Fuzzy System and Genetic Algorithm for Social Media Text Classification

Asst.Prof.Dr. Hayder Mahmoud Salman

Asst.Lect.Mohammed Abduljaleel Shannen

Al-Turath University College

Abstract

Text classification is an issue for some applications. The point of text classification is to arrange text to one or many classes. Web-based media gives ample data to concentrate on individuals' thoughts and feelings about occasions in this world. The issue is to arrange texts dependent on the importance degree to the picked occasion, where the pertinence degree can be highly relevant, moderate relevant, low relevant, or irrelevant. In this paper, this issue is tackled by utilizing a classification system based on fuzzy logic and genetic algorithm. The proposed framework goes through four phases that are information assortment, preprocessing, features extraction, and a classification stage. In the information assortment stage, this framework is rely upon the Twitter text as a contextual analysis. The aftereffects of the proposed framework contrasted with and fuzzy logic-based method and Naïve Bayes classifier dependent on the adjustment rate and gradual rate. The amendment pace of proposed framework for every informational index are (98.75%, 99.45%, 99.31%, 98.70%, 98.55%) however the remedy rate fuzzy logic technique are (99.2%, 99.3%, 98.3%, 98.3%, 99.1%) and Naïve Bayes classifier are (95.7, 97.7, 98.4, 96.3, 96.7) in grouping. At the gradual rate, the proposed framework can extricate tweets more than this technique, where in dataset 1 the number of the tweets removed by the proposed framework is 160 tweets however the quantity of the tweets that separated by the Keyword search strategy, Naïve Bayes classifier and fluffy rationale based strategy are 98, 133 and 141 in a grouping.

Keywords: Data Collection, feature Extraction, and classification

1- Introduction

This research explains the design and implementation of the proposed evolving intelligent system in detail. We proposed evolving intelligent system for twitter text classification based on the fuzzy logic and the genetic algorithm. Evolving intelligent system focuses on classify text data from twitter to a designated occasion where the level of importance to wanted occasion on the off chance that it irrelevant ,relevant, low relevant, moderate relevant and high relevant, based on numbers of steps.

2- Data Collection phase

The first step of evolving intelligent system for text classification is a data collection and data division into training data and test data. Underlying data containing in excess of one million tweets collect through the Twitter API ("Application Programming Interface") collected tweets from twitter.

In social media, twitter is one of the essential tools, let user to cast their more thinner on specific issues and events by tweets that are 140 characters or less than. This is one of the challenges in this system where the tweet contains 140 characters to the maximum extent and this number contains additions and words that are not important in the classification process and if not deleted its will affect the process of classification such as numbers, special characters, URL, hashtag and words stop. So, after the preprocessing will remain a few words in every tweet to classify it.

Data gathered amid period from 10.27.2012 to 11.7.2012. Each record have of location, timestamp, date and text data. This data sifted and get just data of text and are then prepared and then processed and extract features and classify them. After data collection, set of 1000 tweets take from primer information heedlessly pursued as preparing information. Training data used to remove More than 50 words utilized oftentimes and concentrate extract fuzzy rules, which utilized in the derivation stage. Tweets are individuals' thoughts and conclusions of user so it not contain contextual data. Along these lines, physically arrange an arrangement of tweets that utilized as training data for classification procedure. Each tweet utilizing one or zero to demonstrate irrelevant or relevant. Each tweet has gathering of score that interval from 0 to 15. Four score breaks described to portray a significant degree of applicable to a tweet dependent on irrelevance L1 [zero, 5), low relevance L2 [5, 9), moderate relevance L3 [9, 12) and high relevance L4 [12, 16], individually.

For Comparison of defuzzification functions phase we utilize training data classified manually composed of 600 tweets, isolated to 300 irrelevant and 300 relevant (low, moderate and high relevance.

3- Data Preprocessing phase

Tweets contain non-helpful information during the time spent sorting text like a URL (worldwide website page address), a name, numbers, and stop words. For instance, 'Storm Sandy! #Hurricane (Bonnier) http://twittter.com'. Eliminate these increments or control them in tweets so as not to influence the arrangement interaction. It contains some inside cycles, for example, eliminate URL, eliminate exceptional person, increments tokenization, eliminate stop words, stemming, lemmatization and Part of Speech (POS). Gathered and partitioned tweets are the contributions to the preprocessing step and the result of this progression is a progression of significant words that utilized in the component extraction step. The pre-processing contains the accompanying stages:

3.1- Remove additions

Non-supportive information in tweets influence to the grouping system, for instance, URL, (worldwide site page address), a name, numbers and extraordinary person. URL represents Uniform Resource Locator, and used to indicate addresses on the World Wide Web. A URL is the key organization distinguishing proof for any asset associated with the web. For instance, "Hurricane Sandy! 2012 (Bonnier) http://twittter.com". Eliminate these increases or control them in tweets so as not to influence the order cycle.

In this review, utilized example coordinating to dispose of these added substances. For instance, a URL with a static example beginning with "http://" will be erased when it is found and erase numbers, name and Special characters as similarly.

3.2 Proposed Hashtag process

Hashtag via online media sites is a word or expression went before by a hash mark (#) utilized inside a message to recognize a catchphrase or subject of intrigue and work with a quest for it, Like #HurricaneSandy. Client sees straightforwardly, yet the machine and the program cannot recognize them. For this situation. Utilize correlation expressions of Hashtag with actually English words. E.g., can recognize "Sandy" after fifth activity amid #SandyHurricane, in light of the fact that parts of term can be described as "S", "Sa", "San " , " Sand ", " Sandy ", separately . After processing of the hashtag, we get very useful additional words that help us in the classification process

3.3 Tokenization

Tokenization is the demonstration of separating a tweet into pieces called tokens to deal with in a simple and proper manner with it. Tokens can be individual words, expresses or even entire sentences. The tokens become the contribution for another interaction. Algorithm (1) Show tokenization process.

Algorithm (1) Tokenization process

```
Input: Tweets
Output : List  of tokens
Begin:
For Tweet  do
  Split Tweet to token when read: stroke, semicolon, space, comma, solidus,., or dot.
  Convert token to tokens List
End for
Return tokens List
End
```

3.4 Remove Stop Words

The main advance in preprocessing is the erasure of the stop words, these words are the most habitually utilized in English and seldom these words valuable in the grouping system its correlative to the tweet and doesn't influence the importance when eliminate it.

3.5 Proposed Enhancement Stemming algorithm

A stemming cycle attempts to return the entered word to its source without Affix, which are both Suffix and Prefix expansions, as well as setting the language rules on each word. For instance, the words "show", "introduced", "introducing", could be in every way decreased to a predominant portrayal "present".

3.6 Lemmatization

Lemmatization is like word stemming yet it doesn't need delivering a stem of the word to supplant the postfix of a word. It showing up in free text with a (ordinarily) unique word addition to get the standardized word structure. For example, the postfixes of words working, works, worked would change to get the standardized formwork representing the infinitive: work; for this situation, both the standardized word structure and the word stem are equivalent. In some cases the standardized structure might be unique in relation to the stem of the word. For instance, the words registers, figuring, processed would stemmed to process, yet their standardized structure is the infinitive of the action word: figure.

4 Features Extraction phase

Extraction features are an important step in the classification process. Some words appear more frequently than others do. After defining the four periods L1, L2, L3 and L4 in data collection step, the tweets that belong to L2, L3 and L4 are choose from the training data collection then we process each tweet. Frequency of each word in tweets is calculated and we choose more 50 words repeated. Word's importance for every word ai define as:

$$ai = \frac{(Ai)}{(Bi)} \ldots (3-1)$$

Where Ai decides words number in tweets that belong to L2, L3 and L4. Bi represent words number in each tweets i.e. that belong to L1, L2, L3 and L4; ai is a rate that represent word's important i. next step, sort most words used frequently based on ai of smallest size to configure the D list, Then compute similarity between word in main menu D and word in tweet . Process of similarity introduce as mathematical operator. Any tweet has n words, Ti indicate to word ith in tweet and i ϵ {1… n}. CK indicate to kth word in D and k ϵ {one…50}. Highest value from similarity scores chosen to represent Fi's. So, Si score define as:

$$Si = \max(ck \; \times Ti \otimes Ck) \, , i \in [1,n] \;\; \ldots \;\; (3-2)$$

Si is our basic value. This features vector extracted from each tweet and it use through improved fuzzy logic method. Details of features as shown:

Indicates (Gj) to word 's highest score of word in tweet (jth)

$$Gj = \max Si \ldots (3-3)$$

Where Gj decides great score of word in tweet (jth).

Indicates (Kj) to tweet score

$$Kj = sum\ Si \ldots (3-4)$$

Where Kj decides a tweet, accumulate score of words

Indicates (Nj) to tweet length

$$Nj = n \ldots (3-5)$$

Where n indicates to words number in tweet.

Indicates (Mj) to regularly utilized words number in tweet

Mj shows to number of words in tweet and it same to words in list L. list L contain words utilized regularly and use to contrast and all tweets.

Indicates (Wj) to tweet weight

$$wj = \frac{Kj}{Nj} \ldots (3-6)$$

Where Wj is mean of words.

Indicates (Xj) to frequently used words weight in tweet

$$Xj = \frac{Mj}{Nj} \ldots (3-7)$$

Where Xj decides rate of words used frequently for all words in tweet.

Indicates (Vj) to no. of patterns in jth tweet (Vj)

After get a list. There are useful words found in training data more than 50 important words that come on their own but are not on list. e.g. 'not safe ' term beneficial more than just one term such as ' safe'. So, indicates YJ to number of this type of pattern in tweet.

5 Proposed classification Phase

The main step in classification is to give it decision. In classification procedure, we use evolving intelligent system based on fuzzy logic with genetic algorithm. The Fuzzy logic considered as a sort of canny frameworks. It can join some information on human specialists in a type of sensible derivation rules. The fundamental mark of fuzzy logic innovation is its capacity of recommending an inexact answer for a loosely planned issue, which traditional rationale can't offer. A genetic algorithm is an inquiry strategy displayed on the mechanics of normal choice rather than a recreated thinking process. It keeps a populace of up-and-comer answer for the main pressing issue, and causes it to advance by iteratively applying a bunch of stochastic administrators to create best outcome. The Inputs for this arrangement method are a bunch of elements separated from tweet contain eleven worth in include vector. The result of this framework is choice of arrangement for tweet, which is level of relationship for each tweet to a designated occasion where the level of importance to wanted occasion if it irrelevant; low relevant; moderate relevant or high relevant. Figure (1) show system of utilizing an arrangement technique. Grouping strategy Pass through three stages of Fuzzification, induction and Defuzzification>

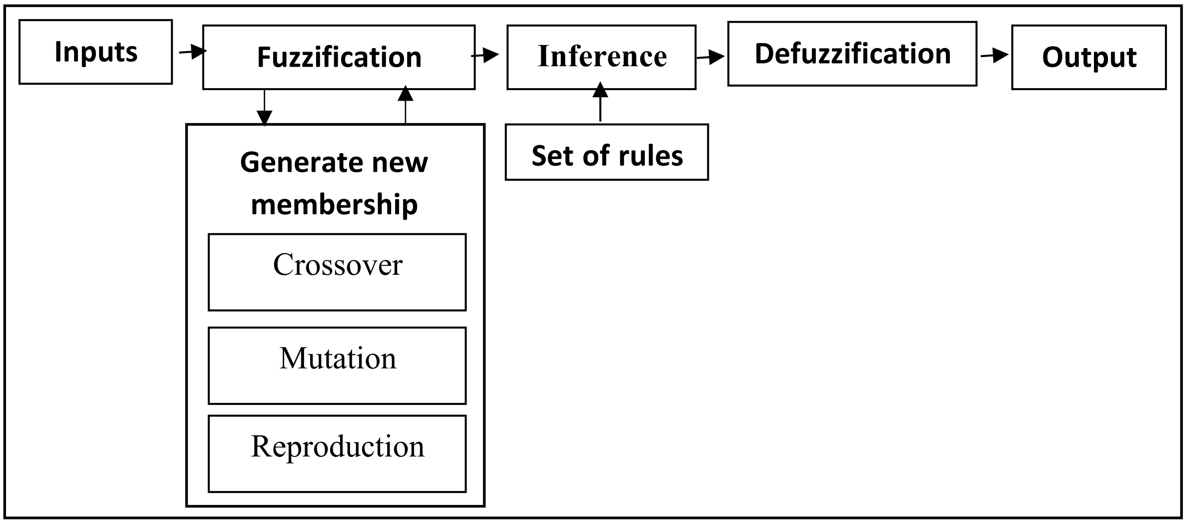

Figure (1) framework of using a classification procedure

5.1 Fuzzification and Genetic algorithm process

Fuzzification used to plan the genuine or fresh inputs to fluffy sets. We register levels of participation for every component utilizing enrollment capacities. For each information and result variable chose, we characterize at least three membership functions (MF), for instance: low, moderate or high. The state of these capacities can be triangles, Gaussian, Singleton and trapezoids .The result of a participation work consistently restricted to somewhere in the range of nothing and one. We utilize the triangular shape participation work since it exact and generally utilized. A triangular MF determined by three boundaries {a, b, c} as follows:

$$\text{triangle}(x; a, b, c) = \begin{Bmatrix} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & c \le x \end{Bmatrix} \dots \quad (2-1)$$

$$\text{triangle}(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \dots \quad (2-2)$$

The oundaries {a, b, c} (with a < b < c) determine the x coordinates of the three corners of the underlying triangular MF.

After highlights extraction step, the levels of membership for each worth in the element vector compute in fuzzification step based on the membership function and determine variable for each value. Table (1) show the eleven inputs and output parameter. For example, five degrees of parameter define for F variable, very low value [0 - 0 .36], low value [0.16 - 0.46], moderate value [0.26 - 0.56], high [0.5- 0.75] and very high [0.65 - 1]. After that, membership function is calculated.

TABLE (1). INPUTS AND OUTPUT PARAMETERS

Variable	No of feature	Linguistic Variables	Range	Linguistic Value	Parameter
Input	1	S	0 - 1	V Low	0 - 0 .36
				Low	0.16 - 0.46
				Mod	0.26 - 0.56
				High	0.5- 0.75
				V High	0.65 - 1
	2	F	0 – 20	V Low	0 – 2.5
				Low	2 – 7
				Mod	4 – 10
				High	7 – 15
				V High	10 - 20
	3	M	0 – 20	Low	0 – 7
				Moderate	5 – 14
				High	12 – 20
	4	I	0 - 10	Low	0 – 3
				Moderate	2 – 7
				High	4 - 10
	5	G	0 - 1	Very Low	0 – 2.26
				Low	0.2 – 0.4
				Moderate	0.3 – 0.6
				High	0.55 – 0.8
				Very High	0.7 - 1
	6	E	0 - 1	Low level	0 – 0.12
				Moderate level	0.06 – 0.23
				High level	0.16 – 1
	7	V	0 - 10	Low level	0 – 4
				Moderate level	3 – 7

				High level	6 - 10
	8	Z1	0 - 20	Low level Moderate level High level	0 – 2 1 – 5 4 - 20
	9	Z2	0 - 20	Low level Moderate level High level	0 – 2 1 – 5 4 - 20
	10	Z3	0 - 20	Low level Moderate level High level	0 – 2 1 – 5 4 - 20
	11	SW	0 - 20	Low level Moderate level High level	0 - 2 1- 5 4 - 20
Out		R	0 - 100	Irrelevance Low Moderate High	0-40 30-65 50-85 75-100

5.2 Inference Process

After the process of fuzzification, the process of inference is the process of drawing inputs to the output and give a decision of classification. Rules are a collection of linguistic Expressions. Inference used rules IF-THEN to transform the fuzzy input into fuzzy output. In the previous work, fuzzy logic based method used human expert knowledge to formulate fuzzy rules, But in this work, we use the human expert knowledge and predefined classified training data used to extract a set of fuzzy rules in addition to other rules. The results are more accurate than the previous method and the number of tweets that classified more than the previous method. Some of these rules define as follows:

1) If I: high level ^ S: very high level ∨ high level ^ Z1: high level →R: high relevant level.

2) If S: high level ^ Z3: Moderate level ^ L3: low level →R: moderate relevant level.

3) If L: moderate level ∨ E: low ^ G: low, → R: low relevant.

4) If M: high level ^ S: very low level ^ Z1, Z2, Z3=zero ^ SW is low level → R is irrelevance level.

As per the above rules, We give an itemized clarification of these principles. oftentimes utilized words and words have serious level in tweet and words number in List D are high this shows that tweet High significant degree to Hurricane Sandy tropical storm. The tweet has a place with a moderate applicable when the level of its words is high worth and tweet's length is low and the quantity of words is moderate inside the 50 words generally utilized in list D, shows that the client posted tweet with basic, significant words and short tweet length. Assuming the heaviness of tweet is low and every now and again used words' weight is low and the quantity of significant words in list D demonstrate to there are minimal significant words or significant words, so the level of tweet is low Relevant to sandy, and the tweets are ordered immaterial on the grounds that significant words Linked to Hurricane Sandy not found. At the point when the pace of words in List D is high and the quantity of words having a place with Hurricane Sandy in the preparation information isn't in the rundown of the main words is moderate For this situation the Tweet is profoundly pertinent to Hurricane sandy.

5.3 Defuzzification process

Defuzzification is cycle of produce quantifiable outcomes in genuine rationale. It must executed to change fuzzy outcomes over to genuine worth dependent on fuzzy sets and comparing participation degrees. There are set of defuzzification capacities recommended in investigates, similar to centroid, Center of Sums Method (COS) , bisector , and mean of the greatest (MOM), smallest of the maximum (SOM) and First of Maxima Method (FOM). Output (R) is unique value defuzzified from overall fuzzy set contain values of output based of defuzzification functions. The essential and troublesome thing are to check the correction rate. Note that users tend to express unique feelings and opinions so the results are extraordinary.

6 Training phase

At this stage, training data contain 1000 tweets that have been previously classified as relevant (low, moderate and high) used to obtain the most frequently used 50 words after passing through the initial processing phase for each tweet. The same training data used to obtain additional fuzzy rules used in the Inference process. Figure(3) show training phase.

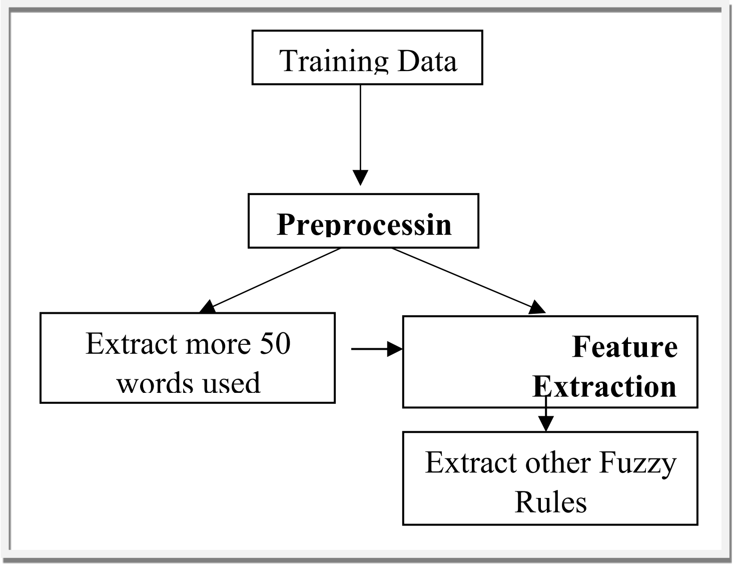

Figure (3) show training phase.

7 Experimental results

7.1 Comparison of defuzzification functions

After the process of extraction feature, the features vector contains number and percentage of each word and eleven feature value for every tweet. Eleven features utilized as the input to the classification procedure. The classification procedure Pass through three steps of Fuzzification process, Inference process, Defuzzification process. The essential and troublesome thing is to check the correction rate. Note that clients tend to express unique feelings and opinions so the results are extraordinary. In the first place, we utilize the training data classified manually composed of 600 tweets, isolated to 300irrelevant and 300 relevant (low, moderate and high relevance). So to look at this strategy through the defuzzification functions and analyze these functions. Table (2) show Polar relevance problem's results. Table (3) show Four-degree relevance problem's results.

Table (2) Polar relevance problem's results

Function	Relationship	First dataset	Second dataset	Third dataset
Centroid	**Irrelevant** **relevant**	99.8 % 99.6 %	99.7 % 95 %	100 % 98.6 %
Bisector	**Irrelevant** **relevant**	99.6 % 99.8 %	99.7 % 95 %	98.6 % 97.2 %
Mean of Maximum	**Irrelevant** **relevant**	99.8 % 99.5 %	99.8 % 98 %	100 % 98 %
Smallest of Maximum	**Irrelevant** **relevant**	99.7 % 98.8 %	99.8 95 %	100 % 96.4 %
Largest of Maximum	**Irrelevant** **relevant**	98.9 % 99.7 %	98.9 % 95 %	96 % 97.4 %

Table (3) Four-degree relevance problem's results

Function	Relationship	First dataset	Second dataset	Third dataset
Centroid	**Irrelevant** **Lowly** **Moderately** **Highly**	99.8 % 78.2 % 72 % 100 %	99.7 % none 80% 100%	100 % 72% 70 % 98.7 %
Bisector	**Irrelevant** **Lowly** **Moderately** **Highly**	99.6 % 70.4 % 95.9% 69%	99.7 % None 79% 59.3%	98.6 % 68.5% 49.6% 69.5%
Mean of Maximum	**Irrelevant** **Lowly** **Moderately** **Highly**	99.8% 59.1% 59.7% 98.7%	99.8% None 79% 59.6%	100 % 68 % 89% 69%
Smallest of Maximum	**Irrelevant** **Lowly** **Moderately** **Highly**	99.7 % 29.8 0. 0 9 99.3%	99.8 None 79 % 100%	100 % 35.5 0.09% 98%
Largest of Maximum	**Irrelevant** **Lowly** **Moderately** **Highly**	98.8 59% 60% 76%	98.9 None 79% 68.6%	96 % 54% 62.5% 66.9%

In the process of testing the defuzzification functions, Three sets of training data are selected. Three sets of datasets utilized from test data. Every dataset contains 200 tweets. The distinction between them is a proportion of irrelevance to relevance, which is 1: 9, 1: 1 and 9:1, separately. This design contains unbalanced and balanced data. The defuzzification functions used to look at amongst them and pick the best function. Note that in the second dataset index, there is no related low example. Therefore, the signs in the tables are "none" And here differences in the results between defuzzification functions and this is naturalistic because the function dependably gives different results.

7.2 Comparison with the fuzzy logic method

The research presents the specifics of the method based on fuzzy logic [3]. First, divided and classified a collected data as training data, the second step is preprocessing. In this step, process each tweet to eliminate the additives that effect to the classification process and then seven input features extracted from each tweet in the feature extraction step. These features used as input to the classification model. Classification model passes with three steps Fuzzification, Inference and Defuzzification. Table (4) shows comparison results between this system and fuzzy logic basedmethod for text classification.

TABLE (4). Results between Fuzzy logic method and proposed system.

No.	Fuzzy Logic Based Method			Proposed system			λ
	A	B	α	A	B	α	
1	141	135	95.71 %	160	158	98.751 %	17.031 %
2	161	157	97.71 %	184	183	99.451 %	16.561 %
3	128	126	98.41 %	147	146	99.311 %	15.871 %
4	137	132	96.31 %	154	152	98.701%	15.151 %

7.3 Comparison with the Naïve Bayes classifier

In research [4], recommended that the issues of immaterial information evacuation and clamor decrease are like the email spam sifting. They prepared a Naïve Bayes classifier for important information recognition. Table (5) shows correlation results between this framework and Naïve Bayes classifier.

TABLE (5) Results between Naïve Bayes classifier and proposed system.

No.	Naïve Bayes classifier			Proposed system			λ
	A	B	α	A	B	α	
1	133	132	99.2 %	160	158	98.75 %	19.69 %
2	149	148	99.3 %	184	183	99.45 %	23.64 %
3	124	122	98.3 %	147	146	99.31 %	19.67 %
4	122	120	98.3 %	154	152	98.70%	26.66 %
5	119	118	99.1	138	136	98.55%	15.25%

8. Conclusion

The aftereffects of the proposed system contrasted with and fuzzy logic-based method and Naïve Bayes classifier dependent on the adjustment rate and gradual rate. The amendment pace of proposed framework for every informational index are (98.75%, 99.45%, 99.31%, 98.70%, 98.55%) however the remedy rate fuzzy logic technique are (99.2%, 99.3%, 98.3%, 98.3%, 99.1%) and Naïve Bayes classifier are (95.7, 97.7, 98.4, 96.3, 96.7) in grouping. At the gradual rate, the proposed framework can extricate tweets more than this technique, where in dataset 1 the number of the tweets removed by the proposed framework is 160 tweets however the quantity of the tweets that separated by the Keyword search strategy, Naïve Bayes classifier and fluffy rationale based strategy are 98, 133 and 141 in a grouping. In this case, the proposed system is better than the above-mentioned methods.

References

[1] Beal, V. (no date) What is IM (instant message)? Webopedia definition. Available at: http://www.webopedia.com/TERM/I/IM.html

[2] Anirban D., Petros D., Boulos H., "Feature Selection Methods for Text Classification." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

[3] Maha A., Saadya F.," Developing a Spam Email Detector", International Journal of Engineering and Innovative Technology (IJEIT) ,Vol. 5, Issue 2, August 2015.

[4] S. Appavu and Ramasamy R.,” Suspicious E-mail Detection via Decision Tree: A Data Mining Approach”, Journal of Computing and Information Technology - CIT 15, 2007, 2, 161–169

[5] Hongwei Mo. "Immune Algorithm Optimization of Membership Functions for Mining Association Rules", Lecture Notes in Computer Science, 2006.

[6] KeYuan Wu, MengChu Zhou, Xiaoyu Sean Lu, Li Huang. "A fuzzy logic-based text classification method for social media data", 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017