Classifying Texts of Twitter Data Using a Modified Fuzzy Logic Method

Yossra H. Ali, Ph.D. (Asst.Prof.)[*] **Nuha J. Ibrahim, Ph.D. (Lect.)**[*]

Mohammed A. Jaleel[*]

Abstract: Social media are a modern web-based application for communication between humans. People share their interests and activities with these Applications. Twitter is a social media site, where people communicate through tweets. People publish their tweets on their profile and send their followers to express their thoughts and opinions about events in this world. In this research, a modified fuzzy logic method to disband text classification problem. The Inputs for this classification system are a set of features extracted from a tweet and the output of this system is a decision of classification for a tweet, which is a degree of correlation for each tweet to an appointed event where the degree of relevance to the desired event if it irrelevant or relevant. The results compared with the keyword search method and the previous fuzzy logic based method based on terms of correction rate and incremental rate. In the incremental rate, the proposed system is able to extract tweets more than a previous fuzzy logic based method, where in dataset 1 the number of the tweets that extracted by the proposed system is 154tweets but the number of the tweets that extracted by the other one are 98 and 141. The correction rate of the proposed system is (98.7) but the correction rates of these methods are (97.9) and (95.7).

Keywords*:* Social media; text classification; fuzzy logic.

[*] Computer Science Department,\ University of Technology, Baghdad, Iraq

1. Introduction

Social media provides a platform for people to express their thoughts and opinions and reciprocity messages among themselves. The data contain a set of features related to the users' opinions and scarce events, it is important to pick benefits from them. In data of social media, there is information in tweets and messages that reduce their usefulness and thus reduce the amount of information extracted. A scarce event, such as Hurricane Sandy, affects the world and make social media users to express their opinions in social media [1].

Text classification is a core problem for many applications, like sentiment analysis, spam detection or smart replies. It is a problem studied widely and several methods implemented to process this complex issue. The goal of text classification is to assign documents to one or many categories. Most of the methods proposed rely on representing the text to classify with a text vector. In the simplest form, this vector contains the frequency with which each word occurs in the text. It can also represent several features of interest that extracted from the text [2].

Fuzzy logic is a method for manipulating and representing uncertain information and deals with ambiguity excellence. It is alike to natural language and it approximates to thinking of human [3]. An important feature to the fuzzy logic is the computerization of words. The methodology can convert words into numerical values of reasoning, computing and deal with linguistic value, so fuzzy logic is a good method to deal with linguistic problems. However, given the unstructured and non-crisp of natural languages like the English language, tools of fuzzy logic can use it to deal with text classification [4].

In this research, a proposed fuzzy logic method used as a modification method for fuzzy logic based method in the research [5] for classification Twitter data and used Hurricane Sandy as a case study. A fuzzy logic uses a set of features derived from each tweet in a data set. The values of these features are converted to linguistic variables, compute a degree of membership based on a membership function and then based on set of fuzzy rules, returns relevant degree of the tweet to the event if it is relevant or not.

2. Related Works

The classification of texts aims to put texts into predefined classes. Researchers suggest a set of algorithms and methods to solve this problem. KeYuan et al. [5] introduced a fuzzy logic-based model to solve the problems of text classification based on social media data. Multiple features extracted from each tweet in the fuzzy logic-based model computes according to predetermined fuzzy rules and fuzzy membership functions. It returns a relevance degree whether a message is related to a particular event or not. Gong et al. [6] used enhanced term frequency-inverse document frequency (TF-IDF) technique for classification text using lemmatization and stemming. Prusa et al. [7] suggested used method of feature selection based on Threshold, Chi-square, and Statistic from First-Order in process of feature selection. Yet, due to the huge data and high dimension of features, there is yet a great space for enhancing the effectiveness of classification. Prusa and Khoshgoftaar [8] utilized Convolutional Neural Networks in addition to the modern encoding way in the classification of text. Wang et al. [9] used an enhanced technique, it's deep feature weight Naïve Bayes via largest Estimation probability to compute conditional and prior likelihood. Bidi and Elberrichi [10] implemented feature selection using Genetic Algorithm (GA), and then combine several classifiers to establish if GA worthy use in this area. Spielhofer et al. [11] suggested the method to solve the problem of noise reduction irrelevant data removal. They trained the classifier of Naïve Bayes for detection relevant data.

3. A Fuzzy Logic Based Text Classification of Previous Method

Text classification method based on fuzzy logic [5]. A collected data are divided and classified as training data, then the first step preprocessing. In this step, process each tweet to eliminate the additives that effect to the classification process. Seven input features extracted from each tweet in the feature extraction step, this features used as input to the classification model. The classification model passes through three steps. The fuzzification is to convert the real input into fuzzy sets containing membership degree using membership functions. The trapezoidal shape function of membership selected because it is accurate, used frequently and simple. Inference process describes the second step, draws the mapping from inputs to outputs, and uses the IF-THEN rules to transform fuzzy inputs to the fuzzy output. The final step is the defuzzification phase to obtain the real output. There are many defuzzification functions, such as centroid, maximum average (MOM) and largest of maximum (LOM).

3.1 Data Collection

Tweets are aggregate through the Twitter Application Programming Interface (API). In social media, one of the important tools is Twitter, let users cast their thinner on certain issues by tweets that are no longer than 140 characters. Data collected during the period from 10.27.2012 to 11.7.2012. Each record based on the location, timestamp, identifier, text data, and date. This information filtered and get the text data only, then processed, extract features and classify them.

After the data is collect, set of 1000 tweets have been randomly selected from initial data and considered as training data. Tweets are people's ideas and opinions so it has not contained contextual information. So, a set of tweets classified manually that are used as training data for a fuzzy logic approach. Score each tweet using zero or one to indicate irrelevant or relevant. Each tweet has a collection of score that interval from zero to 15. Four score intervals are defined to describe a degree of relevance to a tweet based on irrelevance L1 [zero, 5), low relevance L2 [5, 9), moderate relevance L3 [9, 12) and high relevance L4 [12, 16].

3.2 Data Preprocessing

Tweets contain non-useful data in the process of categorizing text such as a label, numbers, and stop words. It is important to remove these additions or manipulate them in tweets so as not to affect the classification process. In this research, a pattern matching have been used to eliminate these additives by checking a given sequence of tokens for the presence of the constituents of some pattern. For example, a URL with a static pattern starting with "http" will be deleted when it is found. Also numbers, label, and special characters. The stop words removal step is the most important step in the preprocessing. These words are rarely useful in the classification process [5][12].

3.3 Feature Extraction

Some words appear more frequently than others do. Each tweet has a collection of score that interval from zero to 15. Four score intervals are defined to describe a degree of relevance to a tweet based on irrelevance L1 [zero, 5), low relevance L2 [5, 9), moderate relevance L3 [9, 12) and high relevance L4 [12, 16], respectively. After defining the four periods L1, L2, L3, and L4 each tweet belong to L2, L3, and L4 is choose from the training data collection then process each tweet. The frequency of each word in tweets is calculated and choose 50 words used frequently [5]. Word's importance for every word ai define as:

$$ai = \frac{Ai}{Bi} \tag{1}$$

where Ai decides words number in tweets that belong to L2, L3, and L4. Bi represent words number in each tweet that belong to L1, L2, L3, and L4; ai is a rate that represents word's important i. Next step is sort most words used frequently based on ai of smallest size to configure D list and then compute a similarity between a word in llist D and word in a tweet. Tweet has n-words, Ti indicates to word jth in a tweet and i ϵ {1… n}. ck Indicates to the kth word in list D and k ϵ {one…50}. Highest value from similarity scores chosen to represent Si. So, Si score defines as:

$$Si = max\ (ck\ \times Ti \otimes ck), i \in [1, n] \tag{2}$$

Si is basic value. This features vector extracted from each tweet and its uses through a modified fuzzy logic method. Details of features as shown:

1. Largest word's score in a tweet (Gj)

$$Gj = \max_{1<=i<=n} \mathrm{Si} \tag{3}$$

2. score of a tweet (Kj)

$$\mathrm{Kj} = \sum_{i=0}^{n} \mathrm{Si} \tag{4}$$

where Kj decides a tweet's accumulate score of words

3. length of a tweet (Nj)

$$Nj = n \tag{5}$$

where n indicates a number of words in a tweet after tweet processing.

4. Frequently used words number in a tweet (Mj)

Mj indicates the number of words in a tweet and it same to words in list L. List L contain words used frequently and use to compare with all tweets.

5. The weight of tweet (Wj)

$$Wj \ = \frac{Kj}{Nj} \tag{6}$$

where Wj is the mean score of words in the tweet j.

6. Indicates (Xj) to frequently used words weight in a tweet

$$Xj \ = \frac{Mj}{Nj} \tag{7}$$

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

7. Indicates (Vj) to the number of patterns in the jth tweet (Vj)

After getting a list, there are useful words found in training data more than 50 important words that come on their own but are not on the list. e.g. ‘not safe' and 'not expected' terms beneficial more than just one term such as 'not' or ' safe'. So, Vj is a number of this type of pattern in a tweet.

3.4 Fuzzy Logic Method

Seven features used as input to the model. The model pass through three steps of Fuzzification, Inference, Defuzzification. Figure (1) shows the framework of using a fuzzy logic-based model in [5].

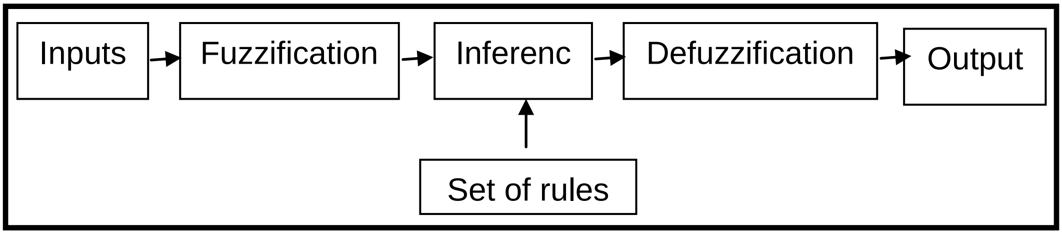

Figure 1. fuzzy logic method-based model (adapted from [11], p. 1945)

Fuzzification used to map a real inputs into fuzzy sets. An inference is a process of drawing inputs to the output and give a decision of classification. It used rules IF-THEN to transform fuzzy input to fuzzy output. Defuzzification is a process of generating result in real logic.

4. Proposed Method

A modified fuzzy logic method used to disband text classification problem In this research. Figure (2) describe a Procedure of this work.

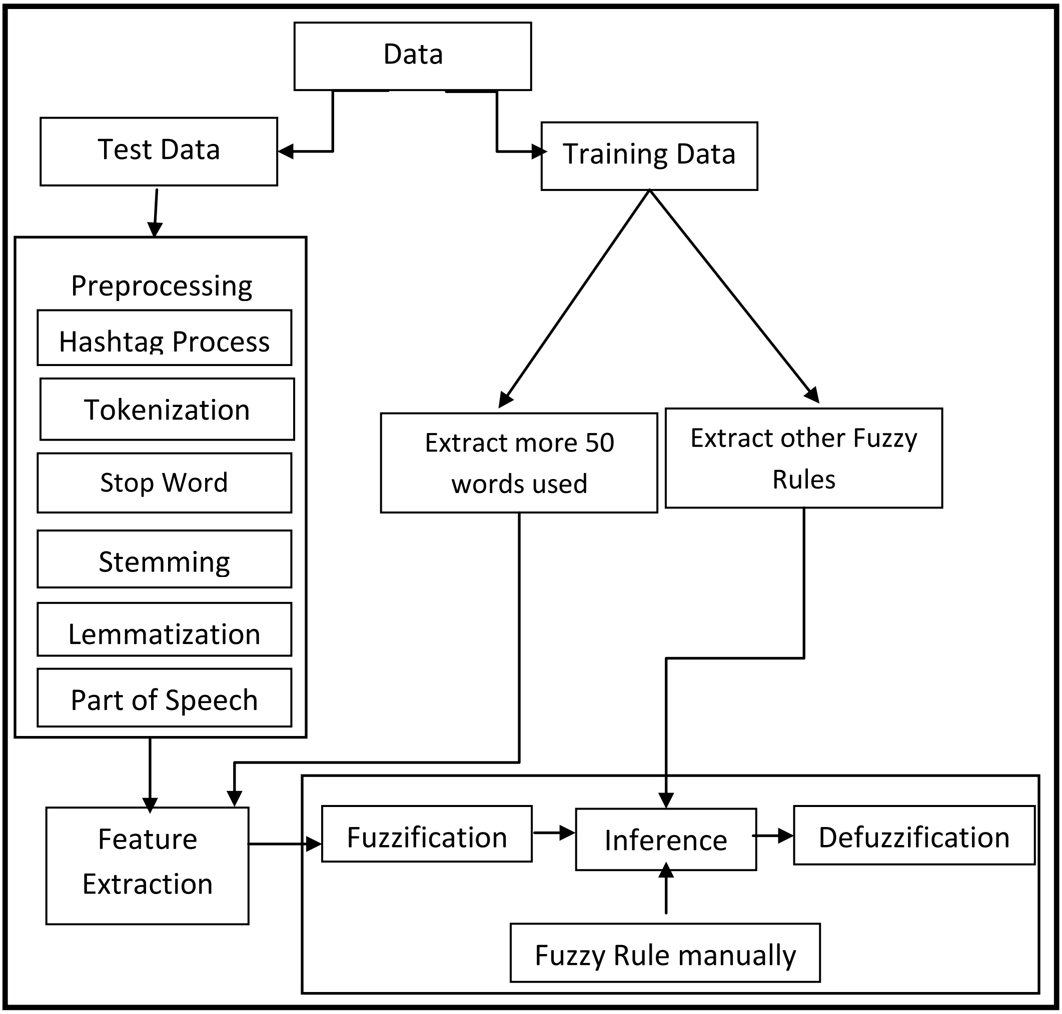

Figure 2. Procedure of this modified method

Below is the details explanation of the proposed method

4.1 Data Collection

This system is based on the Sandy Hurricane event, where the final results are compared previous studies used this event data. data is collected from the time period 10.27.2012 to 11.7.2012. Each record contains the text data, date, location, timestamp and publisher, where text data is extracted only and is grouped into a one dataset of initial data. Set of 1000 tweets taken from initial data chased as training data. This training data used to extract More than 50 words used frequently and extract additional fuzzy rules used in the inference phase. Each tweet utilizing one or zero to refer to relevance degree. Each tweet has an accumulation of score that interim from 0 to 15. Four scores to depict an important level of relevant to a tweet based on irrelevance L1 [zero, 5), low relevance L2 [5, 9), moderate relevance L3 [9, 12) and high relevance L4 [12, 16], respectively. For comparison of defuzzification functions phase, training data classified manually composed of 600 tweets, isolated to 300 irrelevant and 300 relevant (low, moderate and high) utilize to compare between defuzzification functions. Set of 1002 tweets take from initial data as testing data, where this test data divide into five datasets to process it.

4.2 Data Preprocessing

In addition to steps of removing URL, numbers, label, special characters and stop words. another. There are many internal processes proposed in this work for preprocessing of text such as manipulate hashtag, remove URL, remove special character, remove additions, tokenization, remove stop words, stemming, lemmatization and Part of Speech (POS). Collected and divided tweets is the inputs to the preprocessing step and the output of this step is a series of important words that used in the feature extraction step. The Hashtag is a set of words that are not fluent and there are no spaces between them, Like #SandyHurric. User understands directly, but the program cannot distinguish them. In this case. Use comparison words of Hashtag with literally English words. e.g., can identify "Sandy" after the fifth operation in #SandyHurric, because parts of a term can be characterized as "S", "Sa", "San", "Sand", "Sandy", respectively. In addition, using the Stemming, Lemmatization, and Part of Speech (POS) for each word in the tweet to more flexibility and accuracy in classification. The last step is to convert each word to a lower value word.

4.3 Feature Extraction

Extraction features are an important step in the classification process. In this work, seven features are extracted for each tweet, In addition to these features, four additional features are extracted to give a more accurate result and to classify more tweets. Divide the list D into three evenly distributed subgroups that refer to Z1, Z2 and Z3 with different weights ϴ1, ϴ2, and ϴ3, respectively as follows:

1 - More words used in the list D (Z1)
2 - Words used normally in the D menu (Z2)
3 - Less commonly used in the list D (Z3)

$$\theta k = \begin{cases} \theta 1 & for\ Z1 \quad k \epsilon [1,17) \\ \theta 2 & for\ Z2 \quad k \epsilon [17,33) \\ \theta 3 & foe\ Z3 \quad k \epsilon [33,50] \end{cases} \tag{8}$$

4- Number of Sandy words (SW) not found in list D, but this words used frequently in training data that related to Hurricane Sandy.

4.4 Proposed Modified Fuzzy Logic Method

After the feature extraction process, the features vector includes eleven values for each tweet. Features vector passes through three steps which are Fuzzification process, Inference process, and Defuzzification process as shown in algorithm 1.

```
Algorithm 1: Three steps of the proposed fuzzy logic method
Input: Predefined classified training data, Feature vector for each tweet contain eleven values.
Output: Decision of classification.
Step 1: Generate fuzzy rules from predefined classified training data.
Step 2: Fuzzification process
        2.1 Select membership function
        2.2 For each value in the feature vector
                Compute degree of membership using membership functions
        2.3 Map the crisp or real input to fuzzy set
Step 3: Inference process
        3.1 Write a set of IF-THEN fuzzy rules
        3.2 Decision Making based on these fuzzy rules in addition to fuzzy rules
        extracted is step1
Step 4: Defuzzification process
        1-  Select Defuzzification function
        2-  Transform the fuzzy results into real value
Step 5: A print of decision and the real value of the result
End
```

4.3.1 Fuzzification

Fuzzification used to map the real or crisp inputs into fuzzy sets. Degrees of membership for each element are compute using membership functions. For each input and output variable selected, three or more membership functions are define. In this research, the triangular shape membership function used because of it accurate and widely used. The eleven linguistic variables output and inputs are shown in Table 1, offer several domains for parameters.

Table 1. Input and output parameters

Variable	**Feature Name**	**Linguistic Variables of Feature**	**Range**	**Linguistic Value**	**Parameter**
Input	Largest score of word in a twee	Gj	0 - 1	Very Low Low Moderate High Very High	0 - 0 .36 0.16 - 0.46 0.26 - 0.56 0.5- 0.75 0.65 - 1
	score of a tweet	Kj	0 – 20	Very Low Low Moderate High Very High	0 – 2.5 2 – 7 4 – 10 7 – 15 10 - 20
	Length of a tweet	Nj	0 – 20	Low Moderate High	0 – 7 5 – 14 12 – 20
	Frequently used words number in a tweet	Mj	0 - 10	Low Moderate High	0 – 3 2 – 7 4 - 10
	Weight of tweet	Wj	0 - 1	Very Low Low Moderate High Very High	0 – 2.26 0.2 – 0.4 0.3 – 0.6 0.55 – 0.8 0.7 - 1
	Frequently used words weight in a tweet	Xj	0 - 1	Low Moderate High	0 – 0.12 0.06 – 0.23 0.16 – 1
	patterns number in a tweet	V	0 - 10	Low Moderate High	0 – 4 3 – 7 6 - 10
	More words used in the list D	Z1	0 - 20	Low Moderate High	0 – 2 1 – 5 4 - 20

	Words used moderately in the D list	Z2	0 - 20	Low Moderate High	0 – 2 1 – 5 4 - 20
	Less commonly used in the list D	Z3	0 - 20	Low Moderate High	0 – 2 1 – 5 4 - 20
	Number of words not found in the D, but belong to Sandy words	SW	0 - 20	Low Moderate High	0 - 2 1- 5 4 - 20
Output		R	0 - 100	Irrelevance/DK Low Relevance mod Relevance High Relevance	0-40 30-65 50-85 75-100

4.3.2 Inference

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 this work, the human expert knowledge is used and predefined classified training data used to extract a set of fuzzy rules in addition to other rules. Some of these rules define as follows:

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

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

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

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

5) If Z1: high ^ Z2: high ^ SW: Moderate, → R: High relevant.

In accordance with the above rules, the detailed explanation of these rules as follow. Frequently used words and words have a high degree in a tweet and words number in List D are high this indicates that tweet high relevant degree to Hurricane Sandy. The tweet belongs to a moderate relevant when the degree of its words is high value and tweet's length is low and the number of words is moderate within the 50 words most used in list D, indicates that the user posted a tweet with critical, important words and short tweet length. if the weight of tweet is low and frequently utilized words’ weight is low and the number of important words in list D indicate to there are little important words or major words, so the degree of tweet is low Relevant to sandy, and the tweets are categorized irrelevant because important words Linked to Hurricane Sandy not found.

4.3.3 Defuzzification

Defuzzification is the process of generating quantifiable results in real logic used to convert fuzzy results to real value based on fuzzy sets and membership degrees. There are set of defuzzification functions suggested in researches, like centroid, Center of Sums (CoS), bisector, and mean of the maximum (MoM), smallest of the maximum (SoM) and First of Maxima method (FoM). Output (R) is a unique value defuzzified from overall fuzzy set contain values of output based on defuzzification functions.

5. Experimental Results

The proposed method implemented using Python 3.6.3 programming language and windows 7 64 bit Operating System. Results compare with a keyword search method and fuzzy logic based method.

5.1 Comparison of defuzzification functions

The important thing is to check a correction rate. Note that users tend to express different opinions so the results are different. In beginning, The training data classified manually composed of 600 tweets, divided into 300 irrelevant tweets and 300 relevant tweets (low, moderate and high relevance) are used to examine this method through defuzzification functions and compare these functions. Table (2) illustrated the polar relevance problem's results. Table (3) illustrated the four-degree relevance problem’s results.

Table 2. Polar relevance problem's results

function	Relationship	First dataset	Second dataset	Third dataset
Centroid	Irrelevant relevant	99.4 % 99 %	99.8 % 93 %	100 % 97 %
Bisector	Irrelevant relevant	99 % 99.4 %	99.8 % 93 %	97 % 96.4 %
Mean of Maximum	Irrelevant relevant	99.4 % 99 %	99.4 % 97 %	100 % 97 %
Smallest of Maximum	Irrelevant relevant	99.8 % 98.4 %	99.8 93 %	100 % 95.3 %

Largest of Maximum	Irrelevant	98.3 %	98.9 %	95 %
	relevant	99 %	93 %	96.3 %

Table 3. Four-degree relevance problem's results

Function	Relationship	First dataset	Second dataset	Third dataset
Centroid	Irrelevant	99.4 %	99.8 %	100 %
	Lowly	71.2 %	none	72%
	Moderately	72 %	80%	70 %
	Highly	100 %	100%	98.7 %
Bisector	Irrelevant	99 %	99.8 %	97 %
	Lowly	70.4 %	None	68.5%
	Moderately	95.9%	79%	49.6%
	Highly	69%	59.3%	69.5%
Mean of Maximum	Irrelevant	99.4 %	99.4 %	100 %
	Lowly	59.1%	None	68 %
	Moderately	59.7%	79%	89%
	Highly	98.7%	59.6%	69%
Smallest of Maximum	Irrelevant	99.8 %	99.8	100 %
	Lowly	29.8	None	35.5
	Moderately	0. 0 9	79 %	0.09%
	Highly	99.3%	100%	98%
Largest of Maximum	Irrelevant	98.3	98.9	95 %
	Lowly	59%	None	54%
	Moderately	60%	79%	62.5%
	Highly	76%	68.6%	66.9%

Three datasets used from test data. Each dataset contains 200 tweets. Defuzzification functions applied to compare between them and choose the best function. In Table 2 and Table 3, a results show that the modified fuzzy logic method achieves superior accuracies to the problem of the polar relevance and Its ability of deal with the four-degree suitability problem is not as good as that of dealing with the polar relevance and centroid performance is highly efficient but LOM is worse. The centroid defuzzification function is choose because of its ability to give well results and better than other functions.

5.2 Compare with keyword search method

Researches [14-15] used keyword search method; it used for the extraction related tweets from the main dataset. The benefit of Keyword search method is highly accurate, efficient and straightforward for tweet has highly relevant. The fault of this method, that is incapable to extract sufficient tweets relevant. Comparison based on five datasets of tweets from the test data. Table (4) shows the comparison results between the modified fuzzy logic method and keyword search method based on correction rate and incremental rate. A correctness rate is:

$$\alpha = \frac{A}{B} * 100\,\% \tag{9}$$

where B set related tweets' number extracted from every methoda and A is correctly categorized tweets number in B. A is calculated by manually check and it is confirmed by the training of data, i.e. the number of tweets categorized correctly in B is calculate. also, an incremental rate (λ) is describe that fuzzy logic able to exploit the information more than well-known keyword search method, which it is defined as:

$$\lambda = \frac{Af - Ak}{Ak} * 100\,\% \tag{10}$$

where Af compute by the modified fuzzy logic method and Ak compute by keyword search method.

Now, compare between the keyword search method and modified fuzzy logic method based on correction rate and the incremental rate shown in Table 4.

Table 4. Comparison results between modified fuzzy logic method and keyword search

Data Set	Keyword Search method			modified Fuzzy Logic method			λ
	X	Y	α	Y	X	α	
1	98	96	97.9 %	154	152	98.7 %	58.3 %
2	103	103	100 %	178	177	99.4 %	71.84 %
3	86	86	100 %	140	139	99.2 %	61.6%
4	93	92	98.9 %	147	145	98.6%	57.6%
5	99	99	100 %	138	136	98.5%	15.25 %

From the result, it has been show that all extracted tweets by keyword search method are subsets from this modified method. In Table 4, the rates of λ indicate that the modified method successfully revises additional tweets more than the fuzzy logic method for text classification.

5.3 Comparison with the fuzzy logic-based method

The difference between modified fuzzy logic methods with the previous fuzzy logic method for text classification illustrated in table (5).

Table 5. Results between Fuzzy logic based method and modified fuzzy logic method.

Data Set	Fuzzy Logic Based Method			Proposed Fuzzy Logic Algorithm			λ
	X	Y	α	Y	X	α	
1	141	135	95.7 %	154	152	98.7 %	12.59 %
2	161	157	97.7 %	178	177	99.4 %	12.73 %
3	128	126	98.4 %	140	139	99.2 %	10.3 %
4	137	132	96.3 %	147	145	98.6%	9.8 %
5	122	118	96.7	138	136	98.5%	15.25%

Modified fuzzy logic method able to extract tweets more than the fuzzy logic method and well know keyword search method. With considering the incremental rate, the modified method is powerful more than the fuzzy logic based method and keyword search method. With considering correctness rate values, a keyword search method doing a little best than the fuzzy logic based method but the modified method is better than the fuzzy logic method and approximate to keyword search method. Considering both standards, the modified fuzzy logic method choose in research, where relevant tweets are highly required for the analysis step. High correctness rate value and high quantity able to guarantee more informative and useful. So the modified method is better than the previous methods where it can ensure the high rate value and a high quantity value of correction and tweets more that are relevant and classified accurately.

6. Conclusion

In this research, a modified fuzzy logic method for text classification from Twitter data has been proposed. By using a set of categorized training data and test data and derived eleven features from each tweet as inputs to the classification system. This modified method compare with two methods. The first is a method of keyword search and the second method is the fuzzy logic method for text classification. Results show that this modified method is appropriate to classify irrelevant or relevant tweets to hurricane Sandy more than the fuzzy Logic method. Additionally, by comparing Defuzzification functions commonly used, centroid function is more efficient and effective than the other Defuzzification functions. In future works, we aim to detect the best way to classify texts of twitter data, for example, using neural networks through Fuzzification step to evolve membership degrees or uses of the evolutionary algorithm and neural networks through Inference step to improve the construction of fuzzy rules.

References

[1] C. Chen, D. Neal, and M. Zhou, “Understanding the Evolution of a Disaster A Framework for Assessing Crisis in a System Environment (FACSE) “, Natural Hazards, vol. 65, no. 1, pp. 407-422, January 2013.

[2] M. Nogueira, O. Rezende, A. Camargo, “On the Use of Fuzzy Rules to Text Document Classification“, International Conference on Hybrid Intelligent Systems, 2010.

[3] B. Kosko, “Fuzzy Thinking: The New Science of Fuzzy Logic“, International Journal of General Systems, Hyperion, New York, June 1994.

[4] L. A. Zadeh, “ “Fuzzy logic = computing with words“, IEEE, vol. 4, no. 2, pp.103-111, May 1996.

[5] [5] KeYuan Wu, MengChu Zhou, Xiaoyu Sean Lu, and Li Huang, **“**A Fuzzy Logic-Based Text Classification Method for Social Media Data“, International Conference on Systems, IEEE, 2017.

[6] Y. Zhang, L. Gong, and Y. Wang. “An improved TF-IDF approach for text classification“, Journal of Zhejiang University Science, Vol. 6, no. 1, pp. 49-55, August 2005.

[7] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, “Impact of feature selection techniques for tweet sentiment classification“, International FLAIRS Conference, Florida, USA, May 2015.

[8] J. D. Prusa and T. M. Khoshgoftaar, “Designing a better data representation for deep neural networks and text classification“, International Conference on Information Reuse and Integration (IRI), IEEE, USA, pp. 411-416, July 2016.

[9] Q. Jiang, W. Wang, X. Han, S. Zhang, X. Wang and C. Wang, “Deep feature weighting in Naive Bayes for Chinese text classification“, International Conference on Cloud Computing and Intelligence Systems (CCIS), Beijing, China, pp. 160-164, August 2016.

[10] N. Bidi and Z. Elberrichi, “Feature selection for text classification using genetic algorithms“, International Conference on Modelling, Identification and Control (ICMIC), IEEE, Algiers, Algeria, pp. 806-810, November 2016.

[11] T. Spielhofer, R. Greenlaw, D. Markham, and A. Hahne, "Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management", 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), Vienna, Austria, pp. 1-6, December, 2016.

[12] A. Kasun, M. Manic, and R. Hruska, "Optimal stop word selection for text mining in critical infrastructure domain", Resilience Week (RWS), Philadelphia, pp. 1-6, August 2015.

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

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

[15] H. Dong, M. Halem, and S. Zhou, "Social media data analytics applied to hurricane sandy", International Conference on Social Computing (SocialCom), IEEE, pp. 963-966, September 2013.

تصنيف نصوص بيانات تويتر باستخدام طريقة منطقية ضبابية معدلة

أ. م. د. يسرى حسين علي* م .د. نهى جميل أبراهيم* محمد عبد الجليل*

المستخلص: تقدم وسائل التواصل الاجتماعي معلومات وفيرة لدراسة سلوكيات الناس وافكارهم وآرائهم حول ما يدور في العالم مثل الأمور السياسية والاقتصادية والكوارث الطبيعية وغيرها. من المهم دراسة وتحليل العلاقة بين الاحداث التي تؤثر على الانسان ووسائل التواصل الاجتماعي. تستخدم هذه الدراسة بيانات تويتر المرتبطة بأعصار ساندي لتصنيف النص. وبما أن النصوص التي يتم جمعها تحتوي على بيانات مختلفة لأحداث مختلفة، نحتاج إلى تصنيف البيانات التي لها علاقة بإعصار ساندي. في هذا العمل أستخدمنا طريقة محسنة يستند إلى المنطق الضبابي لحل مشكلة تصنيف النص. المدخلات لهذا النظام هي مجموعة من الميزات التي يتم استخلاصها من كل تغريده. الناتج هو مدى ارتباط كل رسالة إلى ساندي. يتم تصميم مجموعة من القواعد غير الواضحة ويتم الجمع بين طرق مختلفة للتشخيص من أجل الحصول على نتائج التصنيف المطلوبة. نحن نقوم بمقارنة النتائج المستخلصة مع دراسة سابقة استخدمت المنطق الضبابي لتصنيف رسائل تويتر المتعلقة بأعصار ساندي ونقارق بين نتائجها ونتائج طريقة البحث عن الكلمات الرئيسية المعروفة من حيث معدل التصحيح والكمية. تظهر النتيجة أن هذه الطريقة المحسنة هي أكثر ملائمة لتصنيف رسائل تويتر من طريقة الكلمات الرئيسية والنهج القائم على المنطق الضبابي.

الكلمات المفتاحية: وسائل التواصل الاجتماعي؛ تصنيف النص المنطق الضبابي.

* قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق.