International Information and Engineering Technology Association

Advancing the World of Information and Engineering

Modified Fuzzy Approach to Automatic Classification of Cyber Hate Speech from the Online Social Networks (OSN's)

Ashok Kumar Nanduri[1*], G.L. Sravanthi[1], K.V.K.V.L. Pavan Kumar[2], Sadhu Ratna Babu[3], K.V.S.S. Rama Krishna[1]

[1] Department of Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, Guntur 522009, Andhra Pradesh, India
[2] Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur 522502, Andhra Pradesh, India
[3] Department of Information Technology, Bapatla Engineering College, GBC Rd, Mahatmajipuram, Bapatla 522102, Andhra Pradesh, India

Corresponding Author Email: jashok.nanduri@gmail.com

ABSTRACT

The extensive use of online media and sharing of data has given considerable benefits to humankind. Sentimental analysis has become the most dynamic and famous application area in current days, which is mainly used in knowing the public's opinion. Most algorithms of machine learning are used as principle methods for sentimental analysis. Even though several methods are available for classification and reviews, all of them belong to a single class of classification which differs among several different classes. No methods are available for the classifying of multi-class instances. Therefore, fuzzy methods are used for classifying the instances depended on multi-class for achieving a clear-cut view by indicating suitable labels to objects during the classification of text. This paper includes the categorization of cyberhate information. If there is a growth in dislike speeches of the online social network may lead to a worse impact amongst social activities, which causes tensions among communication and regional. So, there is the most demand for cyberhate conversation detection automatically through online social media. Generally, an updated process of fuzzy words is designed that includes two stages of training for the classification of cyberhate conversation into 4 forms, race, disability, sexual orientation, and religion. Depended on the types of classification, experiments have been conducted on these four forms by gathering different conversations through online media. Systems based on rules of fuzzy approach have been used. This fuzzy with rule-based is for the classification of features using Machine Learning techniques such as the words that implants for future bag-of-words and extraction methods. In this, the cyberhate conversations are taken from OSN's depended on the attributes defined in a dataset using rule-based fuzzy.

1. INTRODUCTION

In the present day's the motto of sentiment analysis is finding the opinions or views of individuals using the linguistics, analysis of the text, and tongue processing [1]. As we all know that machine learning (ML) approach has become a really powerful tool for sentiment analysis [2] classification by suing various algorithms like support vector machines (SVM), naive Bayes (NB) and a lot more are used with good performance in a wide range of applications that involve sentiment analysis. Some of the sentiment analysis applications are Cyberbullying detection [3], abusive language detection [4], reviewing the movies, online e-commerce site reviews [5], and cyberhate identification [6]. In recent years, neural networks or DNN (deep neural networks) are also utilized for performing sentimental analysis, different sorts of the classification of text. The machine learning concept uses the algorithms (NB, GBT, DT, DNN, and SVM) which are mentioned above are considered as discriminative learning, as all aims to differ among classes. Moreover, the algorithms which are described above are work supported for the assumption that various classes are exclusive in mutual and

every instance is very clear, given a label of truth at ground level. Therefore, within the concept of classification of text, the assumptions mentioned above will not behold every time, specifically while involving the examples subsequently [7].
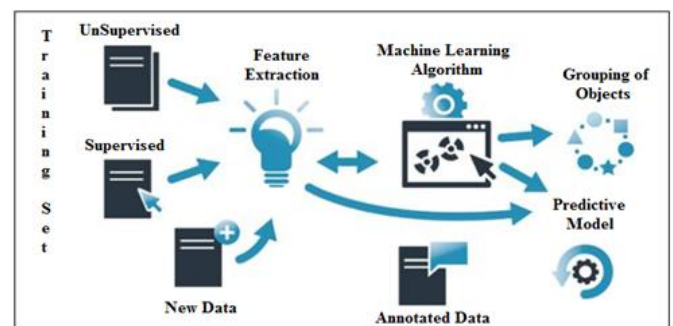

Figure 1. Denote the flow of machine learning algorithm on a sample training dataset

From the above Figure 1, we can clearly identify that in order to apply machine learning algorithms on any type of

training dataset, the data will be initially loaded and main features are extracted from that input data. Once the main features are extracted then they will be sent to the predictive model with the help of appropriate machine learning algorithm. Here we try to group the objects based on two categories like similar category and dis-similar category. Based on the type of prediction the resultant data is extracted and all the main fields are annotated by the system. In this way we can able to use the appropriate machine learning algorithm to perform the given task in an efficient [8].

Here the text classifier of SVM is trained, tested to notice the cyberhate or online hate message(speech) through the information gathered from OSN's. In this, we deal with the (Supervised Machine Learning) SVM tools that are used for the categorization of sample "data set" and how their outcomes are interpreted that is used in decision making and policy. ML proposal BoW (Bag-of-words) as characteristics. A bag-of-words proposal uses words inside a principle as predictive characteristics and neglects sequence among words and also semantic and syntactic content any. For modeling offensive data, a rule set is also produced, showing refinement on ML standard approaches regarding a more decreased false rate of negativity.

According to the previous research, after collecting the data from OSN's, an SML classifier is built to differentiate antagonistic or/and hateful feedback having a focus on disability, race, sexual orientation, and religion, and more common feedbacks based on the event. For finishing the subjective action through data analytics on a large-scale, which is completely needy for volumes of produced data, ML classifiers we used for learning the tweets characteristics which include their class (general response or cyberhate). After learning about characteristics, the model is applied to the entire data set. Later it works as a set of classes or categories in which every information point will be classified. The outcomes of annotation actions will split into testing and training sets of data for ML.

Java Weka ML libraries were used for getting experimental outcomes to enlarge more supervised classifiers, which are tested and trained the characteristics discussed in the earlier section. Every tweet was converted as a characteristic vector-an attribute set that represents tweets for use in classifier training. Every vector contains the original class the tweet related depended on the exercises of human annotation, and a programs list that may contain typed dependencies, words, or both of them, based on the characteristics of the set used for classifier training.

Depended on these principal characteristics are taken out and sent to predictive models with the help of suitable ML algorithms. In this, the objects are grouped depended on 2 categories such as dis-similar and similar category. The resultant information is taken depending on the prediction type, and the system notices overall main fields.

For getting promising outcomes among the purposes of rule-based are used in the classification of antagonistic information. For determining the improvement or enlargement on a rule-based or probabilistic model of classification of spatial model an SVM was also utilized.

In general, if we come with the first assumption, one movie belongs to several categories or sometimes the one book can be similar to several subjects [9]. Similarly, there are lots of examples relatively not required for exclusive mutually, that is two or more classes will contain overlaps, regarding the instances of a class, could have overlapped, in terms of class

instances but not in their real working functionality. In the second case, while several classes will be mutually positive exclusive, their instances can be most complex, maybe very complex task to classify them into one category. For example, a small piece of a sentence like— I LOVE India but I HATE immigrants‖ contains speech of both negative and positive [10]. From the sentence we can identify that sentence may not be clear-cut because this is contained with half positive meaning and half negative meaning. Hence this is not clear- cut message because it is having some challenges inside the sentence.

Additionally, in the domain of sentiment analysis, every instance of an individual contains a label may not be the inner true or positive nature of that message but may represent an acceptance as the view of several people. This is nothing but different people will have different opinions about the sentiment instance. Due to this reason, sentiment analysis is becoming an essential factor to discover the pattern from user opinions [11]. All these examples mentioned above describes that the instances based on the text are fuzzed normally and the learning methods are similar to compute struggle like fuzziness. This mainly motivates me in developing the methods of fuzzy to classify the concept and give a specific result for the uncertain data.

In this current thesis, we try to concentrate on the hate speech detection in online (cyberhate) in short we term as posted text as informal on the platforms of social media. This is the most important research work which is going on to identify the online hate speech to reduce the outcomes of anti-social. In general, we try to design an updated approach of fuzzy which contains learning generatively for classifying the cyberhate messages into four types like, sexual orientation, race, religion, and disability. In this thesis, we use a fuzzy-based approach for getting the clear-cut result by verifying the sentence in two-class classifications.

2. BACKGROUND WORK

In this section, we try to define the background work that is carried out in order to propose this current application.

2.1 Preliminary information

Here we try to analyze the main factors that are required for processing the text. Text Processing can be defined as an important task in major machine learning applications. Now let us look about this and its applications in detail:

Some of the applications which come under text processing are:

(1) Language translation

This is one form of text processing technique used for translating a sentence from one language to another language by using some predefined meanings of these words.

(2) Sentiment analysis

This is another form of text processing technique which is mainly used to identify the sentiment by a text of corpus; either the sentiment towards any concept is positive, negative, or neutral.

(3) Spam filtering

This is also one of the text processing techniques used for identifying the unsolicited and unwanted Email or SMS messages which come from a in distinct locations.

Figure 2. Denote the flow of text processing in machine learning approach

In this application, we try to use this text processing technique on sentiment analysis to analyze the classification of cyberhate conversations, as the increasing of the hate speech through social media may lead to wrong effects on social actions, which may also cause tensions in regional, community.

From the above Figure 2, we try to gather the information from a document and try to extract the features from that document and then convert those extracted features in machine understand format. Now we try to apply sentiment analysis technique to identify the hate speech based on any of the four types like:

Religion, Race, Disability, and Sexual Orientation.

Now the extracted features are mapped with these four types and if any word is matched with these category words then the conversation is identified as hate speech on social media.

Initially the data pre-processing step need to be performed on the input data. Here the input data is nothing but a sample web site containing several products and opinions collected from e- customers. The pre-processing step can be split into two main phases.

In order to deal with text processing we need a huge amount of text to perform classification and also require a lot of work from the back end. Now let us discuss about this text processing in a step wise manner:

(4) Data pre-processing

In this stage we are going to undergo these steps

Tokenization — here the input sentence is converted into words and all these words are combined at the end to form a data set.

Try to remove the unnecessary punctuation, special symbols and tags.

Next we try to check if there is any duplicate words present in these words and try to remove those frequently repeated words like ‖the‖, ‖is‖, etc. They are removed because they don't have a special meaning or importance.

Stemming — In this stage we try to construct a root node and from that root node we try to create stems and this will be continued until leaf words are identified.

Lemmatization — In this stage we try to remove the inflection of the sentence by determining the part of speech.

For example:

The stemmed form of studies is: studi.

The stemmed form of studying is: study.

The lemmatized form of studies is: study.

The lemmatized form of studying is: study.

Finally we came to a clear understanding that stemming & lemmatization [12] greatly help to reduce individual words like _studies', _studying' to a common root word like _study'.

This will greatly reduce the duplicate words which are same in meaning but different in usage.

3. OVER VIEW OF MODIFIED FUZZY APPROACH

In this section we mainly try to demonstrate the overview of modified fuzzy approach to automatic classification of cyber hate speech from the online social networks. Now let us discuss about that in detail as follows:

Motivation:

Fuzzy logic is one kind of rule-based system which is termed as the extension of deterministic logic. In this fuzzy logic, we try to employ the truth values with two numerical values like zero to one (i.e. 0 to 1), they are termed as truth values involved in identifying the truth based on rule-based systems. According to the concept of rule-based systems of fuzzy, the primary operation is to use continuous numerical values in the training stage mapped with the corresponding attributes. This is mainly formed by identifying the total number of attributes that are present in that sentence and then map every attribute with individual identity.

For example, if we take —Age‖ as one attribute in the sentence, this may have numerical values from 0 to greater than 100 and using this fuzzy-based approach we try to classify this age attribute into 3 attributes of qualitative such as — Young Age‖, — Middle Age‖ and — Old Age‖. Here every individual attribute is considered as a set of fuzzy attributes and they must be ranging between [0, 1].

Earlier frameworks mainly aim at distinct abstraction levels for detecting hate speech and decreasing the single task complexity in a single level. Ample works have been done for detecting hate speech generally as a task.

Some works have been concentrated on noticing the absence or presence of multiple forms as a task for hate speech. Hate speech of distinct forms contains potential intersectionality, which specifies that identifying tasks of hate speech will not be formulated easily as regular multi-class categorization issues depended on the mutual exclusion assumptions of distinct classes. Especially, the classification of multi-label will be a better suitable formulation. We achieve the classification of multi-label typically using the training classifiers at a supervised learning setting. However, multi-labeling is experimentally less feasible, for hate speech information, as emphasized. Hence, a new concept for formulating fuzzy depended on purpose to cyberhate speech automatic classification on ONS's. It includes two stages of training for categorizing the cyberhate conversations (messages) in 4 forms such as religion, sexual orientation,

disability, and race. The experiments are conducted on these 4 forms by gathering distinct messages from social media from online and attempt to categorize these characteristics through the ML approach such as words which includes for future extraction methods. Therefore, the fuzzy rule-based is noticing the cyberhate conversations through OSN's depended on the attributes defined in a dataset.

3.1 Stages of feature extraction

The following are the stages of feature extraction, they are as follows:

(1) Identify the number of keywords present in the Sentence:

As we all know that keywords are most powerful words which give brief idea about the text or sentence. Initially from a sentence or document we need to extract the count of keywords and also the distinct number of keywords that are present in the text corpus.

We can extract the count of keywords from the text and also the keywords [13]. This feature could be helpful in many ways, one example — if you are classifying the tweets with and without keywords or want to classify on the types of the keywords. In this proposed thesis we try to classify the online social networks tweets or posts with the cyber hate speech words, so we try to identify every tweet posted by corresponding OSN user and try to identify the important keywords from that tweets.

(2) Identify the number of users who are Tagged in that Tweet or Post:

If we find a post or tweet is tagged with a set of users, then we can identify that tweet is a conversation between two or multiple parties for some kind of acknowledgement. Here for any tweet the users are tagged with a symbol @ in the tweets. In this stage we try to create a new feature which will identify and hold the words which are marked with @ symbol.

(3) Identify the Count of Numerical Values Present in that Conversation

Normally in a conversation we can identify a lot of Numeric values with different meanings. These numerical values can be sometimes a year and sometimes money. So we need to identify all the numerical values which are present in that text corpus and identify as one feature in the method of feature extraction.

(4) Identify the Count of Upper Case Words in that Conversation

In general we try to use UPPERCASE words as an abbreviation or sometimes used to express excitement like anger, happy, OMG, TGIF, TTYL etc. With these types of words we can able to classify the emotion that is present in that conversation sentence. So we need to identify all these type of upper case words and try to store them as one feature in the method of feature extraction.

(5) Data Cleaning

At this stage we will try to clean the unwanted fields which are present in the conversation like spaces, symbols and try to keep remaining features for the process of text classification.

(6) Identify the Average Word Length of that Conversation

Here at this stage we try to identify the different kind of words that usually have different word length and all may be heterogeneous in nature. So average word length will greatly help the users to differentiate the text type and situation.

Average word length = Sum (Length of all the Words in the tweet or document) / (Total Number of Words in the Tweet (or) Document).

(7) Calculate the Sentiment Score:

In this stage we try to calculate the popular words which are having more intensity are identified and they will be marked as main features. Here we try to find out the rank of such words based on the sum function.

3.2 H.Bag of Words (BoW) calculation

This BoW is nothing finding the hate speech related words and then add all these words in the database and they are termed as the word's frequency in a document. Here in the proposed application these bow need to be identified and maintained by the administrator and the decision of choosing a word in this bow database purely depends on the admin [14].
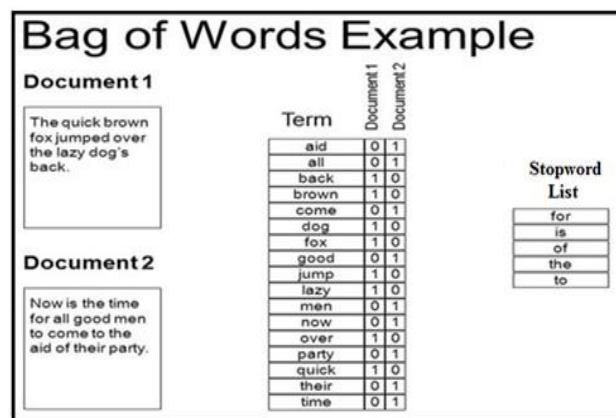

Figure 3. The flow of bag of words stage in the process of feature extraction

From the above Figure 3, we can clearly identify that bag of words is extracted from two individual document and they are arranged in an alphabetical order. They are mapped with terms [0,1]. If the document1 contains those words, then it is marked with value _1 and if the same word is not available in that appropriate document, then it is termed with value '0' [15-17].

Here in BOW for accessing/taking the words in order, that we can use predefined functions. This application is implemented fuzzy-based java application, so we can use an array.sort() method for displaying the words in order. Based on this BOW will classify the model according to user requirement and return a required result.

Fuzzy logic plays a very essential role in real time applications. The Boolean system concept will be replaced by it. In between false and true several values are provided by fuzzy, and to get a better solution to a particular problem it provides flexibility. Only 2 possibilities (1 and 0) are available in a Boolean system, here the exact false value is represented by 0 and the exact true value is represented by 1, which is true and false partially. These values are used to determine the state of cyberhate tweets is a normal or positive tweet. Depended on the 4 fuzzy functional building blocks, all this is done. They are Defuzzification, Rule Base, Inference Engine, Fuzzification. The Rule Base methodology includes rules set and if-then conditions given by the specialists to rule the system of decision making, based on linguistic data, a model of Fuzzification transforms the inputs that are crisp number as a sets of fuzzy. The accurate inputs are known as crisp inputs which are measured using a classifier and sent to processing, the matching ratio of existing fuzzy write to every rule is determined by Inference Engine and determines the rules that to be removed based on the fields of input, then, finally, the

model of Defuzzification transforms the sets of fuzzy acquired through inference engine as crisp value.

4. IMPLEMENTATION STAGE

Implementation is a stage where theoretical design is automatically converted into programmatically manner. In this stage we need to divide the application into number of modules and where each and every individual module has its own methodology. The proposed application is designed in JAVA programming language with HTML, CSS and JSP as front end technologies and MY-SQL as back end database to store the information about the tweets as well as the Bag of Words. The proposed application is mainly divided into two main modules and internally there are several sub-modules are available. Now let us discuss about these modules in detail as follows:

1). Admin/Server Module.
2). User Module

4.1 Admin/server module

In this module, the admin/server has to login with their valid username and password. After login successful they can do some operations such as View All Users, Add Filter, Add Post, View All Posts, View All Reviewed Users, View All Text Classification Details, View All Cyber hate Details, View All Positive Details, View All Users Search History, View All Search Ratio, View Search Ratio Result, View Post Score Results, View Keyword Score Results.

4.2 User module

According to this module, n numbers of people are available. Registration should be done before the user for performing any action. After successful registration, the user can log in to his account by using the valid user id/name and password. If the login is done successfully the user can perform few operations such as Search Post, Search Reviewed Posts, My Search History, My Reviewed Posts. According to this module, the user can able to search the data sets type and click on search then he will show all details of a particular data set.

In the user module, we can also add sub-modules like recent reading, related recommendations to the user. Based on these modules recent user identifies the previous user what they read and is it helpful for us or not. if it is useful then they are trying to learn more on that and Does the application give us any recommendation? Would it be useful for us to give a recommendation? Or not? Based on these we can get more information. So these two modules also play a vital role in this application.

5. EXPERIMENTAL RESULTS

We have conducted experiment on sample online social network web sites by gathering tweets or posts from various online social network users. Here we designed this Offline OSN web site by using Java Platform using JEE as development environment. We tested the application by registering multiple OSN users and try to login all users and try to send friend request with one another. Here we try to login as one user and try to add a post and then wait for the

comments or reply posted by the other users. Here we try to categorize the hate speech based on four categories like: Religion, Race, Disability, and Sexual Orientation. These categories are maintained by the server module in order to add the hate speech related words into the bow database. Here we try to use modified fuzzy approach to classify and cluster the messages which are posted on the OSN walls.

Here we look at four cyberhate messages. But if we want, we can add a lot of cyberhate messages to our application. The final result will be based on the cyberhate messages we have added into an application and the cyberhate messages that have come from an online social network (OSN's). Based on the model finally, we can find out which one is a cyberhate message and which one is a normal message (i.e. negative and positive). This is done purely based fuzzy logic approach based on the fuzzy model. we can apply the sentiment analysis technique to identify hate speech based on any of the other types like color, ethnicity, gender, nationality, etc.

5.1 Admin try to add filtered words into the BoW database

Figure 4. Admin try to add filtered words into the BoW database

From the above Figure 4, we can clearly identify the filter is mainly categorized into 6 categories like Religion, Race, Disability, Sexual, Cyberhate, Positive. Based on these categories admin will try to choose a set of cyberhate speech related words and add those words into the bow database. Once if any words are added into the database the same word cannot be added either in same category or different category, this is a unique constraint maintained by the admin or server.

5.2 Admin can view all the cyber hate filtered words from the BoW database

Figure 5. Admin can view all the cyber hate filtered words from the BoW database

From the above Figure 5, we can clearly identify there are a set of words which are added and maintained by the admin, where each and every word is repeated only once and there are no duplicate values present in that BoW database.

5.3 Comparative analysis of cyberhate tweets and normal tweets based on sample tweets
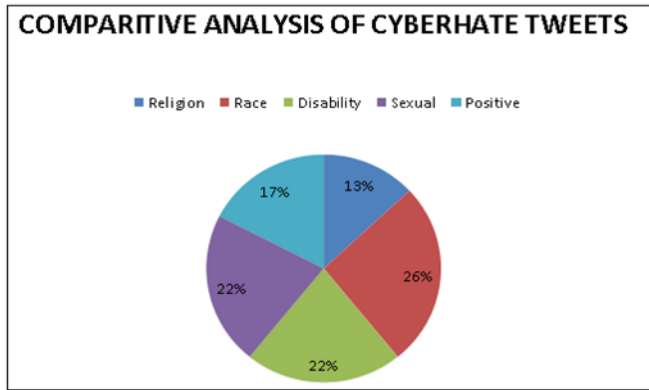

Figure 6. Comparative analysis of cyberhate tweets and normal tweets based on sample tweets

From the above Figure 6 we can able to find out all the distinct categories with the percentage of each and every individual category. Here we collected more than 100 of sample tweets posted by various users on several type of posts and from those tweets we try to classify the type of tweet based on the bow database. Here we can see Race category occupied nearly 26 percent of words matched in the user tweets. In this way automatic classification of tweets is possible from online social media and we can able to find out the percentage of hated users who suffer with these symptoms.

6. CONCLUSION

In this study, for the 1 time, we have created a method for classifying the multi-class instances with the exact response. For this, the fuzzy methodologies have been used for instance classification depended on multi-class for achieving exact representation by allocating suitable labels for objects at the classification of text. For this proposed work, we gathered various posts or tweets through various OSN (online social network) users after that attempted to study every tweet based on the categorization of cyberhate messages, as the growth in hate-speech from an online social network may cause negative effects on social activities, which may either lead tensions among the community and regional.

In this described application, we attempted for designing an updated fuzzy that includes 2 stages of training for the classification of cyberhate conversation into 4 forms, race, disability, sexual orientation, and religion. By performing several experiments on the present process the experimental and theoretical estimation has validated the characteristics through ML approaches such as the words that combined for future methods Bow (bag-of-words) and extraction, which also differentiate cyberhate and normal conversation very appropriately.

REFERENCES

[1] Stone, P.J., Dunphy, D.C., Smith, M.S. (1966). The general inquirer: A computer approach to content analysis. American Journal of Sociology, 73(5): 634-651. https://doi.org/10.1086/224539

[2] Gottschalk, L.A., Gleser, G.C. (1979). The Measurement of Psychological States Through the Content Analysis of Verbal Behavior. Univ of California Press.

[3] Volcani, V., Fogel, D.B. (2001). System and method for determining and controlling the impact of text. United States Patent 7603268.

[4] Das, A., Bandyopadhyay, S. (2010). Subjectivity detection using genetic algorithm. Computational Approaches to Subjectivity and Sentiment Analysis, 14-21.

[5] Hu, M., Liu, B. (2004). Mining and summarizing Customer Reviews. Proceedings of KDD 2004.

[6] Burnap, P., Williams, M.L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 5: 1-15. https://doi.org/10.1140/epjds/s13688-016-0072-6

[7] Burnap, P., Williams, M.L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2): 223-242. https://doi.org/10.1002/poi3.85

[8] Liu, H., Cocea, M. (2017). Fuzzy rule based systems for interpretable sentiment analysis. 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 129-136. https://doi.org/10.1109/ICACI.2017.7974497

[9] Liu, H., Cocea, M. (2017). Fuzzy information granulation towards interpretable sentiment analysis. Granular Computing, 2(4): 289-302. https://doi.org/10.1007/s41066-017-0043-8

[10] Nanduri, A.K., Sravanthi, G.L., Kumar, K.S., Kumar, K.P. (2020). Sentiment analysis of cascading twitter data using Hadoop ECO system approaches. Journal of Critical Reviews, 7(9): 452-460. https://doi.org/10.31838/jcr.07.09.92

[11] Jefferson, C., Liu, H., Cocea, M. (2017). Fuzzy approach for sentiment analysis. 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, pp. 1-6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015577

[12] Plisson, J., Lavrac, N., Mladenic, D. (2004). A rule based approach to word lemmatization. In Proceedings of IS, 3: 83-86.

[13] Zadeh, L.A. (2015). Fuzzy logic—a personal perspective. Fuzzy Sets and Systems, 281: 4-20. https://doi.org/10.1016/j.fss.2015.05.009

[14] Gambäck, B., Sikdar, U.K. (2017). Using convolutional neural networks to classify hate-speech. Proceedings of the First Workshop on Abusive Language Online, pp. 85-90. https://doi.org/10.18653/v1/W17-3013

[15] Ross, T.J. (2004). Fuzzy Logic with Engineering Applications. New York: Wiley. https://doi.org/10.1002/9781119994374

[16] Siva Kumar, K., Rama Krishna, K.V.S.S., Satya Sandeep, K., Ashok Kumar, N. (2020). Classification and feature selection approaches for cancer detection. International Journal of Advanced Science and Technology, 29(5): 4444-4455.

[17] Kotamraju, S.K., Nanduri, A.K., Sujatha, V. (2020). Implementation of intrusion detection system. International Journal of Scientific & Technology Research, 9(2): 1406-1410.