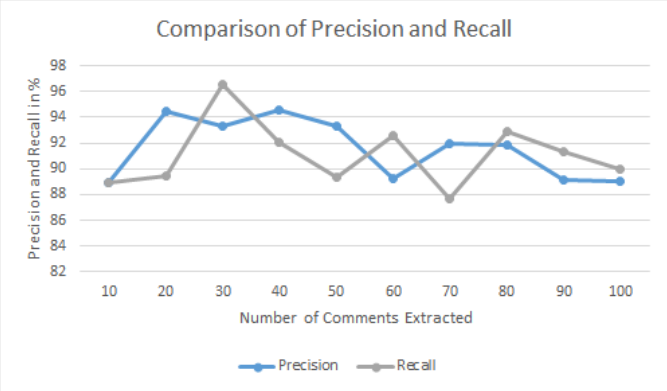

Figure 4: Comparison of Precision and Recall for the Fuzzy classification analysis

The above graph indicates that the Fuzzy classification technique utilized in the proposed methodology achieves expected levels of precision of about 91.58% and recall of 91.09%. The high values of precision and recall indicate an incredibly fair execution of the Fuzzy Classification module in the proposed methodology.

V. CONCLUSION AND FUTURE SCOPE

The classification of comments on an online forum such as social media websites and E-commerce portals has always been a highly difficult procedure. This is mainly because the English language like any other human language has complex nuances and intricacies that are very difficult for a computer to comprehend such as sarcasm which is rarely detected by humans. There is an increased necessity for performing such classifications and assigning labels to the comments as it allows for effective segregation that can be further utilized by machine learning systems to provide valuable insight. Therefore, the proposed methodology implements the NLP or Natural Language Processing techniques that can extract the underlying meaning of the sentence along with the introduction of the Term Frequency – Inverse Document Frequency and assisted by Fuzzy Classification to accurately classify the Comments.

The performance of the proposed methodology is ascertained through the use of extensive experimentation. The results of the experiment indicate that the proposed Fuzzy Classification framework has a comparable performance that is significantly better than the conventional approaches. Precision and Recall parameters were utilized to evaluate the performance metrics of the described comment classification procedure. The experimental results have demonstrated the superior accuracy of the proposed methodology.

For future work, the presented technique can be implemented in a real-time web application. The technique could also be extended for implementation as an API.

REFERENCES

[1] Hui Yang et al, "Identification and Classification of Requirements from App User Reviews", IEEE 19th Conference on Business Informatics, 2018.

[2] Chunting Zhou et al, "A C-LSTM Neural Network for Text Classification",17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019.

[3] W. Jitsakul et al, "Enhancing Comment Feedback Classification using Text Classifiers with Word Centrality Measures", 2nd International Conference on Information Technology (INCIT), 2017.

[4] M. Andriansyah et al, "Cyberbullying Comment Classification on Indonesian Selebgram Using Support Vector Machine Method", Second International Conference on Informatics and Computing (ICIC), 2017.

[5] J. Savigny and A. Purwarianti, "Emotion Classification on Youtube Comments using Word Embedding", International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017.

[6] Siswanto et al, "Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes", International Conference on Applied Information Technology and Innovation, 2018.

[7] N. Chandra et al, "Anti-Social Comment Classification based on kNN Algorithm", International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2017.

[8] A. Ikeda et al, "Classification of Comments on Nico Nico Douga for Annotation Based on Referred Contents", 18th International Conference on Network-Based Information Systems, 2015.

[9] F. Prabowo and A. Purwarianti, "Instagram Online Shop's Comment Classification using Statistical Approach", 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017.

[10] M. Takeda et al, "Classification of Comments by Tree Kernels Using the Hierarchy of Wikipedia for Tree Structures", 5th IIAI International Congress on Advanced Applied Informatics, 2016.

[11] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning", 12th IEEE International Conference on Semantic Computing, 2018.

[12] H. Shen et al, "Log Layering Based on Natural Language Processing", International Conference on Advanced Communications Technology (ICACT), 2019.

[13] K. Sintoris and K. Vergidis, "Extracting Business Process Models using Natural Language Processing (NLP) Techniques", IEEE 19th Conference on Business Informatics, 2017.