Spring 5-31-2017

A fuzzy logic-based text classification method for social media

Keyuan Wu
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

A FUZZY LOGIC-BASED TEXT CLASSIFICATION METHOD FOR SOCIAL MEDIA

by
Keyuan Wu

Social media offer abundant information for studying people's behaviors, emotions and opinions during the evolution of various rare events such as natural disasters. It is useful to analyze the correlation between social media and human-affected events. This study uses Hurricane Sandy 2012 related Twitter text data to conduct information extraction and text classification. Considering that the original data contains different topics, we need to find the data related to Hurricane Sandy. A fuzzy logic-based approach is introduced to solve the problem of text classification. Inputs used in the proposed fuzzy logic-based model are multiple useful features extracted from each Twitter's message. The output is its degree of relevance for each message to Sandy. A number of fuzzy rules are designed and different defuzzification methods are combined in order to obtain desired classification results. This work compares the proposed method with the well-known keyword search method in terms of correctness rate and quantity. The result shows that the proposed fuzzy logic-based approach is more suitable to classify Twitter messages than keyword word method.

A FUZZY LOGIC-BASED TEXT CLASSIFICATION METHOD FOR SOCIAL MEDIA

by
Keyuan Wu

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Electrical Engineering

Department of Electrical and Computer Engineering

May 2017

APPROVAL PAGE

A FUZZY LOGIC-BASED TEXT CLASSIFICATION METHOD FOR SOCIAL MEDIA

Keyuan Wu

Dr. Mengchu Zhou, Thesis Advisor — Date
Distinguished Professor of Electrical and Computer Engineering, NJIT

Dr. Ali Abdi, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Hesuan Hu, Committee Member — Date
Professor in School of Electro-Mechanical Engineering, Xidian University

BIOGRAPHICAL SKETCH

Author: Keyuan Wu

Degree: Master of Science

Date: May, 2017

Undergraduate and Graduate Education:

- Master of Science in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2017

- Bachelor of Science in Electrical Engineering and Automation,
  Shanghai Ocean University, Shanghai, P. R. China, 2014

Major: Electrical Engineering

献给我的亲人

ACKNOWLEDGMENT

Foremost, I would like to express my deepest gratitude to my advisor, Dr. Mengchu Zhou, for excellent guidance, patience and bringing me to the palace of academic. Professor Zhou served as my research advisor, but he was very influential in the academic path chosen I have made and gave me a lot of excellent suggestions.

Special thanks to my committee members Dr. Ali Abdi and Dr. Hesuan Hu who gave their time and expertise in making this possible. To Xiaoyu Lu without whom this thesis would not be complete. He always answered my questions, guided me through this process, and laid the groundwork for this thesis. Also to Li Huang who is my motivation from the beginning.

Furthermore, my sincere thanks also go to my group members, Xiaoyu Lu, Li Huang, Liang Qi, Haoyue Liu, Xilong Liu and Jingchu Ji, and my good friends for giving me many excellent suggestions and assisting me to complete this thesis.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\otimes$	A Similarity Evaluation Process

CHAPTER 1

INTRODUCTION

We live in an era with high-developed Internet which enables human beings to share information swiftly and precisely. Social media, a product of Internet, provide a proper and flexible platform where users exchange messages. Since this sort of data is monstrously large and informative, it is worthy to take advantage of it. However, due to its large quantity, many vital messages could probably be overlooked. Thus, people start to learn patterns in the field of text data to make full use of it. Text categorization or classification is the task of assigning one or more classes to a document based on its content. This could be done manually or algorithmically. For example, there are many news reports about weather, politics, entertainment, etc. Various of methodologies and algorithms are implemented to classify documents into each category, which is efficient and time saving. Presently, researchers use different classifiers to improve the precision and efficiency, such as deep neural networks [Prusa, 2016] and Naïve Bayes methods [Jiang, 2016].

Twitter, one of the most popular social media platform on the Internet, has a large number of active users [Kwak, 2010]. It allows users to send short messages called Tweets that are essentially of 140 characters or less. This sort of data (tweets) can be achieved through Twitter's public Application Programming Interface (API). Tweets are frequently used to share information and post general public events such as sports, personal blogs, natural phenomena, etc. A rare event, for instance, Hurricane Sandy occurred in 2012, impacts the real world as well as produces numerous Twitter's messages. Generally speaking, people's capabilities are beyond finding out that hurricane's path, scale and

calculating how much loss it caused, such as property and life. Path and scale give people an initial evaluation of Hurricane Sandy. What if we want to know the very detail such as the effect on ordinary people? It is hard to achieve it via news because there are only official reports, since reporting each individual problem is unrealistic and costly. For example, "*Trees are knocked over in my back yard.*" would only appear on Twitter. Many informative messages are concealed among the data called hidden information. On the other hand, we can achieve both official reports and hidden information to the specific events via Twitter.

Thus, this thesis uses Hurricane Sandy related Twitter as the background to study patterns and extract relevant and informative data. On account of the large number of raw data that crawled through Twitter's API, there must exist irrelevant and unwanted data. Note that the irrelevant and unwanted data is, specifically, the data that is related to other events' descriptions. Based on our background, irrelevant and unwanted ones are those not related to Hurricane Sandy. Moreover, due to the attributions of tweets, which are simplified, without contexts and full of Internet slang words such as LOL means laugh out loud, people probably could understand these tweets clearly, but machines are different. Considering the quantity of data, it is impossible to do classification manually. Hence, people try to train machines to classify the data automatically.

Presently, researchers set a keyword search method to filter out the irrelevant Twitter's data [Lu, 2016] [Guan, 2014]. This method is straight forward and simple. It can indeed achieve majority of relevant tweets with a high accuracy. Nevertheless, the limitation of using the keyword search method is obvious, i.e., it fails to extract the hidden information that we discussed above. Therefore, this work intends to present a Fuzzy Logic

approach which can well deal with vagueness and ambiguity. Several fuzzy rules are designed to determine if those indirect descriptions are relevant. Note that the indirect description is one forms of the hidden information. Figure 1.1, a real example, is used to explain what the indirect description is. According to President Trump's tweet, he is supposed to sustain his viewpoint of "Muslim ban". However, he depicts other things instead of talking about Muslim ban directly.

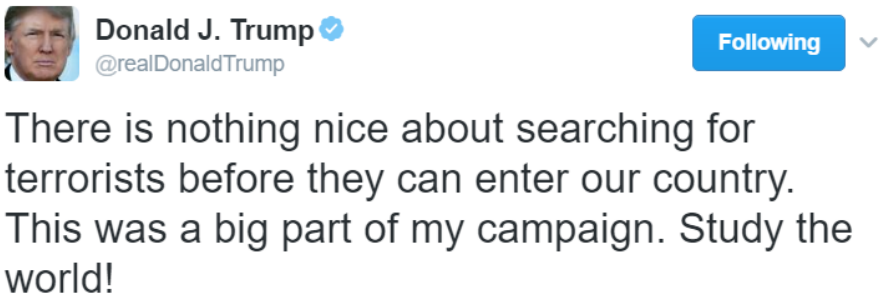


Figure 1.1 President Trump's real tweet on 30th Jan., 2017.
Source: https://twitter.com/realDonaldTrump/status/826044059647107073

Fuzzy logic is capable of handling with the partial truth, where the truth value may range between completely true and completely false [Al-Najjar, 2003]. In Boolean algebra, 1 represents truth and 0 is falsity. For other algorithms and methodologies, such as Naïve Bayes methods and Support Vector Machines, only completely true and completely false are acceptable, while fuzzy logic is able to accept partial truth or partial false. Taking water temperature as an example, traditional Boolean logic could return either cold or hot, while fuzzy logic would return very cold, cold, warm, hot somehow, extremely hot, etc.

The objective of this work is to classify whether the hidden information is relevant to Hurricane Sandy or not. The rest of this thesis is organized as follows. Chapter 2 presents

related works including social media, text classification and fuzzy logic. Chapter 3 describes the details of fuzzy logic-based approach and text processing in social media data. The process of design is discussed. Chapter 4 is experiments and evaluation for the proposed methods. The dataset, experiment design, evaluation method and experimental results are specified in this section. Chapter 5 concludes the whole thesis and indicates the future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Social Media

Social media platforms like Facebook, YouTube and Twitter have greatly improved our lives. These platforms provide users a chance to gather and share messages, opinions and experiences. The reason why they are "social" is that they focus on communication and participation. They can be easily accessed via computers and smartphones. Presently, these platforms are still gaining popularity among people, which attract researchers to learn and exploit the features of the social media in order to make use of it. Social media differs from newspaper or traditional media such as TV broadcasting. Traditional media operates under a model that one source deliver information to many receivers, such as a TV program which is broadcast to numerous users. In contrast to traditional media, social media is more flexible.

Social media can support information exchanges before, during and after an emergency in several ways. In study [Knuth, 2016], researchers take advantage of social media's real-time character and multi-media integration to study emergency. They regard social media as practical emergency communication tools. However, in the large quantity of data, unconfirmed and unreliable messages limit its usefulness. Thus, researchers obtain data in different phases of emergencies as well as in different media. They recruit participants to complete a survey which is the ground truth for their experiments. The main question in such a survey is what a participant's next move is during and after the emergency. Based on the analysis of survey data, researchers can learn people's behavior patterns in the answers, and then apply them into the real world. The results show that

social media can help establish connections to the potential helpers in the preparedness phase, and warn about upcoming emergencies such as stormy weather or heavy traffic in advance. Before an emergency happens, social media can be used to reduce the negative consequences of an event or a disaster.

The studies in [Caragea, 2014] [Lu, 2016] [Shekhar, 2015] [Guan, 2014] [Dong, 2013] [Spielhofer, 2016] pay attention to rare events and natural disasters which not only effect world significantly, but also influence social media.

Caragea *et al*. [2014] study how Hurricane Sandy influences people's moods with the consideration of distance between the disaster and people. They propose a sentiment classification method for user posts in Twitter during the Hurricane Sandy. According to Starbird *et al.* [2012], social media data can be identified as coming from bystanders to direct responders. Analyzing social media data can help people understand people's concerns, panics and emotional impacts among users. Caragea *et al.* [2014] take this problem as a classification problem and use supervised machine learning approaches to classify a tweet into a positive, negative or neutral one. After applying stemming and lemmatization method, and stop words and punctuation removal, they use frequency of words as features. The sentiment features are calculated with the SentiStrength algorithm that returns values between -5 to 5. [-5, -1], [-1, 1] and [1, 5] represent negative, neutral and positive attitudes, respectively. They manually label 602 tweets in total as positive, negative and neutral, which is the ground truth for the experiments. After the data pre-processing, the Support Vector Machine (SVM) and Naïve Bayes are designed to proceed to the classification. The best accuracy is 75.91%, which is calculated with SVM.

One challenging problem is whether information extraction is effective and accurate. It is accepted that the quantity of social media data is vast. Because of this, researchers make much effort to find valuable information and try to avoid and filter out unwanted data. The works [Lu, 2016] [Guan, 2014] [Dong, 2013] performed by focusing on natural disasters. Hurricane Sandy in 2012 did produce countless data on social media, which has draw many researchers' attentions to study and find patterns between the Twitter activities and hurricane. All these works use a Keyword Search method to extract relevant data by setting various keyword lists. The advantages of Keyword Search method are straight forward and effective. The dataset used in [Lu, 2016] [Guan, 2014] [Dong, 2013] is crawled through Twitter's API, which contains more than 9 million unlabeled records. After filtering time and geo-coordinates and applying keyword search method, the researchers are able to obtain the relevant information.

The drawbacks of the keyword search method are also obvious. This method cannot extract the information that beyond its keyword list. Namely, the words in the list are the limitation of this method. However, the emphases in these works is not the information extraction. They just want to obtain the data in the specific period, and make use of it to proceed to other researches. The work [Lu, 2016] proposes a data processing method based on a k-means clustering algorithm and analyze the evolution of a rare event via social media. The research [Guan, 2014] develops a new metric, disaster-related ratio (DRR), which simply illustrates the relationship between a disaster and social media activities. DRR calculates the ratio of the number of relevant to that of irrelevant ones at a same time span in the same area. For example, a high DRR means that the disaster is occurring. The work [Dong, 2013] performs statistical analysis to find the causality correlation between

an approaching hurricane and the response of the public. They propose a system that can automate the process of extracting, analyzing and visualizing social media data. They define two lists of keywords to identify two behaviors. For example, the tweets containing "evacuate" and "leave" indicate that users intend to evacuate; the tweets containing "stay", "prepare" and "power" indicate that users intend to stay. After the noisy data eliminated, the Latent Semantic Indexing (LSI) algorithm is designed to produce a topic model of the Hurricane Sandy data. However, its drawback of LSI is its huge computation [Zhang, 2011].

In addition, some researchers try to use other methods to extract relevant data. Korolov *et al*. [2016] study the possibility of predicting a social protest based on social media data. They are aware that the hashtag is also worthy to utilize. They point out that only about 1/3 of tweets contains hashtag, while the hashtag usually indicates the topic directly. For those tweets without hashtags, an algorithm, cosine similarity, is used to calculate the similarity between tweets for clustering. Besides, the keyword search method is still used in their work. After combing these three methods, they train an SVM classifier by using 6521 manually labeled tweets for extracting relevant data. Alsaedi *et al*. [2016] use an idea similar to the study [Korolov, 2016]. Their novel point is that they put more emphases on online stream data for event detection. They combine temporal, spatial and textual features in their system to detect disruptive events in a given place. Additionally, in the phase of data pre-processing, they remove stop words; calculate the term frequency-inverse document frequency (TF-IDF) and term frequency. The keyword search method, geographic tags and hashtag are also used for a performance guarantee.

Moreover, Spielhofer *et al*. [2016] propose that the problems of irrelevant data removal and noise reduction are similar to the email spam filtering. They manually label 1000 tweets in “spam” or “not spam” for training and testing. The Naïve Bayes classifier is trained for relevant tweet detection.

In the meanwhile, social media benefits human resources management when recruiting. Sewwandi *et al*. [2017] provide a web application in order to detect an individual’s personality by using social media data. They realize this method by considering three features: ontology based personality detection, personality detection through linguistic analysis, and questionnaire based personality detection. The first feature is calculated with protégé OWL; the second one extracts some basic characteristics such as the percentage of personal pronouns in the textual data; the third one is prepared for those people who do not register Facebook which is the source of the data. The Trait classification algorithm is designed to classify three kinds of personalities which are extrovert, neuroticism and psychoticism. The final accuracy is remarkable, which reaches 91% when it tests against a real world questionnaire.

2.2 Text Classification and Feature Selection

2.2.1 Text Classification

The increasing availability of text documents in digital forms draws researchers’ attentions as well as interests in developing automatic analysis methods for them. Text classification or categorization is introduced for this purpose. Text classification aims to assign text documents into predefined categorizes or classes, such as news and academic topics. Recently, researchers propose many algorithms in the field of text classification, like a

Naïve Bayes-based approach introduced in the study [Jiang, 2016], Convolutional Neural Networks (CNN) used in the study [Prusa, 2016], and other algorithms combined with different classifiers. Researchers make every effort to improve the performance of text classification.

Different from those well-structured data such as excel documents, text data contains more semi-structured or even unstructured data such as tweets and text messages. Generally, text data is a kind of high-dimension data if we treat each word in a document as a feature or dimension [Zhang, 2005]. Thus, researchers propose algorithms, such as TF-IDF, LSI and bag-of-word, to convert text data into the forms of mathematics for better processing and analysis. In the meanwhile, feature selection gains its importance in this domain. The job of feature selection is to reduce the number of features or dimensions and finding a subset of the original data to improve the performance and decrease the computation, which is appropriate to be applied in the domain of text classification. Feature selection can be used with other robust classifiers, such as Genetic Algorithm (GA) combined with SVM [Bidi, 2016].

Zhang *et al*. [2011] compare the algorithms among TF-IDF, LSI and multi-word in information retrieval and text classification. Because these three methodologies have been introduced for a long time, but there is no comparison of performance among them. Thus, they study the effectiveness of different representation methods. Indexing and weighting are two tasks in the domain of text representation. Indexing is to assign the indexing term to documents. Weighting is to assign a weight to each term, which measures the importance of a term in a document. They claim that TF-IDF and LSI have a common drawback, which is the huge computation. For instance, the size of the text documents decides the number

of features and dimensions with which TF-IDF will compute. Normally, the text data is large, which causes the huge computation. A multi-word algorithm consists of two methods: statistical and linguistic methods.

TF-IDF is commonly used to weight each word in a text document. It can obtain the feature matrix or called training dataset feature. The more often a word appears in a document, the more important it is; the more text in which the word is, the less discrimination it is. Because of the limitation of TF-IDF mentioned above, some researchers introduce an improved TF-IDF for text classification [Zhang, 2005]. In order to reduce the computation, namely, the matrix dimension, the researchers use stemming technique and synonyms. This process can be regarded as a data pre-processing step, which improves the performance of the entire algorithm.

Regardless of TF-IDF techniques, Naïve Bayes algorithm is a popular method in text classification because of its simplicity and effectiveness. Jiang *et al*. [2016] introduce an improved Naïve Bayes method called deep feature weighting Naïve Bayes (DFWNB). It is a combination of TF-IDF and an improved Naïve Bayes method. The weight of each feature is computed by TF-IDF. In the Naïve Bayes method, the novel point is that they use Maximum Likelihood Estimate (MLE) to calculate the prior probability and conditional probability.

Furthermore, Yin *et al*. [2015] claim that people put little emphasis on the short text analysis. Short text is widely used in micro-blog and short reviews. They propose a semi-supervised learning method and train an SVM classifier to improve the traditional methods such that it can be applied in the field of big data. Their main work focuses on data pre-processing and semi-supervised learning. In the step of data pre-processing, they use an

existing invalid data dictionary information Z to fuzzy match the word and remove the useless and non-informative data. During the step of semi-supervised learning, they first train an SVM classifier using labeled samples. Then the trained SVM classifier is used to label those samples without label. Iteratively, all the samples are labeled.

2.2.2 Feature Selection

Due to the text data's high feature dimension, a feature selection technique is applied to remove irrelevant data and reduce the dimension for improving performance. The flowchart of text classification is shown in Figure 2.1 [Shah, 2016].

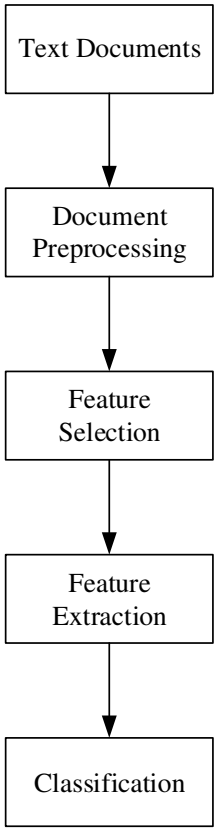


Figure 2.1 The flow of text classification.

According to the work [Shah, 2016], researchers apply different algorithms in each step. Stop word removal and stemming technique are used in document pre-processing; Document Frequency, Gini Index, Chi-square Statistic and Information Gain are applied in feature selection; feature extraction contains Principal Component Analysis (PCA) and LSI; classification can be done with a Decision Tree (DT) classifier, K Nearest Neighbor (KNN), Naïve Bayes and SVM. Shah *et al*. [2016] also make a comparison among algorithms in each step as shown in Tables 2.1-2.3.

Table 2.1 Comparison of Feature Selection Methods

Method	Advantage	Disadvantage
Document Frequency	Simplest method with lower cost in computation	Assuming that the rare terms are non-informative, it contradicts with a principle of IR, where rare terms are more informative
Gini Index	Select the features efficiently	Selecting a large number of features while eliminating redundant features
Chi-square Statistic	The value of this is comparable with the same category terms	Not reliable for low frequency terms
Information Gain	Biased towards multi-valued attributes	Eliminating no redundant features

Table 2.2 Comparison of Feature Extraction Methods

Method	Advantage	Disadvantage
PCA	Simplicity of the technique and robustness in approximating the covariance or correlation matrix	The covariance matrix cannot be evaluated in an accurate manner
LSI	Easy to implement, understand and use	It is a linear model, and cannot handle nonlinear dependencies well

Table 2.3 Comparison of Classifiers

Method	Advantage	Disadvantage
Decision Tree	Easily picking the best feature from the set of data	Overfilling and not so accurate
K Nearest Neighbor	Robust to noisy training data	Classification time long and difficult to find the optimal k
Naïve Bayes	Easy to implement and requiring less training data	Assuming independence of the class and losing accuracy
Support Vector Machine	Robust and very accurate	High algorithmic complexity and extensive memory requirements

Additionally, in [Bidi, 2016], researchers try to introduce GA to perform feature selection, and then combine different classifiers to prove whether GA is qualified to be used in this field. The main problem in handling text data is the feature dimension. Hence, a superior feature selection tool and a robust classifier do matter. However, it does not mean that the best feature selection tool combined with the best classifier is the best solution. It needs to be proved by the experiments and results. Feature selection aims to find the smallest subset of the original data such that it can improve the classification performance and reduce the computation time. In their work, they apply three different popular classifiers, Naïve Bayes, KNN and SVM. The results indicate that the combination of SVM and GA outperforms the others.

Moreover, researchers have used Deep Neural Networks (DNN) to analyze text data, which is a hot topic presently [Prusa, 2016]. They intend to use CNN that is a DNN method to automatically learn features from raw text data. CNN and a new encoding approach are used to reduce memory requirements and training time. Note that CNN is mainly used in image processing. The new encoding method can convert text data into an image form such that CNN can be used. Besides, they apply Bag-of-Words to create word

vectors to describe each sample through word presence or frequency. The drawbacks are loss of information and requirement of high dimensional features.

Feature selection can also be applied to social media applications. Prusa *et al*. [2015] utilize feature selection techniques for tweets sentiment classification. Many studies are conducted to construct sentiment classifiers, but few apply feature selection with sentiment classification. Since the volume of Twitter's data is too large, feature selection is necessary in order to reduce the computation and improve the performance. They choose Chi-squared, Threshold-Based Feature Selection techniques and First-Order Statistic in the feature selection. After combining with other classifiers, such as KNN and Logistic Regression, they use cross-validation to verify their proposed methods.

Besides, stop word removal is a crucial process during data pre-processing. However, removing all stop words may cause the loss of information, which further influences the accuracy and performance. The work [Kasun, 2015] proposes a novel method to reserve some stop words to improve the performance. Differing from the traditional methods, in the process of removing stop words, their proposed method is to split one sample into two groups of words. One group aims to process stop words only, the other is to analyze the remaining words without stop words. After that, GA is introduced to calculate the subsets, which are in turn used in training the classifiers.

2.3 Fuzzy Logic

Fuzzy logic was introduced a few decades ago. Its foundations is solid, its applications are numerous, and its influence has been widely spread. Different kinds of applications are utilized in cybernetics, linguistic domain and software programming. Kosko [1993] argues that the superiority of fuzzy logic is its similarity with natural language and natural

thinking, which is close to human brains. The main contribution of fuzzy logic is the methodology for computing with words [Zadeh, 1996]. As "computing with words" suggests, it is a methodology that words can be converted into numerical values for computing and reasoning. Zadeh *et al*. [1996] also claims that no other method serves this purpose. The majority of methodologies deal with completely true and completely false cases, while fuzzy logic is able to accept partially true cases, such as 70% true. Fuzzy logic keeps gaining its fame because we face uncertainty and vagueness which traditional methods fail to deal with adequately.

Fuzzy logic stems from simulating human brain, such as uncertainty of judgment, the way of reasoning and thinking, description of uncertain or unknown systems and nonlinear objective functions. Through simulating the thinking way of human being, fuzzy logic can be used to implement fuzzy comprehensive judgment. In addition, it is good at expressing knowledge and experience with unclear boundaries. It takes advantage of the concept of membership functions to distinguish fuzzy sets and deal with fuzzy relations, in order to solve the linguistic problems.

2.3.1 Fuzzy Logic in Linguistics

Natural language processing is a field of computational linguistics with interactions between machines and human natural languages. It is common knowledge that many words and sentences in natural language lead to descriptions with fuzzy mathematics, such as fuzzy set, fuzzy relations and fuzzy logic.

Sun *et al*. [2002] propose to create a fuzzy logic system that can learn semantic relations among the concepts represented with words from a linguistic corpus. Further using such relations to process the word sequences generated by speech recognition

systems aims to improve the system. This method is designed to predict those words that speech recognition systems fail to recognize. It is able to boost the performance of speech recognition systems. Generally, words and their meaning are often vague and in a fuzzy relation. They design their fuzzy logic system to accomplish two major tasks. One is to evaluate whether a word belongs to certain corpus based on sentiment. The other is to predict the missing words in a given corpus based on some fuzzy rules. Thus, they have two sets of rules for evaluation and prediction. Besides, they aim to find the core word, e.g., the main verb in a verb phrase.

A fuzzy logic-based approach can be used for sentiment classification. Sathe *et al*. [2017] propose a novel method for performing sentiment classification by using Neural Network (NN) and fuzzy logic theory. The Neural Network can deal with numeric and quantitative information whereas a fuzzy logic system is able to handle symbolic and vague information. Their datasets and reviews contain imprecise information, ambiguity and vagueness, which can be well processed by the fuzzy logic system. They fuzzify the input reviews by using a Gaussian membership function and build a fuzzification matrix. In this matrix, every element represents the possibility of one certain sample belonging to one certain class. This matrix is transposed to a Multilayer Perceptron Backpropagation Network, a subclass of NN. Then, it takes this matrix as an input and modifies the weights for better accuracy.

Text summarization and intelligent tagging utilize the fuzzy logic as well. The study [Suanmali, 2009] uses fuzzy logic to deal with the problem of differentiation between the important and unimportant features. During the preprocessing process, sentence segmentation, tokenization, stop words removal and stemming are used. Nine features are

used to score each sample. In the meanwhile, the input membership function for each feature is divided into several membership functions which are composed of values. In [Damaševičius, 2016], four fuzzy logic-based models are proposed for tag recommendation. They aim to assign tags to texts automatically by using fuzzy logic-based models. They design four fuzzy logic-based models with 5, 7, 9 and 12 features, respectively. More features make rules comprehensive and thus improve the performance of the system.

Internet of Things (IoT) is a hot topic today. IoT is a concept to connect every object with other objects and make them communicate. It needs to respond according to the inputs given by the users and produce suitable and reasonable actions that are convenient to users. A fuzzy logic-based approach can be combined with IoT. For example, Patel *et a*l. [2016] propose the architecture of context awareness in the middleware portion of the IoT systems, where the fuzzy logic is applied. The object connected in an IoT region is distributed hierarchically. The fuzzy logic is used in the process where upper objects transform data to lower objects. For instances, the sprinkler operates automatically based on soil moisture and environment sensors, and the air condition will be started based on outside temperature and room temperature. This method mainly deals with the linguistic information and imprecise data, which further boosts the system performance.

2.3.2 Fuzzy Logic in Cybernetics

The power of fuzzy logic also draws some researchers' attentions in the field of cybernetics. In the domain of robotics, robot control must be subject to certain constraints, such as uncertain position and velocity. Fuzzy logic control can deal with those conditions under which the parameters are dynamic or uncertain. Thus, a fuzzy control scheme can be

developed for nonlinear systems according to unknown changes of the environment [Sun, 2016].

The studies [Kulkarni, 2007] [Quiros, 2016] introduce fuzzy logic into traffic light controller design and traffic condition analysis. Kulkarni *et al*. [2007] suggests that the conventional traffic light changes at a constant cycle time. Some methods are proposed to predetermine the patterns of light in a certain period of a day according to the historical data. However, this is not the optimal solution. Hence, they apply fuzzy logic to improve this system by using many empirical protocols and rules. For instance, the extension or termination of a green light depends on the number of vehicles approaching the intersection and the number of vehicles in the queue. In [Kulkarni, 2007], vehicle detectors that are installed on different positions are used to obtain real-time traffic condition data. Fuzzy rules are used to evaluate how suitable it is for an extension green light, such as "if approaching vehicles are few, then extension is zero". Quiros *et al*. [2016] propose a fuzzy logic-based method to evaluate the traffic state of a road. The IP cameras installed at different roads are used to obtain data, which can obtain the location and size of a vehicle. Therefore, the density of vehicles, the distance between two neighboring vehicles and sizes can be computed, and then all the information are transposed to the fuzzy logic system. They have designed 27 fuzzy rules, which can identify six kinds of traffic conditions.

In [Salah, 2016], fuzzy logic is used in model order reduction so as to improve the accuracy while preserving the basic properties of the original model. In the domain of model order reduction, many methodologies have been introduced, but the researchers claim that nobody try fuzzy logic presently. They conclude that the Gaussian membership function is more suitable for this proposed method.

An intelligent drilling system can be commercially profitable in terms of reduction in crude material and labor. Zulkifli *et al*. [2016] design a fuzzy logic controller (FLC) to precisely select cutting parameter for the drilling operations. FLC allows the machining parameters such as spindle speed to be automatically selected under different conditions, which can reduce the risk of spoiling the drill bit, and then increase productivity. For an existing drilling system, there are only two parameters whereas FLC can deal with five inputs such as material hardness, tool hardness, depth of drill, hole diameter and cutting fluid flow rate. Each input is assigned into either three or five membership functions. By designing fuzzy rules, Zulkifli *et al*. [2016] can compute the spindle speed and feed rate. By using Matlab and Simulink, they prove that the proposed FLC can successfully improve the performance.

CHAPTER 3

PROPOSED METHODOLOGY

This work aims to introduce a fuzzy logic-based text classification method during the evolution of a rare event. After data collection, data preprocessing and analysis follows. Then, the proposed model is used to decide whether the data items (tweets) are relevant or not. Figure 3.1 shows a framework of this work.

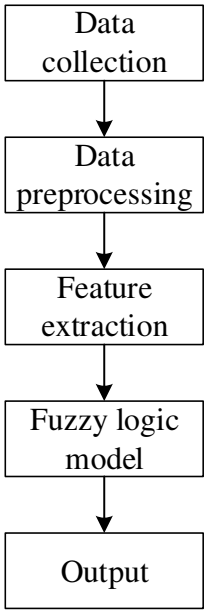


Figure 3.1 The framework of the proposed method.

This section is divided into three parts, data preprocessing and analysis, feature extraction and fuzzy logic model design. The first part explains in details how the dataset is preprocessed and analyzed. The second part is to introduce in details how we extract useful features in each tweet for text classification with a fuzzy logic model. The third part will present how we design the fuzzy logic model.

3.1 Data Resource

3.1.1 Manually Labeled Data and Data Analysis

We try to use a labeled text dataset to build a model. Thus, we request volunteers to manually label a set of randomly selected tweets from the original dataset as a training dataset, which is treated as the ground truth data in the proposed study. They grade each tweet using 0 or 1to indicate irrelevance/don't know (DK) or relevance, respectively.

For the jth tweet, there are N volunteers to grade it. We define C_{ij} for the jth tweet scored by the ith volunteer, where $i \in \{1,\ldots,N\}$. Based on this, the summation score of the jth tweet O_j is defined as:

$$O_j = \sum_{i=1}^{N} C_{ij}, C_{ij} \in \{0,1\}. \tag{3.1}$$

We predefine four score intervals $D_1 \in [0, N/4)$, $D_2 \in [N/4, N/2)$, $D_3 \in [N/2, 3N/4)$ and $D_4 \in [3N/4, N]$ to represent a relevant degree to a certain tweet, which stand for irrelevance/DK, low relevance, moderate relevance and high relevance, respectively. To be more specifically, the definitions are shown as follows:

(1) if $O_j \in D_1$, it means the jth tweet is irrelevant/DK to a rare event.

(2) if $O_j \in D_2$, it means the jth tweet is lowly relevant to a rare event.

(3) if $O_j \in D_3$, it means the jth tweet is moderate relevant to a rare event.

(4) if $O_j \in D_4$, it means the jth tweet is highly relevant to a rare event.

3.1.2 Data Processing

Data preprocessing is a vital stage for further analysis and processing. Especially, this is an inevitable step in the domain of text analysis. It reduces the risk of producing errors and increase system performance, because text data usually contains many useless and meaningless words. In this work, we remove stop word and URL (the address of a World Wide Web page), and adopt a lower-case scheme i.e., converting all capital letters into lower-case ones.

According to the route of Hurricane Sandy, irrelevant geographic coordinates are filtered out at the very beginning. By reviewing tweets, we find out that tweets are full of noise such as URL. It is a common knowledge that URL has a fixed format which starts with “http://”. Hence, we apply a pattern match algorithm to filter text [Sheshasayee, 2015]. When pattern, “http://”, is found, the program automatically removes the information following it.

Stop words are the most commonly used words in a language. In English, for example, “the”, “this” and “that” appear frequently in the sentences, which are seldom used for analysis even though they are essentially for some sentences to be grammatically correct. In our method, we split each sentence into words, and then compare each word with a list of stop words. The word is removed if it belongs to this list. The lower-cased scheme allows one to treat two words with an identical sequence of letters but some different case letters as a same word, e.g., “HURRICANE” and “Hurricane” are treated as one same word after decapitalizing every letter in both.

3.2 Feature Extraction

In order to use a fuzzy logic-based method, it is necessary to quantify the sentences as few numerical features that are convenient to calculate. These features are attributes that tend to represent the data used for the proposed fuzzy logic model. Seven features from each tweet are designed.

At the beginning, the training dataset is worthy to be exploited. After the above processes are done, the tweets in the training dataset are ready for analysis. Note that some words appear more frequently in a specific topic or context than others. For example, in the airport, such words as "time", "arrival" and "airline" are more frequently broadcast than "bids" and "price" in an auction case. Following this idea, we select the tweets belonging to D_2, D_3 and D_4 from the training dataset to obtain the most frequently-used words. For each most frequently-used word i, word importance α is defined as:

$$\alpha_i = \frac{P_i}{Q_i} \times 100\% \tag{3.2}$$

where P_i decides the number of word i appearing in those tweets which belong to D_2, D_3 and D_4; Q_i decides the number of word i appearing in all the tweets, i.e., belonging to D_1, D_2, D_3 and D_4; α_i is a percentage that indicates how important word i is. Generally, the larger α_i, the more important word i. The next step is to sort all the most frequently-used words according to its α_i from the largest to smallest, select top l words to build a key list L, and then split L into three evenly distributed subsets denoted as L_1, L_2 and L_3 with different relevant weights θ_1, θ_2 and θ_3, respectively, where $L=L_1 \cup L_2 \cup L_3$.

L is our ground truth data for the following computation. Therefore, in order to calculate the similarity between a word in a tweet and a word in the key list L, we introduce a similarity function in [Natural Language Toolkit]. We define a similarity evaluation process as a mathematical operator: $\otimes$.

For the jth tweet that has n words, T_i denotes the ith word in this tweet, where $i \in \{1,\ldots,n\}$. In addition, W_k denotes the kth word in the key list L, where $k \in \{1,\ldots,l\}$. We choose the highest one among the similarity scores to represent T_i 's score. Thus, the highest similarity score S_i of T_i is defined as follows:

$$S_i = \max_{1 \le k \le l} \left(\omega_k \times T_i \otimes W_k \right), i \in [1, n] \tag{3.3}$$

where

$$\omega_k = \begin{cases} \theta_1 & if\ k \in \left[1, l/3\right) \\ \theta_2 & if\ k \in \left[l/3, 2l/3\right) \\ \theta_3 & otherwise \end{cases} .$$

S_i is our basic value because the following seven features are based on it. These features are extracted from each tweet and to be used in the proposed fuzzy logic-based model. The details and definitions are shown as follows.

(a) The highest word score in the jth tweet (H_w)

$$H_j = \max_{1 \le i \le n} S_i \tag{3.4}$$

where H_w decides the highest word score in the jth tweet.

(b) The score of the jth tweet (F_j)

$$F_j = \sum_{i=1}^{n} S_i \tag{35}$$

where F_j decides a tweet's score accumulated by each word's score.

(c) The length of the jth tweet (M_j)

$$M_j = n \tag{3.6}$$

where n is the total number of words in the jth tweet.

(d) The number of frequently-used words in the jth tweet (I_j)

Algorithm 3.1 The procedure designed to compute I_j

Input:
T_i, the ith word in the jth tweet
L, the key list L

Output:
I_j, the number of high frequently used words in the jth tweet.

Procedure:
1: For each word T_i in the jth tweet
2: For each word W_k in L
3: If $T_i == W_k$
4: I_j += 1
5: End
6: End
7: End
8: Return I_j

As computed via Algorithm 3.1, I_j is a feature that indicates how many words in the jth tweet are the same to those words in the key list L. L consists of the most frequently-used words in our ground truth data. It is reasonable that we use L to compare with all tweets.

(e) The weight of the jth tweet (G_j)

$$G_j = \frac{F_j}{M_j} \tag{3.7}$$

where G_j is the mean score of the words in the jth tweet.

(f) The weight of frequently-used words in the jth tweet (E_j)

$$E_j = \frac{I_j}{M_j} \tag{3.8}$$

where E_j decides how many proportion of the frequently-used words in all the words is in a tweet. The larger E_j, the more informative a tweet contains.

(g) The number of patterns in the jth tweet (V_j)

By reviewing the key list L and training dataset, we find that some combinations of words are useful while they may usually be ignored when they are alone. For example, “no power”, “power off” and “no school” are more useful than only single word like “no” or “power” alone. Thus, V_j indicates the number of those patterns in a tweet. The procedure is shown in Algorithm 3.2.

Algorithm 3.2 The procedure designed to compute V_j

Input:
T_i, the ith word in the jth tweet
L_1, the subset of key list L
L_2, the subset of key list L
L_3, the subset of key list L

Output:
V_j, the number patterns used in the jth tweet

Procedure:

```
For each T_i in the jth tweet
    If T_i in L_1
        Count1 += 1
    End
    If T_i in L_2
        Count2 += 1
    End
    If T_i in L_3
        Count3 += 1
    End
    V_j = 0
    For i = 1:3
    For j = i:3
        If i == j
            Continues
        Else
            Sum = counti + countj
            If sum > V_j
                V_j = sum
            End
        End
    End
End
Return V_j
```

3.3 Fuzzy Logic-based Model Design

The framework of using the proposed model is shown in Figure 3.2.

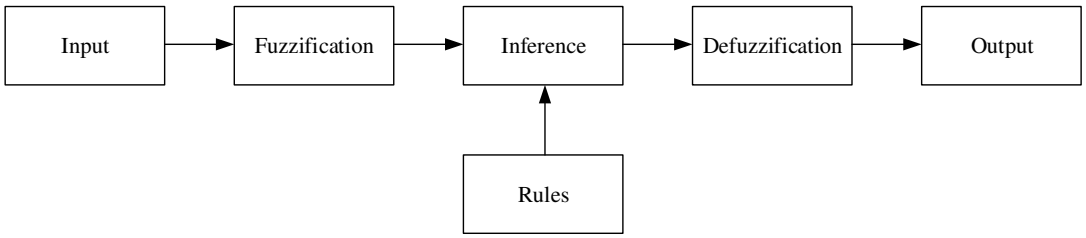


Figure 3.2 The framework of using a fuzzy logic-based model.

(1) Input: We extract seven features discussed above from each tweet as inputs for the fuzzy logic model.

(2) Fuzzification: Fuzzification is the process where the crisp or real inputs are mapped to fuzzy sets by using membership functions. A membership function is designed for fuzzy inputs, and there are several membership functions available, such as trapezoidal-shaped, generalized bell-shaped, triangular-shaped, Gaussian curve and sigmoidal membership function. The details of each membership function are defined as follows.

(i) Trapezoidal-shaped membership function

$$f(x,a,b,c,d)=\begin{cases} 0, & x\le a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ 1, & b\le x\le c \\ \dfrac{d-x}{d-c}, & c\le x\le d \\ 0, & d\le x \end{cases} \tag{3.9}$$

where x is the input, and a,b,c and d are four scalar parameters as shown in Figure 3.3. The scalar parameters a,b,c and d decide the size of trapezoidal-shaped area.

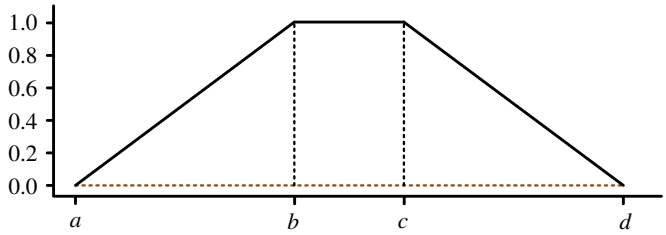


Figure 3.3 An example of trapezoidal-shaped membership function.

(ii) Generalized bell-shaped membership function

$$f(x,a,b,c)=\frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \tag{3.10}$$

where the parameter c is located at the middle of the curve as shown in Figure 3.4.

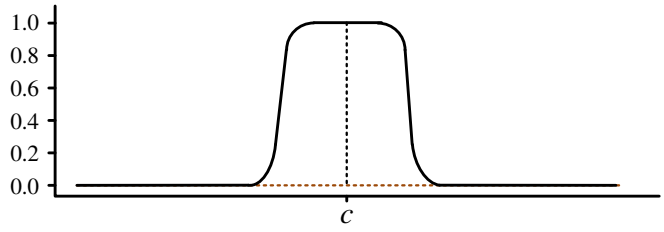


Figure 3.4 An example of generalized bell-shaped membership function.

(iii) Triangular-shaped membership function

$$f(x,a,b,c)=\begin{cases} 0, & x\le a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ \dfrac{c-x}{c-b}, & b\le x\le c \\ 0, & c\le x \end{cases} \tag{3.11}$$

where the parameter *b* decides the location of the maximum value. An example shape is shown in Figure 3.5.

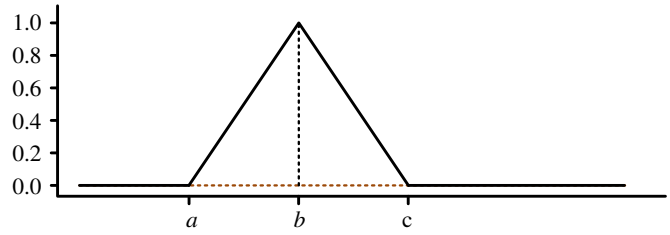


Figure 3.5 An example of triangular-shaped membership function.

(iv) Gaussian curve membership function

$$f(x,\sigma,\mu) = e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.12}$$

where *x* is the input, μ is mean or expectation of the distribution and σ is standard deviation.

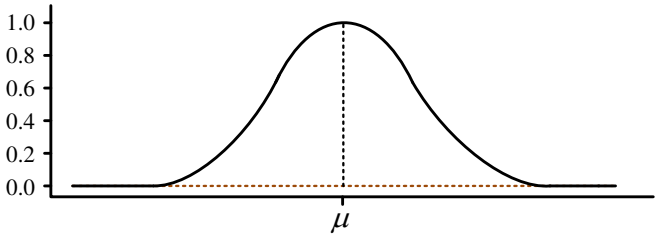
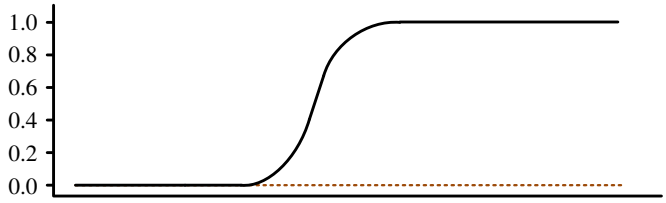


Figure 3.6 An example of Gaussian curve membership function.

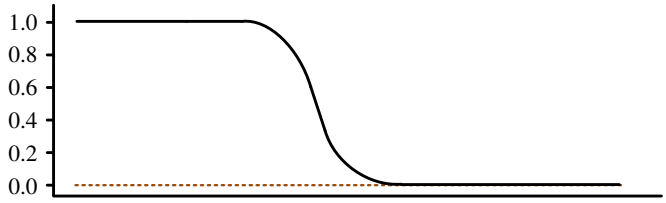
(v) Sigmoidal membership function

$$f(x,a,c) = \frac{1}{1+e^{-a(x-c)}} \tag{3.13}$$

where the parameter *a* decides whether the curve is open to right or left. If $a \geq 0$, the sigmoidal membership function is open to right as shown in Figure 3.7 (a). Otherwise, the sigmoidal membership function is open to left as shown in Figure 3.7 (b).



(a) The Sigmoidal membership function is open to right.



(b) The Sigmoidal membership function is open to left.

Figure 3.7 Examples of a sigmoidal membership function.

In the fuzzification process, we select trapezoidal-shaped membership function because it is simple and commonly used. The input membership function for each feature is divided into several membership functions that are decided by empirical values. For example, we use the trapezoidal-shaped membership function to divide the feature of the highest word score H_{w_j} into five degrees which are composed of very low (VL), low (L), moderate (M), high (H) and very high (VH).

(3) Inference: It is the process of simulating and evaluating a human decision based on a fuzzy logic concept. It is the mapping from a given input to an output. Normally, it uses IF-THEN fuzzy rules to convert the fuzzy input to the fuzzy output.

(4) Rules: Rules are a set of linguistic statements based on IF-THEN statements, which normally follow human expert knowledge or empirical rules. It is easy for us to interpret

the rules as they are similar to our natural language. The details of rules are to be shown in Chapter 4.

(5) Defuzzification: Defuzzification is a method that produces a number that best represents, and consistently represents the fuzzy set. The input for the defuzzification process is an aggregate fuzzy set and the output is a single value. The aggregate of a fuzzy set contains a range of output values, thus it needs to be defuzzified in order to obtain a single output value from the fuzzy set. There are many defuzzification methods proposed in other literature, such as centroid, bisector, mean of maximum (MOM), smallest of maximum (SOM) and largest of maximum (LOM) defuzzification methods [Ma, 2000]. For this work, we compare all these defuzzification methods. Because we cannot conclude which one is the best for our data before the experiments. The definitions and details of defuzzification methods are discussed next.

(i) Centroid: Centroid is the geometric center of a shape. The centroid defuzzification returns the corresponding x axis value.

(ii) Bisector: The bisector is a vertical line that divides a region into two areas with equal proportion. It is sometimes, but not always coincident with the centroid line.

(iii) SOM, LOM and MOM: SOM is the point along the x axis where the region first reaches the maximum value; The LOM is the point along the x axis where the region first decreases from the maximum value. In addition, MOM is the point at the middle of the smallest and largest of maximum values. Figure 3.8 shows the examples of these methods.

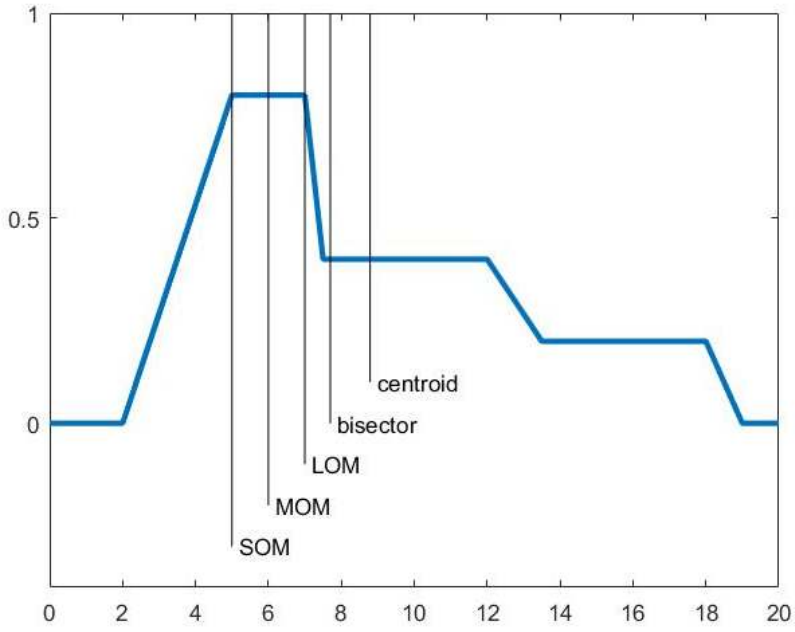


Figure 3.8 Illustration of defuzzification methods.

CHAPTER 4

DATA SOURCE AND EXPERIMENTAL RESULTS

The initial dataset, which is composed of more than 9 million records, is obtained with the help of Twitter's API. For the period from 10.27.2012 00:00:00 to 11.7.2012 24:00:00, each record contains its identifier, timestamp, geographic coordinates and text data. Timestamp and geographic coordinates, i.e., certain dates and places, can be filtered through simple sets of rules, what we need to analyze is the text data called tweets. Different from many other data, this kind type of data does not have any context information. People often have no idea what the tweet actually means without context. For example, if our work's background is not provided, people hardly make any connections with other events like blackout or school cancellation. Consequently, we cannot simply classify tweets into relevant or irrelevant. Therefore, four fuzzy variables, e.g., irrelevant including don't know (DK), lowly relevant, moderately relevant and highly relevant, are designed to describe the degrees of the correlation between tweet and Hurricane Sandy event in our proposed fuzzy logic approach.

4.1 Data Collection, Processing and Analysis

In order to verify the classification result based on the proposed method, 15 volunteers are assembled together to grade 600 randomly selected samples from the original data that are used as ground truth data in this work. For each sample, we adopt two-class classification method, 1 represents relevant and 0 means irrelevant. Then, each sample has a summation score that ranges between 0 to 15. For example, if one tweet's score is 15, which means everyone agrees that this tweet is highly related to the hurricane event. Since each person

has his/her own perspective, the absolute relevant and irrelevant can spilt into four different results that we discussed above. Based on this idea, a total score falling in [0,4) represents irrelevant, [4,7) stands for lowly relevant, [7,11) means moderately relevant and [11,15] is highly relevant.

At the first glance, tweets are not such clean that we can make use of them directly or efficiently. It is not hard to imagine that these tweets are full of Internet slangs and noise such as URL (the address of a World Wide Web page). Hence, after crawling the tweets, the next step is to pre-process them. Any piece of information that we do not want may disturb the outcomes more or less. For instance, "All systems active! #BucksSandy #Sandy (@ Hurricane Bunker) http://t.co/y1U0FlYp" is the original one which is obtained from the Twitter directly and has such inference information mentioned above. Under this circumstance, pattern matching is a suitable and efficient way to handle this problem. Pattern matching, in computer science, is the step of checking a given sequence of expressions for the presence of some patterns. Usually, it is widely used in text mining applications to identify the correct patterns given a large amount of text [Sheshasayee 2015]. After applying pattern matching, a clean version is produced: "All systems active! #BucksSandy #Sandy (@ Hurricane Bunker)".

Moreover, a special part is hidden after the hashtag. It usually provides important information, but not easy to extract them precisely because only about 1/3 of messages have hashtags and the usage is not consistent [Korolov 2016]. For example, in "#sandycantstopme Don't let her stop you. #floodproof #keepgoing", one can find that the words following hashtags are useful while there is no space among words. For human beings, it is easy to understand what this sentence or phrase means, but for the machines,

it is quite difficult and challenging. Since every tweet in this work consists of words, it is reasonable to believe that more useful words in one sentence or tweet lead to a better outcome. Therefore, we must make every effort to mine informative words as many as possible. For this situation, pattern matching should be adopted. But there is another method applied to solve this issue, which is to compare a hashtag's words with English dictionary letter by letter. For example, the word "rain" can be recognized after the fourth operation done in the hashtag "#raintomorrow", because the first, second and third operation can recognize "r", "ra" and "rai", respectively.

The final aspect of data pre-processing is to filter stop words and convert every word into a lower-cased one. The stop words usually refer to the most commonly words in a language. There is no single uniform list of stop words used by all natural language processing tools [Kasun, 2015]. Usually, the stop word is rarely meaningful or useful even though it is essentially for some sentences to be correct or even meaningful grammatically, such as "the", "this", "a" and "on". These words, which appear multiple times in the sentences, do produce redundant and useless information. For instance, when we count the most common words in a corpus without removing the stop words, "the" is definitely among the top ones. Making words lower-cased helps us guarantee a good solution. It can prevent programs from regarding the same word as two different words due to the problem of lower and upper cases, e.g., "storm" and "Storm" are going to be treated as a same word after decapitalizing every capital letter.

4.2 Fuzzy Logic-based Model Design

After the prior mentioned processes are done, the tweets are clean enough to proceed to the next stage of analysis. In the following step, we make effort to extract useful fuzzy values with the consideration of a fuzzy logic approach. Fuzzy logic is capable of dealing with vagueness, which is on the basis of transferring vagueness into fuzzy sets mathematically. A fuzzy set allows for its members to have degrees of membership. Seven parameters discussed in Chapter 3 are implemented in the proposed fuzzy logic-based model. Based on these parameters, the corresponding membership functions are shown in Figure 4.1. The membership function defines the fuzzy set for the possible values underneath of it on the horizontal axis. The vertical axis, on a scale of 0 to 1, provides the corresponding membership values in the fuzzy set. The horizontal axis stands for the domain of four classes. The shape of the used membership functions define the fuzzy set and the decision on which type to use is dependent on our purposes. For this work, trapezoidal-shaped function is utilized owing to its popularity and easy computation.

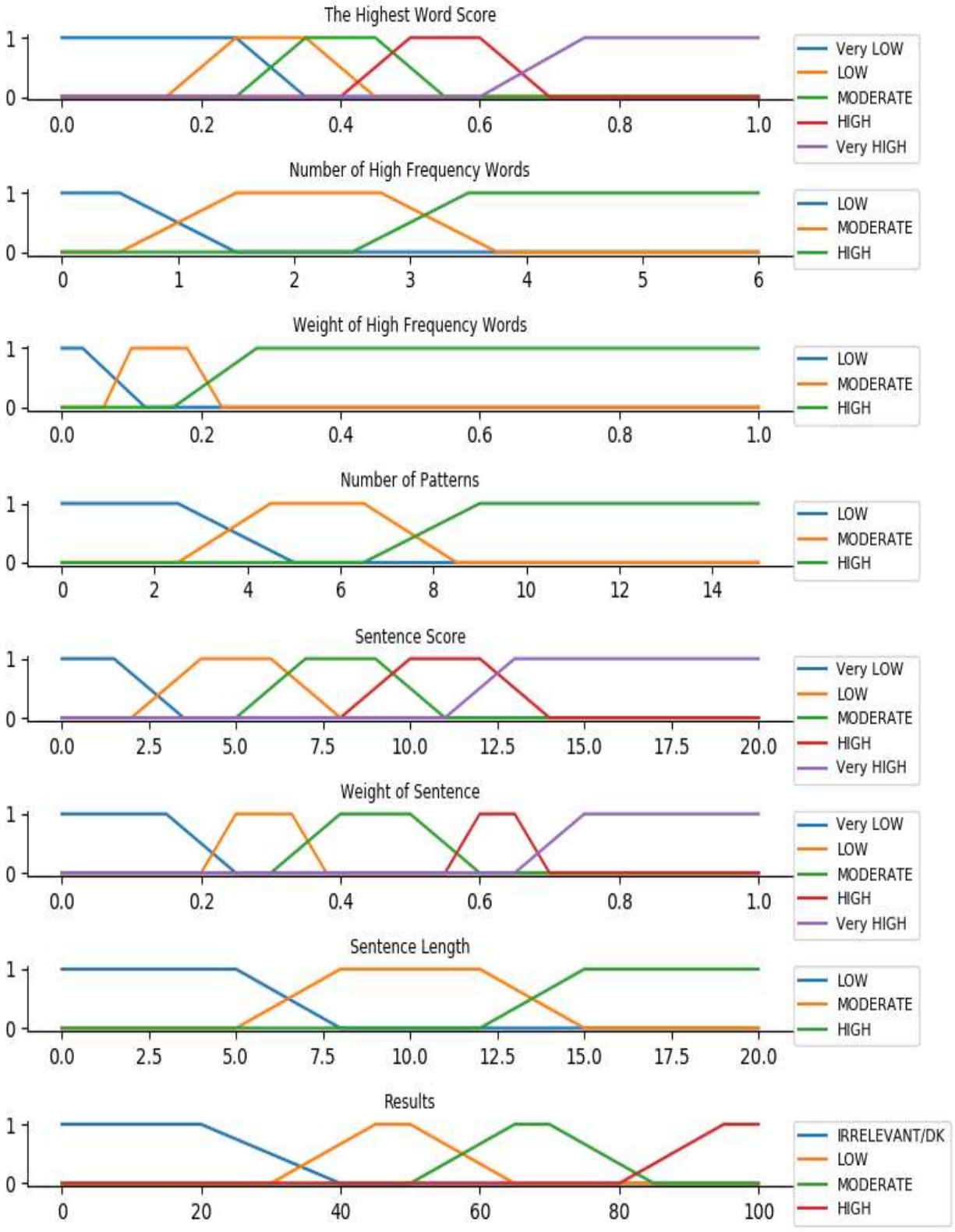


Figure 4.1 Membership functions using trapezoidal shape.

Since the sets of membership functions are prepared, we have to design some rules which are applied to the fuzzy logic system. There are sets of rules representing four corresponding degrees of results. For example, we can regard a sample as highly relevant if its level of the highest word score or the number of high frequency words reaches the demands of the highest domain. The rules are detailed as follows:

Rule 1. *If* $H_j \in [0.4,1.0]$ *and* $I_j \in [2.5,6.0]$, *then* $R \in [75,100]$;

Rule 2. *If* $H_j \in [0.6,1.0]$ *and* $I_j \in [0.5,3.75]$, *then* $R \in [75,100]$;

Rule 3. *If* $H_j \in [0.6,1.0]$ *and* $E_j \in [0.16,1.0]$, *then* $R \in [75,100]$;

Rule 4. *If* $H_j \in [0.6,1.0]$ *and* $N_p \in [2.5,15]$, *then* $R \in [75,100]$;

Rule 5. *If* $E_j \in [0.15,1.0]$ *and* $G_j \in [0.55,1]$, *then* $R \in [75,100]$;

Rule 6. *If* $N_p \in [6.5,15]$ *and* $I_j \in [0.5,3.75]$ *or* $E_j \in [0.15,1.0]$, *then* $R \in [75,100]$;

Rule 7. *If* $N_p \in [2.5,8.5]$ *and* $M_j \in [0,8]$ *or* $E_j \in [0.15,1.0]$, *then* $R \in [75,100]$;

Rule 8. *If* $I_j \in [2.5,6]$ *and* $E_j \in [0.06,0.23]$, *then* $R \in [50,85]$;

Rule 9. *If* $I_j \in [0.5,3.75]$ *and* $G_j \in [0.3,0.6]$, *then* $R \in [50,85]$;

Rule 10. *If* $H_j \in [0.25,0.55]$ *and* $G_j \in [0.3,0.6]$, *then* $R \in [50,85]$;

Rule 11. *If* $H_j \in [0.4,0.7]$ *and* $G_j \in [0.2,0.38]$, *then* $R \in [50,85]$;

Rule 12. *If* $H_j \in [0.15,0.45]$ *and* $G_j \in [0.65,0.1]$, *then* $R \in [50,85]$;

Rule 13. *If* $I_j \in [0.5,3.75]$ *and* $M_j \in [0,8]$, *then* $R \in [50,85]$;

Rule 14. *If* $H_j \in [0.25,0.55]$ *and* $M_j \in [0,8]$, *then* $R \in [50,85]$;

Rule 15. *If* $N_p \in [2.5, 8.5]$ *and* $I_j \in [0.5, 3.75]$ *or* $H_j \in [0.25, 0.55]$, *then* $R \in [50, 85]$;

Rule 16. *If* $H_j \in [0.15, 0.45]$ *and* $I_j \in [0.5, 3.75]$, *then* $R \in [30, 65]$;

Rule 17. *If* $G_j \in [0.2, 0.38]$ *and* $E_j \in [0, 0.23]$, *then* $R \in [30, 65]$;

Rule 18. *If* $I_j \in [0.5, 3.75]$ *and* $E_j \in [0.06, 0.23]$, *then* $R \in [30, 65]$;

Rule 19. *If* $H_j \in [0, 0.35]$ *and* $I_j \in [0, 1.5]$, *then* $R \in [0, 40]$;

Rule 20. *If* $F_j \in [0, 3.5]$ *and* $G_j \in [0, 0.25]$ *and* $I_j \in [0, 1.5]$, *then* $R \in [0, 40]$;

Rule 21. *If* $M_j \in [12, 20]$ *and* $I_j \in [0, 1.5]$ *or* $H_j \in [0, 0.35]$, *then* $R \in [0, 40]$;

Rule 22. *If* $F_j \in [0, 3.5]$ *and* $I_j \in [0, 1.5]$, *then* $R \in [0, 40]$;

Rule 23. *If* $V_j \in [0, 5]$ *and* $I_j \in [0, 1.5]$, *then* $R \in [0, 40]$;

Rule 24. *If* $V_j \in [0, 5]$ *and* $I_j \in [0, 1.5]$ *or* $H_j \in [0, 0.35]$, *then* $R \in [0, 40]$;

Rule 25. *If* $V_j \in [0, 5]$ *and* $M_j \in [5, 20]$, *then* $R \in [0, 40]$;

where

(1) H_j is the highest word's score in a tweet;

(2) E_j is the weight of frequently-used words in a tweet;

(3) I_j is the number of frequently-used words in a tweet;

(4) G_j is the weight of a tweet;

(5) F_j is the score of a tweet;

(6) V_j is the number of patterns in a tweet;

(7) M_j is the length of a tweet;

(8) R is the result.

4.3 Case Studies

Taking a real tweet as an example, "DISASTER UPDATE: NYC #timessquare is deserted. GOOD NIGHT AND BE SAFE!! SO ANXIOUS FOR TONIGHT! #hurricanesandy @ Times Square http://t.co/weAV66o3", the system is able to convert its vagueness and text information into mathematic values corresponding to the seven parameters respectively. Based on this work's background, "DISASTER" is related to the hurricane event; "#hurryupsandy" is the most vital information in this case, because it contains the most critical word, "sandy", which is highly related to the Hurricane Sandy. Therefore, this sample will be definitely ranked by the proposed method. With the activity of each output membership function known, all output membership functions must be combined together. It can be regarded as aggregation. The output memberships are shown in Figure 4.2.

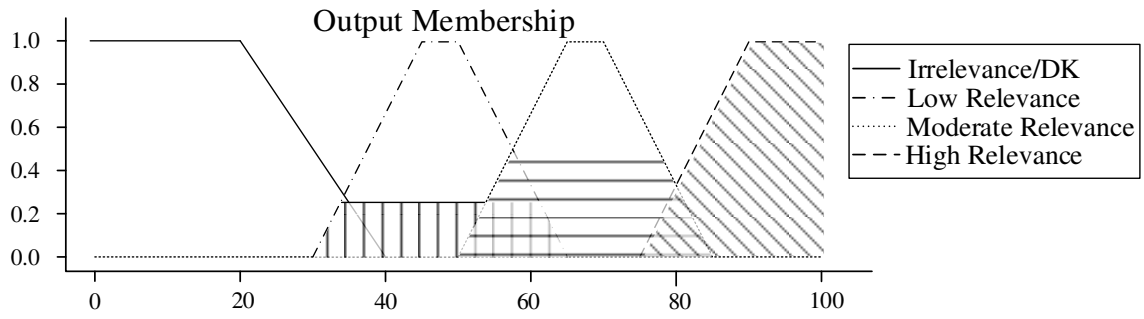


Figure 4.2 An example of output membership.

We can observe that the downward diagonal pattern fully fills the domain of high relevance, the vertical pattern and the horizontal pattern partly fill the domain of low relevance and moderate relevance, respectively. It means that this tweet is completely satisfied with high relevance and partly meets the demands of low relevance and moderate

relevance. By transferring these fuzzy variables into mathematic values, in this case, it has a membership of 1.0 with high relevance, a membership of 0.5 with moderate relevance and a membership of 0.2 with low relevance.

Finally, in order to obtain a real numerical answer, we have to return to crisp logic from the fuzzy membership functions. This step is called defuzzification. For the purposes of this example, the multiple defuzzification methods, i.e., centroid, bisector, MOM, SOM and LOM, are chosen to be used as different ways to conduct the computation. The results of these methods are shown in Figure 4.3. The solid black line indicates the answer.

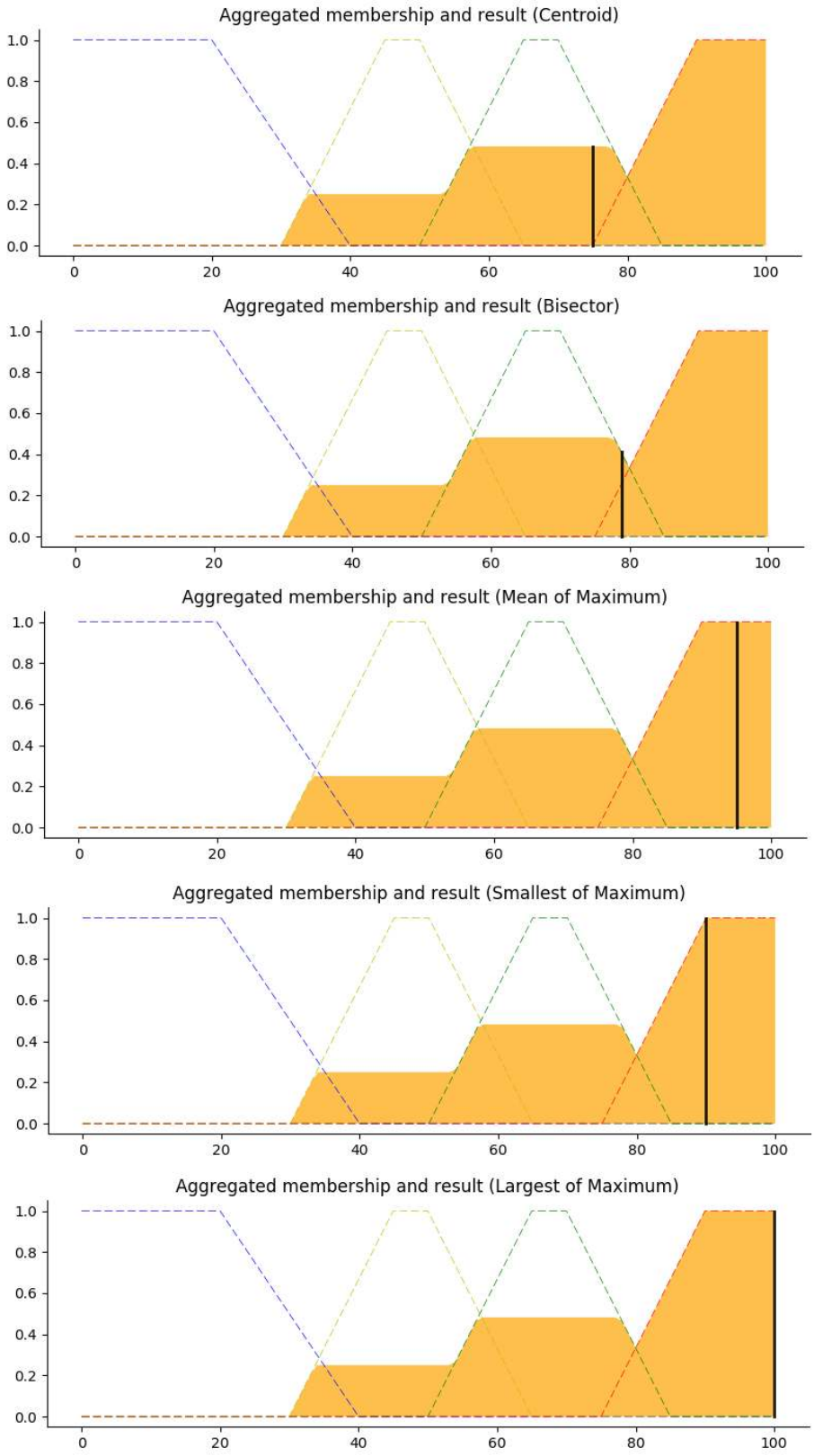


Figure 4.3 Defuzzification methods.

4.4 Experiment of the Proposed Method Results

One challenging problem is to verify the accuracy. Considering an event's background and the type of data, people tend to have different opinions, thereby leading to different answers. In other words, there does not exist a standard or uniform answer given a tweet in this work. Therefore, we have manually labelled a portion of tweets and use that as ground truth data to compare with the automatically computed outcomes.

We purposefully select the data with 300 "relevant" tweets and 300 "irrelevant" tweets from the original dataset. Then 15 volunteers are requested to grade these tweets into four degrees which fit to our proposed fuzzy logic model. After this step, there remains 291 relevant (including low, moderate and high relevance) tweets and 309 irrelevant/DK tweets. In order to widely verify this model, we use them to obtain three sets of testing datasets. Each dataset contains totally 200 randomly selected tweets. The only difference among them is the ratio of relevance to irrelevance, which is 1:1, 1:9 and 9:1, respectively. To be more specific, the first dataset contains 100 relevant tweets and 100 irrelevant tweets, the second one contains 20 relevant tweets and 180 irrelevant tweets and the third one is opposed to the second one. This design includes balanced and imbalanced data items. The class imbalance problem is significant in machine learning where the total number of a class of data (positive) is far less than the total number of another class of data (negative) [Kang, 2016] [He, 2009].

The accuracy compared between computed and manually labelled outcomes are shown in the following tables. Since there are different defuzzification methods, the centroid defuzzification method is taken as an instance for explaining and analyzing.

Table 4.1 shows the accuracy of a polar relevance problem, and Table 4.2 shows the accuracy of a four-degree relevance problem.

Table 4.1 Results of Polar Relevance Problem

Method	Relevance	1st Dataset	2nd Dataset	3rd Dataset
Centroid	Irrelevant/DK	99%	99.4%	100%
	Relevant	97%	90%	95%
Bisector	Irrelevant/DK	98%	99.4%	95%
	Relevant	98%	90%	94.4%
Mean of Maximum	Irrelevant/DK	99%	98.3%	100%
	Relevant	97%	95%	95%
Smallest of Maximum	Irrelevant/DK	99%	99.4%	100%
	Relevant	96%	90%	93.4%
Largest of Maximum	Irrelevant/DK	97%	97.7%	95%
	Relevant	97%	90%	94.4%

From Table 4.1, for the centroid method, the results show that the fuzzy logic-based model performs well with the polar relevance problem. In other words, this model is good at processing two-class classification. Yet, the flaws exist. By reviewing the results, we find out some special tweets which are completely misclassified. For example, "No work. No class. Wtf am I going to do with my day.", this sample is graded to be lowly relevant, and the fuzzy logic-based model misclassifies it into irrelevant. To some extent, people are able to imagine that "No work. No class." is caused by the hurricane. However, there is no direct evidence that the hurricane caused this event. Imagination does not exist in our fuzzy

logic-based model. Since we fail to reconcile this kind of problems, there inevitably exists faulty or inaccurate results.

Table 4.2 Results of Four-Degree Relevance Problem

Method	Relevance	1st Dataset	2nd Dataset	3rd Dataset
Centroid	No	100%	99.4%	100%
	Lowly	57.1%	None	62.5%
	Moderately	62.5%	75%	56.4%
	Highly	100%	100%	97.6%
Bisector	No	98%	99.4%	95%
	Lowly	71.4%	None	62.5%
	Moderately	93.9%	75%	84.6%
	Highly	65%	56.3%	64.5%
Mean of Maximum	No	99%	98.3%	100%
	Lowly	57.1%	None	50%
	Moderately	57.6%	75%	56.4%
	Highly	96.7%	100%	94.4%
Largest of Maximum	No	97%	97.7%	95%
	Lowly	28.6%	None	31.3%
	Moderately	0.06%	75%	0.05%
	Highly	98.3%	100%	97.6%
Smallest of Maximum	No	99%	99.4%	100%
	Lowly	57.1%	None	50%
	Moderately	57.6%	75%	61.5%
	Highly	70%	62.5%	66.9%

From Table 4.2, the centroid defuzzification method's performance on low and moderate relevance is not as good as the polar relevance's performance. The major problem is that the boundaries among these degrees are unclear. People are unable to unify the answers or opinions. For instance, "seriously worried about my beach house #probablygone" is graded to be lowly relevant by 7 people, moderately relevant by 6 people and irrelevant by 2 people. We adopt the largest score among the different degrees to represent its rank. Based on this idea, this sample is labelled to be lowly relevant then.

However, the vote result shows that it is close to the moderate relevance. Thus, it is hardly to claim that this sample completely belongs to low relevance. Fuzzy rules should identify it belongs to these two degrees with varying memberships. This is the main reason leading to the low accuracy for both situations of low relevance and moderate relevance.

In Table 4.1, comparing the other defuzzification methods, the results show that the fuzzy logic-based model achieves quite high accuracies on polar relevance problems. The performance of the centroid method and the MOM method is approximately same and effective. LOM method is the worst one among these defuzzification methods.

In Table 4.2, we can find out that the fuzzy logic-based model's ability of handling the four-degree problem is worse than that of handling polar problems. Note that, there is no lowly relevant sample in the second dataset. Hence, the marks in tables are "None". The main reason caused low performance is mentioned above, because we cannot reconcile the unclearness of boundaries. The performance of the centroid method and MOM is still very close, and the LOM is again the worst. By further comparing centroid and MOM, we select the centroid defuzzification method in the remaining work because its capability of classifying polar relevance problem is quite better than MOM's.

4.5 Comparison with Other Methods

The works [Lu, 2016] [Guan, 2014] [Dong, 2013] use a keyword search method to extract relevant tweets from the original dataset. The keyword list in [Lu, 2016] includes "sandy", "hurricane" and "storm". Its advantage is straight forward, effective and high accurate for those highly relevant tweets. However, because of the limitation of the keyword list, the defect of this method is that it is unable to extract enough relevant tweets.

With the consideration of quantity and correctness rate, we decide to conduct a comparison between the keyword search method and the proposed fuzzy logic-based model. Five sets of tweets from the original dataset are prepared for this experiment. Each set of data contains 10,000 randomly selected tweets. We realize the keyword search method and use the same keyword list used in [Lu, 2016]. The comparative analyses of quantity and correctness rate are shown in Table 4.3. In this experiment, fuzzy logic-based model returns the tweets including lowly, moderately and highly relevant tweets, i.e., we use the polar relevance classification method.

Since we cannot calculate the number of correctly identified irrelevant and relevant tweets, a correctness rate is defined as:

$$\gamma = \frac{X}{Y} \times 100\% \tag{4.1}$$

where Y decides the total number of relevant tweets extracted by each method, X is the number of correctly classified tweets in Y. Note that X is calculated by manually double check, i.e., we examine how many tweets are classified correctly in Y.

Additionally, an incremental rate λ describes that the proposed model can exploit more information than the keyword search method, which is defined as:

$$\lambda = \frac{X_f - X_k}{X_k} \times 100\% \tag{4.2}$$

where X_f is calculated by the proposed fuzzy logic-based model, and X_k is calculated through the keyword search method.

Table 4.3 The Comparison Results of Two Methods

Dataset	Method						λ
	Keyword search			*Fuzzy logic-based*			
	Y	X	γ	Y	X	γ	
1	98	96	97.9%	141	135	95.7%	40.6%
2	103	103	100%	161	157	97.7%	52.4%
3	86	86	100%	128	126	98.4%	46.5%
4	93	92	98.9%	137	132	96.3%	43.5%
5	99	99	100%	122	118	96.7%	19.1%

By manually reviewing and analyzing the results computed by both method, we find out that all tweets extracted by keyword search method appear in the results of fuzzy logic-based model. In other words, the results of keyword search method are the subsets of the results calculated by fuzzy logic-based model. The latter successfully mines more tweets than the former does. In Table 4.3, the values of λ indicates that the latter successfully mines more tweets than the former does. After sorting all the results of the latter according to the values of relevance, we manually double check the results to find those misclassified tweets. Additionally, we find out that there exists a common error in both methods. "I'm at Sandy Hook Diner (Sandy Hook, CT)" is extracted by both algorithms. This sample is unrelated to our background, because it talks about a location's name which includes "sandy".

It is unsurprised that fuzzy logic model still succeeds in extracting many other informative tweets while the keyword search method fails to do so. For instance, "The wind

is picking up and so is the rain! Luckily the power is still on." and "3 trees & a basement GONE" are graded to be moderately relevant by the fuzzy logic-based model. Keyword search method fails to extract them for lacking of the corresponding keywords in the two tweets.

To summarize, the fuzzy logic-based approach can extract much more tweets than the keyword search method. Only considering the quantity, the fuzzy logic model is more powerful than the latter. With the consideration of correctness rate, the keyword search method works slightly better than the former. Considering both criterion, we claim that the fuzzy logic model is preferred in research context where more relevant tweets are highly desired for the analysis stage, such as [Lu, 2016][Guan, 2014]. High quantity and high correctness rate can guarantee more informative and useful data.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary of Contributions of This Thesis

This work proposes a fuzzy logic-based text classification method based on social media data. It is meaningful to analyze the correlation between social media and human-affected events. Since the quantity of social media data is large and countless, many useless messages should be filtered. Through the analysis of experimental results, we can claim that fuzzy logic-based method is qualified in text classification. This thesis makes the following contributions.

(1) Making literature review about social media, text classification and fuzzy logic.

Social media platforms provide users an opportunity to exchange messages, opinions and experiences, which produce numerous quantity of text data. Then, the increasing availability of text documents in Twitter draws researchers' attentions as well as interests in developing automatic analysis methods for them because of the potential connection between posted tweets and various events. Consequently, text classification is introduced to solve this problem of classifying event-related or non-related messages. However, handling the vagueness and ambiguity of text data is a challenging issue. Fuzzy logic provides us a good solution. Thus, its possible use to build fuzzy rules to convert words into numerical values for computing and reasoning motivates this thesis work.

(2) Proposing a fuzzy logic-based method for relevant and irrelevant message classification for Twitter data sets.

We extract seven features from each tweet message, and treat them as inputs to the proposed fuzzy logic-based model. Then, we obtain a value indicating the relevance degree

to Hurricane Sandy for each tweet message.

(3) Conducting comparisons with other methods and analyzing experimental results.

According to the experimental results, the fuzzy logic-based model performs well with the polar relevance problem, i.e., it is good at two-class classification. The performance of a four-degree relevance problem is not satisfied. We conduct a comparison with the well-known keyword search method which is effective and highly accurate. By analyzing the experimental results, we claim that the proposed method can extract 40% relevant tweet messages more than the keyword search method

5.2 Limitations

However, this proposed method has some limitations.

(1) The proposed model can only classify Hurricane Sandy 2012 related Twitter text data.

Namely, the fuzzy logic-based model designed in this work cannot be used in other contents or events. The reason is that the proposed model is built based on the specific background. Social media data do not have context information. Because of this, we cannot use them directly to solve problems.

(2) The fuzzy logic-based model cannot well classify the four-degree relevance problem.

The imperfect fuzzy rules and insufficient input parameters result in a low performance of classifying a four-degree relevance problem, especially on low and moderate relevance classification. Nevertheless, by reviewing the training dataset, we observe that many tweets confuse the volunteers, which means people are unable to make a clear and definite decision, nor machines do.

5.3 Future Work

As future work, we attempt to find a better way to calculate the similarity score between words and between two tweets. The fuzzy rules need to be improved such that we can obtain a better classification result on a four-degree relevance problem. Besides, contrasting to the quantity of original data crawled from Twitter, the volume of the training dataset in this work is quite small. Hence, we intend to use a larger volume of training dataset to build a better fuzzy logic-based model. Natural language processing techniques, such as stemming and lemmatization, and sentiment analysis, can be applied during data preprocessing. Moreover, we can conduct comparisons with many other methodologies, such as TF-IDF and LSI. Currently, we focus on analyzing the rare events. The further work can cover other events, such as sports and pre-determined events such as New Year eve celebration.

REFERENCES

B. Al-Najjar and I. Alsyouf, "Selecting the most efficient maintenance approach using fuzzy multiple criteria decision making," *International Journal of Production Economics*, vol. 84, no. 1, pp.85-100, 2003.

N. Alsaedi, P. Burnap, and O. Rana, "Sensing real-world events using Social media data and a classification-clustering framework," in *Proc. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI),* Omaha, Nebraska, USA, October 12-16, 2016, pp. 216-223.

N. Bidi and Z. Elberrichi, "Feature selection for text classification using genetic algorithms," in *Proc. 2016 IEEE 8th International Conference on Modelling, Identification and Control (ICMIC),* Algiers, Algeria, November 15-17, 2016, pp. 806-810.

C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, and A. Tapia, "Mapping moods: geo-mapped sentiment analysis during hurricane Sandy," in *Proc. 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, state college, PA, USA, May 18-21, 2014, pp. 642-651.

N. Chawla, N. Japkowicz, and A. Kolcz, "Special issue on learning from imbalanced datasets, sigkdd explorations," in *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*ACM SIGKDD)*, Seattle, WA, USA, August 22-25, 2004, pp. 1-6.

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

R. Damaševičius, R. Valys, and M. Woźniak, "Intelligent tagging of online texts using fuzzy logic," in *Proc. 2016 IEEE Symposium Series on Computational Intelligence (SSCI),* Athens, Greece, December 6-9, 2016, pp. 1-8.

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

H. He and E.A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp.1263-1284, September 2009.

Q. Jiang, W. Wang, X. Han, S. Zhang, X. Wang and C. Wang, "Deep feature weighting in Naive Bayes for Chinese text classification," in *Proc. 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Beijing, China, August 17-19, 2016, pp. 160-164.

R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, and H. Ji, "On predicting social unrest using social media," in *Proc. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* San Francisco, CA, USA, August 18-21, pp. 89-95.

Q. Kang, X. Chen, S. Li, and M. Zhou, "A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification," *IEEE Transactions on Cybernetics*, to appear in 2017.

A. Kasun, M. Manic, and R. Hruska, "Optimal stop word selection for text mining in critical infrastructure domain," in *Resilience Week (RWS)*, Philadelphia, PA, August 18-20, pp. 1-6.

D. Knuth, H. Szymczak, P. Kuecuekbalaban, and S. Schmidt, "Social media in emergencies: How useful can they be," in *Proc. 2016 IEEE 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM),* Vienna, Austria, December 13-15, 2016, pp. 1-7.

H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media," in *Proc. 19th International Conference on World Wide Web*, Raleigh, NC, USA, April 26-30, 2010, pp. 591-600.

G. H. Kulkarni and P. G. Waingankar, "Fuzzy logic based traffic light controller". In *Proc. IEEE International Conference on Industrial and Information Systems* (*ICIIS*), University of Peradeniya, Sri Lanka, August 9-11, 2007, pp. 107-110.

R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han, and H. Ji, "On predicting social unrest using social media," in *Proc. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, USA, August 18-21, 2016, pp. 89-95.

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

M. Ma, A. Kandel, and M. Friedma, "new approach for defuzzification," *Fuzzy Sets and Systems*, vol. 111, no. 3, pp.351-356, May, 2000.

Natural Language Toolkit (NLTK) [Online].
Available: http://www.nltk.org/, accessed Apr. 1, 2017.

J. D. Prusa and T. M. Khoshgoftaar, "Designing a better data representation for deep neural networks and text classification," in *Proc. 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, Pittsburgh, PA, USA, July 28-30,2016, pp. 411-416.

J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Impact of feature selection techniques for tweet sentiment classification," in *28th International FLAIRS Conference*, Hollywood, Florida, USA, May 18-20, 2015, pp. 299-304.

A. Patel and T. A. Champaneria, "Fuzzy logic based algorithm for Context Awareness in IoT for Smart home environment," in *Proc. 2016 IEEE Region 10 Conference (TENCON),* Marina Bay Sands, Singapore, November 22-25, 2016, pp. 1057-1060.

A. R. F. Quiros, R. A. Bedruz, A. C. Uy, A. Abad, A. Bandala, and E. P. Dadios, "Machine vision of traffic state estimation using fuzzy logic," in *Proc. 2016 IEEE Region 10 Conference (TENCON)*, Marina Bay Sands, Singapore, November 22-25, 2016, pp. 2104-2109.

F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Proc. IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),*Chennai, India, March 22-24, 2017, pp. 2264-2268.

K. Salah and A. Adel, "Model order reduction using fuzzy logic algorithm," in *Proc. 2016 IEEE 28th International Conference on Microelectronics (ICM),*Cairo, Egypt, December 17-20, 2016, pp. 13-16.

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," in *Proc. 2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Vienna, Austria, December 13-15, 2016, pp. 1-6.

H. Shekhar and S. Setty, "Disaster analysis through tweets," in *Proc. 2015 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI),* Kochi, India, August 10-13, 2015, pp. 1719-1723.

W. Sun and W. He, "Fuzzy logic control of an uncertain robot with output constraint," in *Proc. IEEE Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Wuhan, China, November 11-13, 2016, pp. 67-72.

L. Suanmali, M.S. Binwahlan, and N. Salim, "Sentence features fusion for text summarization using fuzzy logic," in *Proc.* 2009 *IEEE 9th International Conference on Hybrid Intelligent Systems*, Shenyang, China, August 12-14, 2009, pp. 142-146.

J. B. Sathe and M. P. Mali, "A hybrid Sentiment Classification method using Neural Network and Fuzzy Logic," in *Proc. 2017 IEEE 11th International Conference on Intelligent Systems and Control (ISCO),*Coimbatore, India, January 05-06, 2017, pp. 93-96.

D. Sewwandi, K. Perera, S. Sandaruwan, O. Lakchani, A. Nugaliyadde, and S. Thelijjagoda, "Linguistic features based personality recognition using social media data," in *Proc. IEEE 6th National Conference on Technology and Management (NCTM*), Malabe, Sri Lanka, January 27-27, 2017, pp. 63-68.

A. Sheshasayee and G. Thailambal, "A comparitive analysis of single pattern matching algorithms in text mining," in *Proc. 2015 IEEE International Conference on Green Computing and Internet of Things (ICGCIoT)*, Delhi, India, 2015, October 08-10, 2015, pp. 720-725.

J. Sun, F. Karray, O. Basir, and M. Kamel, "Fuzzy logic-based natural language processing and its application to speech recognition," in *3rd WSES International Conference on Fuzzy Sets & Systems*, Interlaken, Switzerland, February 11-15, 2002.

C. Yin, J. Xiang, H. Zhang, J. Wang, Z. Yin, and J. U. Kim, "A new SVM method for short text classification based on semi-supervised learning," in *Proc. 2015 IEEE 4th International Conference on Advanced Information Technology and Sensor Application (AITS),* Washington, DC, USA, August 21-23, 2015, pp. 100-103.

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

W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758-2765, March, 2011.

Z. U. A. Zulkifli, M. F. R. Tasriva, and M. A. Matiur, "Fuzzy logic controller design for intelligent drilling system," in *Proc. 2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS),* Shah Alam, MY, October 22-22, 2016, pp. 208-213.

L. A. Zadeh, "Fuzzy logic = computing with words," *IEEE transactions on fuzzy systems*, vol. 4, no. 2, pp.103-111, May, 1996.