International Journal of Fuzzy System Applications
Volume 13 • Issue 1

Fuzzy SVM With Mahalanobis Distance for Situational Awareness-Based Recognition of Public Health Emergencies

Dan Li, Zhejiang College, Shanghai University of Finance and Economics, China
https://orcid.org/0000-0003-4519-2483
Zheng Qu, Shanghai Business School, China*
Chen Lyu, Shanghai University of Finance and Economics, China
Luping Zhang, Zhejiang College, Shanghai University of Finance and Economics, China
Wenjin Zuo, Zhejiang College, Shanghai University of Finance and Economics, China

ABSTRACT

In public health emergencies, situational awareness is crucial for swift responses by governments and rescue organizations. In this manuscript, a novel framework is proposed to identify and classify event-specific information, aiming to comprehend concepts, characteristics, and classifications associated with situational awareness in social media emergencies. First, a statistical approach is employed to extract a set of standard features. Second, a category-based latent dirichlet allocation to vector (LDA2vec) model is leveraged to extract topic-based features to enhance accuracy, particularly for unbalanced datasets. Finally, a fuzzy support vector machine (FSVM) classifier utilizing the Mahalanobis distance kernel is introduced to improve the detection accuracy of event-specific information. The framework's effectiveness is evaluated using the social media public health dataset, achieving superior filtering capabilities for non-informative data with a precision of 89% and an F1-Score of 91%, surpassing other standard methods.

KEYWORDS

FSVM, LDA2vec, Mahalanobis Distance, Public Health Events, Situational Awareness

INTRODUCTION

Public health emergencies are characterized by their suddenness, speed, and unpredictability, presenting significant challenges to emergency management (An et al., 2018). Governments and voluntary relief organizations should strive to collect and understand relevant disaster information to aid emergency response operations (Fu et al., 2020). Situational Awareness (SA) (Huang & Xiao, 2015), which involves gathering and comprehending relevant crisis information (i.e., what is occurring in impacted communities during an event), is critical to this process. Social media has become a primary mode of disseminating information online, owing to its speed, versatility, and interactivity.

DOI: 10.4018/IJFSA.342117
*Corresponding Author

It also serves as a substantial communication channel, particularly for situational awareness during emergencies like natural calamities.

However, due to the diversity of online user communities, online news content varies widely, posing challenges for relevant agencies in swiftly gaining situational awareness of events. The key to enhancing the speed of emergency response to unforeseen events and minimizing associated losses is efficiently collecting pertinent information related to situational awareness from vast amounts of data in the shortest possible time frame. The use of social media for situational awareness during unforeseen events typically involves tasks such as social media text classification and semantic mining, which includes parsing concise and informal messages, managing information overload, and prioritizing different types of information identified within these messages. These tasks can be mapped to classical information processing operations, such as filtering, categorization, sorting, aggregation, extraction, and summarization (Imran et al., 2015; Liang & Li, 2020).

In recent years, an increasing number of scholars have explored the implementation of techniques, including natural language processing, machine learning, and deep learning, for the automated processing of social media breaking news messages (Xia et al., 2021). However, there still needs to be a framework to identify and classify event-specific information, primarily due to the complexity of the task and the dynamic nature of online information (Nan et al., 2022). Such a framework is necessary for the ability of relevant agencies to efficiently process and make sense of the vast amount of data generated during unforeseen events. This gap in existing methodologies arises from the following two factors: the complexity of information and unbalanced data issues.

On the one hand, the diverse and dynamic nature of online content, particularly during emergencies, poses challenges in developing a comprehensive framework. The sheer volume of information and the rapid evolution of events demand a sophisticated approach to extracting relevant details. On the other hand, dealing with unbalanced data sets, where informative and non-informative data may be unevenly distributed, adds another layer of complexity. A robust framework should account for these imbalances to ensure accurate and unbiased results.

In light of these challenges, developing a comprehensive and adaptable framework becomes imperative. This paper proposes an automated and comprehensive framework for identifying informative information. Using situational awareness, our approach aims to comprehend the concepts, features, and categories of informative information related to public health emergencies on social media. First, we define the concepts, characteristics, and components of informative information regarding public health emergencies based on situational awareness on social media. Second, we employ statistical methods to extract traditional features, including linguistic, numeric, punctuation, and source-based features. Third, we enhance our framework by extracting topic-based features using a category-based latent Dirichlet allocation to vector (LDA2vec) model, tailored explicitly for addressing unbalanced data sets. Finally, we introduce a fuzzy support vector machine (FSVM) classifier designed to handle unbalanced and noisy data, utilizing a kernel based on Mahalanobis distance rather than the traditional Euclidean distance kernel. The effectiveness of our framework is assessed through comparisons with traditional machine learning models and state-of-the-art methods. To further validate our approach, we leverage a BERT pre-trained model to cluster the classification results, demonstrating that informative information has a superior situational awareness effect.

The main contributions of our work could be summarized as follows:

- We proposed an automated and comprehensive framework for the identification of informative information. This framework addressed the complexities associated with public health emergencies on social media, aiming to provide a holistic solution to the challenges posed by the dynamic nature of online information.
- We leveraged a category-based LDA2vec model to extract topic-based features. This approach enhanced the capability of our framework to grasp nuanced information related to public health

emergencies, which is particularly beneficial for handling unbalanced data sets and improving overall classification accuracy.

- To improve the detection accuracy of informative information, we designed an FSVM classifier utilizing the Mahalanobis distance kernel, departing from the conventional Euclidean distance kernel. The FSVM is adept at handling unbalanced and noisy data, making our framework robust and effective in real-world scenarios.

The rest of our paper is organized as follows. The related work is introduced in Section 2. Informative information in public health emergencies is described in Section 3. The construction of our framework is presented in Section 4. We perform many experiments and analyze the comparative results in Section 5. Situational awareness analysis is done in Section 6. Finally, Section 7 summarizes and concludes our work.

RELATED WORK

Essentials of Situational Awareness Information

During emergencies, various types of information on the web, including updates, personal opinions (e.g., on the adequacy of rescue operations), and emotions (e.g., expressions of sympathy for those affected by the disaster), are posted by users in huge quantities and at incredible speeds. This situational information can help the government, as well as the rescue organizations, gain an overview of the overall situation of the event (Rudra et al., 2018).

Hornmoen et al. (2021) argued that social media can significantly enhance risk and crisis management by facilitating situational awareness systems. Vieweg et al. (2010) encoded social media data using geo-location, location referencing, and scenario updating to elevate situational awareness. Imran et al. (2013b) classified social media data into five categories: warnings and advisories, casualties and losses, donated goods and services, missing and found persons, and sources of information to guide manual annotation, categorize the information, and extract valuable information. Wang and Ye (2018) examined situational awareness from the three dimensions of time, space, and semantics in social media data. Imran et al. (2014) categorized messages posted during disasters into a set of user-defined information categories (e.g., "needs," "damages") and, with the help of crowdsourcing, manually extracted various tweet features to identify information in tweets. Bouzidi et al. (2022) reported that social media information related to warnings, alerts, situational awareness, and disaster education is valuable for situational awareness in public health emergencies. Scheele et al. (2021) proposed a geospatial context-aware text mining approach that combines spatial and temporal information from social media and authoritative data sets with text information. This approach was used to classify disaster-related social media posts, enhancing emergency situational awareness. Rudra et al. (2018) differentiated informative information across events, languages, and perspectives and proposed a method for identifying and extracting non-informative information. Castillo et al. (2011) studied how to classify disaster management-relevant information in tweets based on message, user, and message propagation features. They conducted extensive manual evaluation of credibility, achieving precision and recall rates ranging from 70% to 80%.

Classification of Social Media Text Information

Feature Extraction

Currently, the research of text information classification methods on social media mainly focuses on feature extraction and the use of classifiers. Feature extraction aims to obtain a feature system with the most influential category differentiation ability. Conventional feature extraction involves basic features based on lexical properties, bag of words, among other things, and statistical features extracted through methods such as TF-IDF and sorted indexing, among others. Additionally, features can be

extracted through the word2vec model of deep learning. Salton and Yu (1973) proposed a novel method for determining the weight of words by exploring TF-IDF. The importance of a word for the text was determined by analyzing its frequency in a corpus and original text. Then, content was selected based on keyword information to improve the accuracy of extracted content. Chouigui et al. (2018) applied this approach to extract information about events in Arabic news using co-occurring words, varying performance by corpus type and domain. Sriram et al. (2010) also explored this approach. Manually annotating various categories of tweets using multiple classification techniques based on lexical and structural features alongside a limited number of domain-specific characteristics extracted from authors' profiles and texts effectively classifies the text into a pre-defined set of generalized categories. Linguistic features, such as those used by Freitas and Ji (2016) were also employed. The use of slang and emotive vocabulary must be avoided when identifying the value of tweets; instead, opt for active learning and content-based features to maintain recall and improve trend classification precision. However, this method is only suitable for popular topics. Hasan et al. (2019) developed a real-time event detection system, which uses inverted indexing and incremental clustering methods to detect newsworthy events at low computational cost. These studies predominantly employ conventional word frequency statistics methods. However, these methods are sensitive to the particular corpus utilized, which may lead to restricted generalization of their models. Furthermore, given the vast amount of COVID-19 data accumulated, machine learning-based approaches may exhibit suboptimal performance in extracting critical information.

Word embedding methods such as word2vec (Mikolov et al., 2013) are standard techniques for extracting topic-based features for text recognition and are largely dependent on the corpus used. Khatua et al. (2019) tested the accuracy of domain-specific word vectors for recognizing crisis-related actionable tweets in the context of the Ebola virus in 2014 and the Zika virus outbreak in 2016, and found that extracting meaningful domain-specific semantic relations was better than training via word2vec or via NLP, thus they argued that it is better to train a context-specific corpus for each burst of information. Wiedmann (2017) devised a joint learning model of structural and textual features for web event extraction, which applies to a wide range of domains. These studies can obtain both textual formal features and content-based features. However, most Word2Vec methods fail to discover relationships between text contexts, and their effectiveness is closely related to the corpus used. Other studies using topic models ignore the imbalance of samples in natural environments, which may lead to biased results of their topic models.

Based on these works, we adopt a category-based LDA2vec model to extract topics from different categories of samples to solve the problem of topic extraction model failure caused by sample distribution bias, which effectively complements the deficiency of traditional features in feature extraction.

Classification of Social Media Information

Training a classifier is the final and especially crucial step in the information classification. A classifier model with numerous parameters is generated through extensive training on large corpora. Its input is a text feature vector, and its output is the predicted text category. Machine learning and deep learning algorithms are commonly used for classification. Kumar et al. (2020) analyzed massive tweets, images, and videos based on the long short-term memory (LSTM) model to identify information related to sudden events on Twitter, thus assisting people in making timely decisions. Lu et al. (2014) utilized unsupervised clustering methods to identify valuable related topics based on pages and posts on social networks and achieved promising results through a combination of deep neural network models. Liu et al. (2016) employed a combined approach of supervised and unsupervised learning to achieve text classification for network services. They started with a small training set with basic classifiers and then iteratively request labels from the most informative services outside the initial training set. They addressed issues arising from sparse term vectors in service descriptions by combining probability topic models and using an SVM to enhance the effectiveness of text classification on the corpora.

Classification of Public Health Emergencies

In the process of risk communication in public health emergency management, the timely and accurate information of sudden public health events is of significant importance in crisis propagation (Li et al., 2022). Therefore, accurately distinguishing the value of different information has become a hot research topic among scholars. In disease surveillance, Unankard et al. (2015) proposed a method based on support vector regression (SVR) to extract influenza-related information from social networks to predict the development of influenza in the United States. Y. Wang et al. (2020a) used word embeddings to encode Twitter texts and input them into a convolutional neural network (CNN) structure to train a classifier for personal health identification. However, this method tended to favor classes with a large number of training samples, as the reliability and classification performance improves with more data. Broniatowski et al. (2013) used an SVM to better differentiate between actual influenza-related tweets and tweets that appear to be relevant but are not actually about influenza and found that the SVM has higher accuracy than other methods. Byrd et al. (2016) used a naive Bayes classifier to classify tweets based on influenza-related keywords, achieving a classification effectiveness of up to 70%. Alessa and Faezipour (2018) described the potential of social media posts in detecting disease outbreaks and providing early warning. Their research mainly focused on predicting disease transmission and classifying disease-related posts using lexical statistical methods. Lee et al. (2013) devised a novel real-time influenza and cancer monitoring system that utilized spatial, temporal, and text mining of Twitter data to detect disease outbreaks and employed machine learning methods to identify influenza and cancer based on frequency. Missier et al. (2016) designed a model for tracking dengue fever outbreaks on Twitter, evaluating 1,000 tweets using naive Bayes and language models based on LDA topic models and analyzing the dengue fever epidemic in the Philippines, achieving classification of epidemic-related tweets.

Most of the studies mentioned above utilized traditional machine learning models for identification and feature extraction. However, traditional machine learning models exhibit certain limitations for public health emergencies on social media with a large volume of information and significant sample imbalances. To address sample imbalances, we carefully considered the relevance of features and proposed an FSVM classifier based on the Mahalanobis distance kernel.

INFORMATIVE INFORMATION IN PUBLIC HEALTH EMERGENCIES

Definition of Informative Information

Situational awareness first emerged in the military field as a way to enhance the discovery, recognition, understanding, analysis, and response capabilities of security threats from a global perspective, based on security big data. According to Endsley(1995), situational awareness refers to "perceiving elements in the environment within a certain time and space, understanding their meaning, and predicting their future state in the near term". In 1995, Endsley proposed the operational framework of situation awareness theory in dynamic decision-making, with the core being the situation awareness model, which includes three levels: perception of situational elements, understanding of the situation, and prediction of the situation.

Emergency decision-makers need to formulate response plans quickly in the face of complex environments, urgent time constraints, missing information, and immense pressure. Therefore, it is crucial to accurately understand the status of unexpected events, which involves identifying information with situational awareness elements from a large amount of news. Perception of situational elements refers to grasping the overall picture of an event and obtaining its essential elements and attributes. In the field of natural disasters, this mainly includes identifying and confirming disasters, damage intensity, disaster behavior, and causes and ways of the current state. In information security, situational elements refer to understanding and acquiring data, as well as collecting capabilities. Different fields have different understandings of situational elements, but they all have a common goal: to help relevant

entities grasp the full picture and status of an event. In this work, we define informative information based on situational awareness as information that contains situational awareness elements and assists in understanding and predicting the situation for relevant entities.

Classification Methodology of Public Health Emergencies Based on Situational Awareness

Different types of emergency events have different elements of situational awareness, and scholars provide no specific standards. For example, based on the elements of the event, the elements of the event object, and the elements of the event environment, the elements of situational awareness can be summarized for sudden natural disasters, accidents, public health emergencies, and social security events (Bruns, 2020; Olteanu et al., 2014). Alternatively, situational elements can be summarized based on the subjective emotions toward factual information. Imran et al. (2013a) categorized informative information of emergency events into eight categories, based on whether the information could be helpful to others, and provided examples of some situational elements. The statement from Imran et al. (2016) suggested that elements contributing to situational awareness in sudden public events encompass information related to infrastructure damage, missing persons, individuals trapped or injured, casualties, available shelters, volunteers, and rescue operations, among other relevant factors. Some researchers also consider non-situational awareness elements. Qu et al. (2011) believed that information related to donations or charity, praise or criticism of rescue actions, tragic views on preventing similar incidents in the future, and post analysis of the ways and causes of disasters were non-situational awareness elements. No directly applicable information is available for situational awareness elements of public health emergencies. In this work, we reference existing literature and combine practical data to organize the elements of situational awareness information for sudden public health emergencies, to guide the construction of manual annotation and feature systems.

The use of social media data for situational awareness primarily focuses on three dimensions of social media: time, space, and semantics (Z. Wang & Ye, 2018). Time information is mainly used to help detect when a disaster erupts and the entire development process of the disaster. Spatial information is mainly used to help government departments understand where the crisis and losses are more severe (Huang et al., 2015; Kryvasheyeu et al., 2016; Y. Wang et al., 2020b). Semantic information can help managers understand what happened during the disaster. In this work, we start from the three dimensions and summarize each situation's situational awareness elements related to public health emergencies. The time dimension focuses on the time of the outbreak of the epidemic, the development process of the epidemic, who is affected, and what advice relevant departments give at what time, among other things. Spatial information mainly helps relevant departments understand what happened during the epidemic, starting with the construction of facilities and public utilities and understanding the basic situation and scope of the event. Semantic information is used for rescue and donation based on whether the information content can assist relevant groups. Some other special situations can be summarized in semantic information, as shown in Table 1.

FRAMEWORK FOR INFORMATION CLASSIFICATION OF PUBLIC HEALTH EMERGENCIES

Overview of Our Framework

To accurately identify informative information regarding situational awareness in the public network, we seek to extract text features of public emergencies and classify them rapidly and accurately. With the situational awareness theory as the foundation, our feature extraction stage surpasses the technical methods based solely on text coding. Consequently, our methodology enables rapid identification of critically informative information, particularly in the initial stages of emergent situations. The framework we propose in this paper is shown in Figure 1.

Table 1. Classification of situational awareness elements of public health emergencies

Dimension	Concrete elements	Examples
Time	Confirmation, death, cure, warning and advice, vaccination guidance, tips, reports, vaccines etc.	"The latest epidemic map # This is the special data of @Alipay: Figure 1 is February 10, 2020, at 9:30, Figure 2 is February 12, 2020, at 10:00 suspected cases have decreased significantly, and the cure rate has increased!"?
		"How to ride the elevator safely? (Xinhua News Agency) L Jiangxi Daily Weibo video?"
Space	Road closures, closures and services, changes to traffic, hospitals, clinics, locations, ranges, repeated, secondary outbreaks etc.	"Type of coronavirus pneumonia COVID-19ARS-CoV-2# O is the first in the country! Hubei Shiyan Zhangwan Area announced the implementation of wartime control of all buildings closed management? "
		"After cases of the novel coronavirus pneumonia have spread to 25 countries and more than 40,000 people have been diagnosed, it finally has an official name - COVID-19. The Art of traveling in China? "
Semantics	Donation, volunteer, rescue, Foundation etc.	The new coronavirus pneumonia COVID-19ARS-CoV-2# O Wuhan after 90 volunteer team initiators: Some people do practical things, and some people scratch materials?
		Who names COVID-19: 'COVID-19' stresses name to avoid geographical reference Vaccine expected to be ready in 18 months WHO names COVID-19: COVID-19?

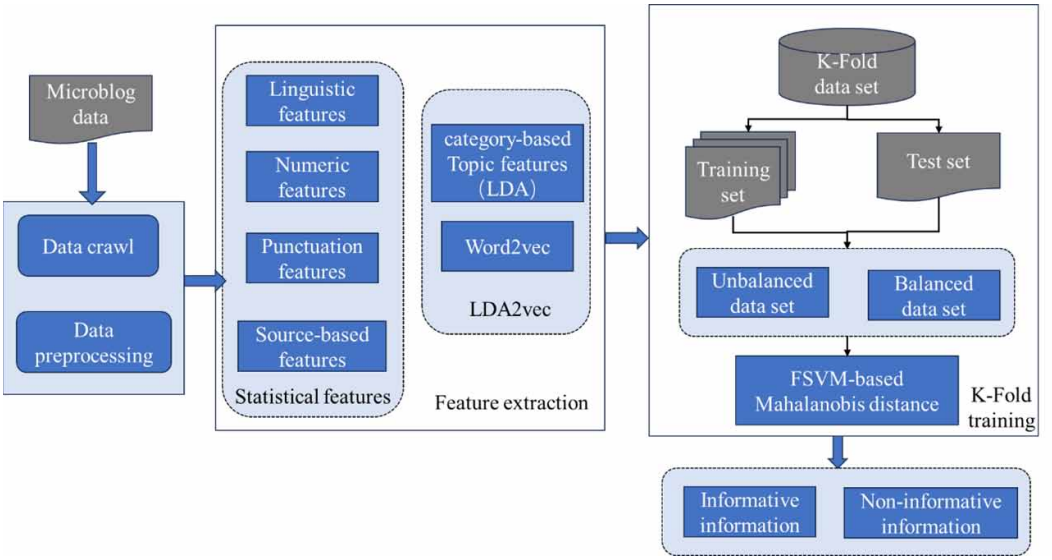
First, we extract information features, including linguistic, numeric, punctuation, and source-based features. These features serve as fundamental elements for illustrating distinct information in textual form. Second, we design a category-based LDA2vec model to extract the topic-based feature for acquiring the feature vector of the information data. The feature vectors are then deployed to represent the variations in informational details across different situations in the text. Finally, we input the extracted feature vectors related to information into an FSVM classifier based on the Mahalanobis distance kernel for classification, categorizing the information into informative and non-informative information.

Although traditional SVM classifiers are proficient at text classification, they encounter specific issues. First, the distribution of relevant news samples on social networks is often distorted, with only a small amount of valuable news among a large amount of irrelevant information. Second, the segmentation of SVM hyperplanes is impacted by numerous noise points in the data set, resulting in potentially biased classification outcomes. Third, conventional SVM or FSVM classifiers that use Euclidean-distance kernels provide equal weighting to various attributes, which impairs the efficiency of the classifier. Thus, we introduce the Mahalanobis distance instead of the Euclidean distance to amend the FSVM model for enhanced model detection accuracy.

Information Feature Extraction

Traditional statistical features include linguistic features, numeric features, punctuation features, and source-based features. In previous studies, these features have been proven to solve tasks such as event extraction and text classification. Topic-based features complement text features by capturing potential semantic aspects in social networks. Since word2vec focuses on local features between words, and LDA addresses global features between contexts, our design involves a category-based LDA2vec model that integrates these two features. Finally, we obtain a combination of statistical and topic-based features, shown in Table 2.

Figure 1. The overview of our framework



Category-Based LDA2vec Model

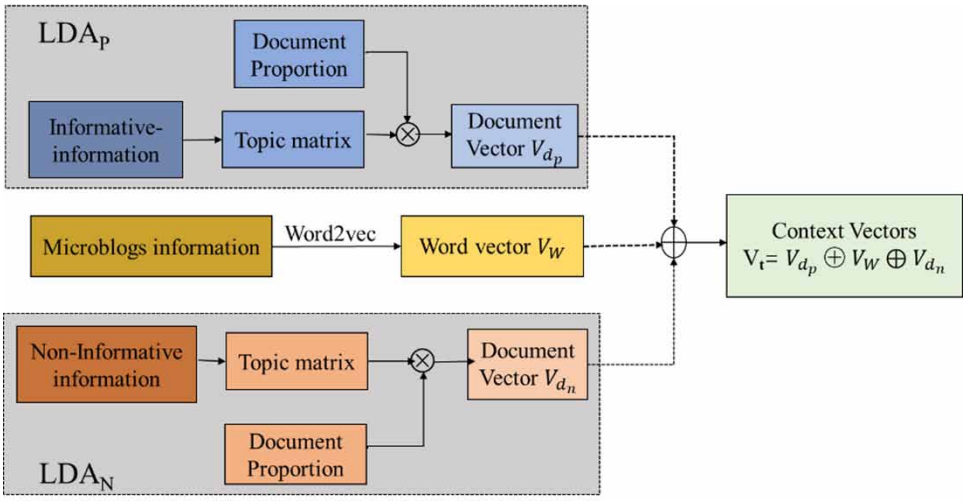
The potential semantic features in social media may represent what will happen in the future, which is of great significance for situational awareness prediction, so obtaining topic features is crucial for situational awareness. First, a word segmentation tool is used to exclude stop and low-frequency words, and microblogs with the same label are connected with documents. Because there is a certain distortion between informative information and non-informative information (about 20% is informative information), the training results of the model will inevitably be biased toward the valueless microblogs in the data set due to the unsupervised attribute of LDA. Therefore, two LDA models are used to train the data set before classification. One is trained on the informative information microblog, denoted as LDA_P. The other is trained on the non-informative information microblog, abbreviated as LDA_N. The category-based LDA2vec model is shown in Figure 2, and the steps are as follows.

1. The samples of the two categories are trained respectively to obtain two topic models. Both document Vector V_{d_p} and V_{d_n} are weighted sums of the subject vectors.
2. All microblog information is trained by word2vec model to get word vector V_W.

Table 2. Information feature extraction

Category	Specific description
Topic-based features	LDA2vec
Statistical features	Linguistic features: nouns, adjectives, text length, etc.
	Numeric features: Arabic numerals, Chinese numerals
	Punctuation features: '!', '? ', '%', '$', etc.
	Source-based features: government enterprises, schools, media and other large and small organizations or individuals, etc.

Figure 2. Category-Based LDA2vec model



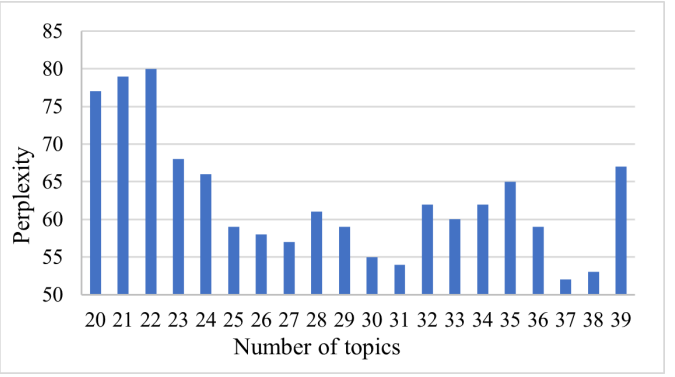
3. The context vectors are the sum of the V_{d_p} and V_{d_n} and V_W. Because length of information in informative and non-informative information is very similar, we assume that V_{d_p} and V_{d_n} have the same dimension.

Then, the normalized confusion is calculated according to the topic data, as shown in Figure 3. The five topics with the lowest confusion are selected as the number of candidate topics, which are 30, 31, 36, 37, and 38. Finally, 30 is selected as the number of topics. Therefore, V_{d_p} and V_{d_n} have a dimension of 30.

Statistical Features

Statistical features are those features of the data set that can be defined and calculated via statistical analysis.

Figure 3. Normalized perplexity



Linguistic features (S_L) can reflect the user with more psychological activity and behavioral intention. Therefore, we calculate the frequency of each word class (such as noun or adjective) in the microblog, denoted as S_{pos}, and S_{len} represents the length of the microblog, whose value is equal to the number of words in the microblog. Finally, we obtain the 10-dimensional vector of linguistic features.

Numerical features often reflect accurate and reliable information. For example, the number of confirmed cases, cures, and deaths is a good representative of the development trend of the epidemic, and the phone number may be an emergency call. We get the numerical feature $S_{snum} = S_{num} / S_{len}$, indicating the proportion of numbers in the number of characters in microblogs, and the numerical feature is assigned $S_N = \left\{ S_{num} \cup S_{snum} \right\}$.

Punctuation features often represent people's emotional orientation; exclamation marks can be used for warning messages, question marks may represent questions, and so on. We use the number of punctuation marks to measure the statistical characteristics of punctuation marks. These punctuation marks include common symbols such as '!', '?') and special symbols such as '%' and '$'. This paper uses S_P is used to represent the usage of punctuation marks in each microblog.

Source-based features affect users' perceptions of text reliability. On the microblog platform, information sources or microblog senders are divided into two categories: official information sources (such as government enterprises, schools or media, and other large organizations) and non-official information sources (such as small organizations or individuals). This paper uses authentication information (Weibo authentication is marked by blue V) and the number of fans to assist in judging the reliability of the information source. We extract a source-based feature S_{gov} to indicate whether a Weibo is published by a public account. S_{gov} is a dummy variable, 1 represents an official Weibo, and 0 represents an unofficial Weibo.

We conduct regression analysis on some representative items from the aforementioned features to verify the significance of the features extracted based on statistical methods in the former. Additionally, we calculate their mean, standard deviation, and p-value across different categories in Table 3.

From Table 3, we can observe that most attributes demonstrate statistical significance (p-value less than 0.05), while some attributes exhibit vital significance (p-value less than 0.01). The p-values for text length are all less than 0.01 among the language features, indicating its highest significance. Both symbolic features and source features are also important. The symbolic of '%' and '?' is the most significant feature within the symbolic feature. This is because most Weibo posts that include '%' also contain numerical features and valuable information, while posts without informative information rarely include such symbols. It is worth noting that the p-value for the source-based features is also less than 0.001 since most official Weibo accounts are valuable sources of information.

FSVM With Mahalanobis Distance Kernel

The standard SVM employs Euclidean distance to calculate the distance between a sample and the hyperplane during classification. However, Weibo data are characterized by large amounts of noise and data volume, which leads to the following issues when employing the Euclidean distance method. On one hand, the importance of different features for calculating distance varies due to significant differences in dimensions. On the other hand, some relevant feature dimensions may interfere with each other during distance calculation. Fuzzy membership can fully solve the fuzzy relationship between samples and categories (Yu et al., 2021). Therefore, in this study, we use an FSVM model based on the Mahalanobis distance kernel, which replaces the Euclidean distance to balance the weight of sample features. The formula for calculating the Mahalanobis distance is as follows:

Table 3. Traditional statistics features

Information type Statistical features		Informative information		Non-informative information		P-value
		Mean value	Standard deviation	Mean value	Standard deviation	
Linguistic Features(S_L)	S_{pos}	3.017	2.860	3.083	3.056	0.07
	S_{len}	90.35	28.67	106.2	25.58	<0.001*
Digit Feature (S_N)	S_{num}	1.078	1.321	1.112	1.332	0.07
	S_{snum}	2.669	3.407	2.481	3.559	0.012
Symbolic Feature (S_P)	?	1.616	2.185	0.411	0.875	<0.001*
	%	0.002	0.001	0.006	0.009	<0.001*
	The other	0.058	0.118	0.052	0.112	0.021*
Source Feature(S_{gov})	Official type	0.045	0.328	0.529	0.232	<0.001*

$$D_M^2\left(x_i, x_j\right) = \left(x_i - x_j\right)^T \sum^{-1} \left(x_i - x_j\right), \tag{1}$$

where $D_M\left(x_i, x_j\right)$ represents the distance between two samples, and $\sum^{-1}$ is the covariance matrix of multi-dimensional random variables. Mahalanobis distance can eliminate the correlation and dimension difference between different feature dimensions in the samples, so that the improved fuzzy membership weight can effectively reflect the weight of the samples.

The RBF kernel function of Mahalanobis distance is expressed as follows:

$$K\left(x_i, y_j\right) = \exp\left(-\frac{x_i - y_j^{\,2}}{\sigma^2}\right) \tag{2}$$

$$\varnothing_{M-RBF} = e^{-\frac{D_M\left(x_i, x_j\right)^2}{\sigma}} \tag{3}$$

where $K\left(x_i, y_j\right)$ represents the kernel of the RBF function with parameter σ, $\varnothing_{M-RBF}$ represents the RBF kernel of the FSVM classifier.

The fuzzy membership degree of positive and negative samples is defined as:

$$S_{i^+}^{'} = 1 - \left[D_M \left(x_i, x_j \right) / \left(r_P + \delta \right) \right]^{\frac{1}{2}}, i = 1, 2, \cdots, m \tag{4}$$

$$S_{i^-}^{'} = 1 - \left[D_M \left(x_i, x_j \right) / \left(r_N + \delta \right) \right]^{\frac{1}{2}}, i = m+1, \cdots, n \tag{5}$$

$S_{i^+}^{'}$ and $S_{i^-}^{'}$ represent fuzzy membership calculated from Mahalanobis distance. The fuzzy membership degree of Mahalanobis distance is introduced into Equation (1), and the fuzzy FSVM representation based on Mahalanobis distance is obtained based on the fuzzy degree:

$$\min_{\omega,\xi} \frac{1}{2} w^2 + C \sum_{i=1}^{m} S_{i^+}^{'} \xi_i^{+} + C \sum_{i=m+1}^{n} S_{i^-}^{'} \xi_i^{-} \tag{6}$$

$$subject\ to \begin{cases} y_i \left(w^T \varnothing_{M-RBF} \left(x_i \right) + b \right) \geq 1 - \xi_i \\ \xi_i > 0, i = 1, 2, 3 \ldots, t \end{cases} \tag{7}$$

Since the deviation between positive and negative samples of public health emergencies related data is not severe, and the feature dimension of the model in this paper is high, the initial penalty coefficient C is the majority of samples (non-informative information samples) divided by the minority of samples (informative information samples).

DATA COLLECTION AND EXPERIMENTS

Data Acquisition and Preprocessing

This study collected the original Weibo posts related to COVID-19 topics and themes using the Weibo search API and a Weibo crawler from January 21, 2020, to May 1, 2020. The search API utilized keywords such as "COVID-19." Retweets and shared posts were excluded as they were considered duplicates. Additionally, a crawler program was developed to randomly select Weibo posts based on the search API results in order to avoid the influence of PageRank. A total of 15,612 Weibo posts were collected, and duplicate posts were removed since most of the information was replicated from the original Weibo data. We excluded Weibo posts without text content, which were unreadable or written in a language other than Chinese. We also excluded posts that only contained tags or specific "[]" tags. As a result, we obtained a data set of 12,600 samples, including informative and non-informative samples.

Data Annotation

We manually annotated the 12,600 Weibo posts, categorizing them as informative or non-informative. The annotators consisted of doctors of information technology, computer science, or public health research, as well as undergraduate and graduate students with extensive experience in social network usage. If a Weibo post met any of the keywords shown in Table 1, it was labeled as informative.

The annotators were given sufficient freedom to search for any information related to the pandemic and apply their judgment. In cases of controversy, the annotators used voting to determine

Table 4. Performance of different feature sets

Model	Acc	Pre	Rec	F1
MMS	0.69	0.71	0.66	0.69
MST	0.73	0.73	**0.77**	0.76
MMS +MST	**0.81**	**0.77**	0.76	**0.77**

the final result. For example, a message would be marked as informative only if there were five votes for PPPNN, PPPPN, or PPPPP, where P represents informative information and N represents non-informative information. The Cohen's kappa coefficient among the data annotators was 0.79, indicating a high level of agreement among multiple independent annotators (Lee et al., 2013). The final research data set comprised 2,863 informative samples and 9,737 non-informative samples.

Evaluation

Evaluation of Feature Extraction

Identifying information through situational awareness and extracting its feature vector from a vast network data are crucial for achieving accurate text classification. We extract topic-based features for the first time, based on statistical features, to capture the possible semantic attributes of the text. We evaluate a common SVM model to assess the effectiveness of topic-based features in information classification.

Statistical features entail linguistic, numeric, punctuation, and source-based features. We use MMS to represent the classification model utilizing solely statistical features. Meanwhile, topic-based features involve segregating manually marked text into informative and non-informative information and extracting topics through our LDA2vec topic model. We use MST to represent the classification models utilizing solely topic-based features. Table 4 illustrates the performance of different feature sets with the same classifier.

As can be seen from Table 4, the MST model performs better than the MMS model, indicating that topic-based features are more accurate than traditional statistical features. MST performs best regarding recall rates, at 77%, because traditional statistical features are surface features that are often more relevant to the users themselves, especially linguistic features. However, topic-based features are more relevant to what is being discussed and contribute to situational understanding and situational prediction.

Our combined model (MMS +MST) and MST model's F1-score is at least 7% superior to the MMS model, which only encompasses traditional features. This indicates the benefits of extraction based on topic features in identifying informative information. Regarding accuracy, our combined model outperformed all other models, surpassing MMS by 12%, 6%, and 8% in terms of accuracy, precision, and F1-score, respectively. Thus, extracting topic-based features and incorporating conventional statistical features significantly enhance the recognition of situational-awareness-based data.

Evaluation of the Balanced Data Set

In this study, we evaluate the efficacy of the Mahalanobis distance kernel-based FSVM classifier for eliminating non-informative data on a balanced data set. To further assess the model's performance, we select the traditional SVM and FSVM based on the Euclidean distance kernel as baseline models for comparison. The comparison results are shown in Table 5. It is evident that our framework significantly outperforms the latter two techniques.

Meanwhile, we compare our framework with other traditional text classification models and literature-based approaches. Table 6 presents the performance results, demonstrating that our framework outperforms other models in terms of evaluation metrics (achieving an 86% accuracy and

Table 5. Performance evaluation of different kernel classifiers on the balanced data set

Model	Acc	Pre	Rec	F1
SVM	0.81	0.76	0.76	0.77
FSVM	0.83	0.75	0.72	0.76
Our Framework	**0.86**	**0.79**	**0.81**	**0.81**

an 82% recall rate). Notably, among the language models, the lowest performing is the naive Bayes classifiers, whereas Xgboost's classification outcomes are the most outstanding. The method outlined by Misser et al. (2016) demonstrates favorable performance when taking high-dimensional vectors as inputs, returning an accuracy of 83% and an F1-score of 74%. However, these figures fall 3% and 7% behind our framework. These findings substantiate the efficacy and efficiency of our proposed model to a significant degree.

Evaluation of Unbalanced Data Set

Most data extracted from real microblogging lacks informative value, resulting in an extremely unbalanced data set. Hence, informative and non-informative information displays differing classification performance. In examining the FSVM classification model based on the Mahalanobis distance kernel on the unbalanced data set, all samples are considered in both the training and test data sets. The ratio of informative to non-informative information is set to 3.5:1 to reflect real-world conditions. The performance results are illustrated in Table 7.

Our proposed framework outperforms SVM and FSVM when informative samples are limited, increasing the recall rate by 4% and F1-score by over 3%. From the findings of classifying non-informative information in unbalanced data sets, there is a high proportion of non-informative information samples, and all models demonstrate better classification performance than informative information samples. Experimental results indicate that our FSVM model, which utilizes the Mahalanobis distance kernel, exhibits strong classification capabilities in practical scenarios and effectively filters out most unnecessary information.

We compare our framework's performance with other baseline models to demonstrate its superiority. Table 8 presents the comparison findings for the unbalanced data set. The baseline models exhibit a higher recall rate than the balanced data set when filtering out non-informative information. Their performance in identifying informative information is not satisfactory. However, despite a considerable difference in scale, the performance of our framework remains relatively stable. Simultaneously, our model significantly outperforms other baseline models regarding precision and

Table 6. Results of comparison with baseline model on the balanced data set

Model	Acc	Pre	Rec	F1
Random forest	0.73	0.69	0.75	0.72
Xgboost	0.77	**0.81**	0.77	0.79
Adaboost	0.68	0.74	0.68	0.70
LR	0.71	0.55	0.63	0.59
Naive Bayes	0.66	0.59	0.72	0.60
(Misser et al., 2016)	0.83	0.70	0.78	0.74
Our Framework	**0.86**	0.79	**0.82**	**0.81**

Table 7. Performance evaluation of different kernel classifiers on the unbalanced data set

Model	Informative information			Non-informative information		
	Rec	Pre	F1	Rec	Pre	F1
SVM	0.67	0.72	0.69	0.91	0.85	0.88
FSVM	0.67	**0.75**	0.70	0.9	0.86	0.89
Our Framework	**0.71**	0.74	**0.72**	**0.92**	**0.89**	**0.91**

F1-score when filtering non-informative information. Regarding identifying informative information, our model achieves the highest recall rate among all models, leading to a marked improvement in classification effectiveness. The filtering efficacy is optimal with a precision of 89%, and the F1-score is as high as 91%.

Parameter Evaluation

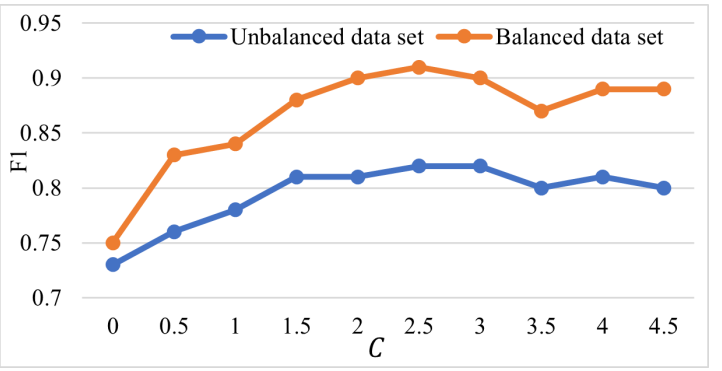
We investigate the impact of hyperparameter penalty coefficients C and kernel function variance *gamma* on the model's final performance. The parameter C serves as the regularization parameter, governing the model's complexity and its fit to training and test data. A higher C value imposes a stricter penalty for misclassification, favoring smaller margins to ensure accurate classification of more training samples. Conversely, a lower C value imposes a lighter penalty, allowing for larger margins and tolerance of misclassification. Specifically, a large C tends to yield more complex decision boundaries while a small C tends to produce simpler boundaries. The parameter *gamma* controls the precision of sample mapping in the feature space. A higher *gamma* shrinks the influence range of individual training samples, enhancing the model's fit to the training data but also increasing sensitivity to noise, potentially causing overfitting. Conversely, a lower *gamma* expands the influence range, reducing the model's fit to training data and possibly leading to underfitting. A large *gamma* tends to generate complex decision boundaries, possibly while a small *gamma* yields simpler boundaries. The optimal parameters are determined through grid search. The values of F1-score for both balanced and unbalanced data sets are presented in Figures 4 and 5, respectively, for the C and *gamma* .

As shown in Figure 4, our model's performance is impacted by the penalty coefficient C . The F1-score increases linearly for both unbalanced and balanced data sets when the penalty coefficient

Table 8. Results of comparison with baseline model on the unbalanced data set

Model	Informative Information			Non-informative Information		
	Rec	Pre	F1	Rec	Pre	F1
Random forest	0.65	0.8	**0.72**	0.9	0.8	0.86
Xgboost	0.55	0.78	0.64	**0.94**	0.85	0.89
Adaboost	0.58	**0.81**	0.69	0.91	0.82	0.86
LR	0.44	0.78	0.56	0.84	0.77	0.81
Naive Bayes	0.47	0.72	0.57	0.83	0.75	0.79
(Misser et al., 2016)	0.69	0.76	**0.72**	0.91	0.82	0.86
Our Framework	**0.71**	0.74	**0.72**	0.92	**0.89**	**0.91**

Figure 4. The value of F1-score with different C on balanced and unbalanced data sets



is below 2.5. It is important to note the sensitivity of the model to the penalty coefficient C. In contrast, the F1-score stabilizes when the penalty coefficient C exceeds 2.5.

As can be seen from Figure 5, for an unbalanced data set, F1-score significantly increases from 0.82 to 0.91 as *gamma* increases. When *gamma* is large, the effect of *gamma* on the unbalanced data set is often greater than that of balanced data sets. According to grid search, the optimal values of parameters are obtained: for an unbalanced data set, $C = 2.5$, *gamma* $= 0.4$; For a balanced data set, $C = 2.5$, *gamma* $= 0.25$.

SITUATIONAL AWARENESS ANALYSIS

Our study identifies informative public health emergencies based on the theory of situational awareness. In this section, we further analyze the differences between informative and non-informative texts in several perceptual elements.

Following the method by Sun et al. (2019), the text is based on a pretrained model to calculate the corresponding emotional indicators for informative and non-informative texts. In emotional factor classification, informative and non-informative texts are classified into positive, neutral, and negative emotions through three classifications. Table 9 provides examples of the corresponding textual content for each emotional value.

Figure 5. The value of F1-Score with different *gamma* on balanced and unbalanced data sets

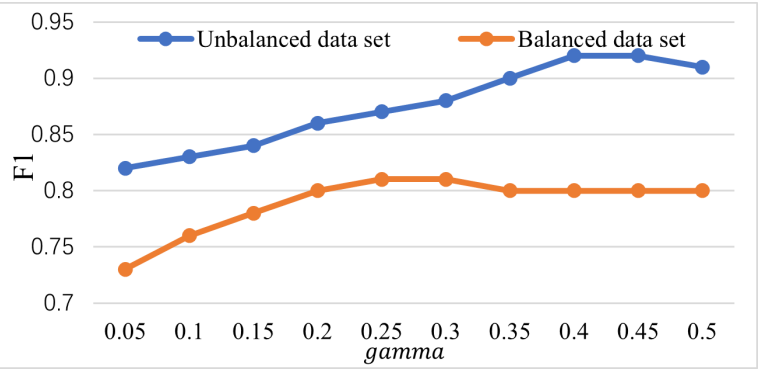


Table 9. Examples of text content corresponding to three types of affective values

Instance	Text content	Emotion classification	Is informative text
1	Today it has been raining non-stop COVID - 19 outbreak... I feel at a loss...	Negative	No
2	Trends of COVID-19 in Singapore (as of 21:00 on February 15), 5 new cases were confirmed, bringing the total number of confirmed cases to 72. # Singapore COVI -19 Pneumonia Vaccine Update # (as of February 15, 21:00)	Neutral	Yes
3	The dawn of the east, Modao Jun early. People are not old, scenery here alone. Come on, hang in there. Victory is in sight.	Positive	No

Table 10 shows the results of emotion classification of different information categories. Most of the COVID-19-related texts convey positive and neutral emotions. Positive emotional samples account for over 48%, neutral samples account for 32%, and negative emotional samples account for 20% of the total samples. This result is similar to the results of previous studies on general social network emotional expression, which indicates that the public's emotional tendency toward the COVID-19 epidemic was positive during the corresponding event stage of the data set, and the attitude toward the epidemic was positive. In the informative samples, positive emotional samples account for 24.2%, negative emotional samples account for 6.4%, and neutral samples account for 69.4%. Unlike the overall emotional tendencies, informative samples show obvious emotional neutrality characteristics, which is mainly due to the fact that informative samples appear in the form of news in practical environments. For most of such samples, the primary goal is to convey important information, so emotional factors are rarely added to the text. Positive informative samples are almost four times more than negative samples and, compared with non-informative samples, informative samples are more positive. The proportion of neutral emotional samples is significantly less for non-informative samples than informative samples, which is mainly because users bring personal emotional elements when posting related to Weibo.

We use feature reduction visualization based on t-SNE to show the classification results of sample sets under situational characteristics, as shown in Figure 6. In the situational awareness-based classification model, informative and non-informative samples can be easily distinguished. Situational awareness characteristics separate the clustering centers of the two types of samples, and informative samples appear in Quadrants 1 and 3, while non-informative samples mainly appear in Quadrants 2 and 4. Compared with the situational awareness-based classification results based on SVM, our framework has more explicit boundaries in classification tasks.

CONCLUSION

This study proposes an automated and comprehensive framework for identifying informative information related to public health emergencies on social media. First, we extract traditional statistical features, encompassing linguistic, numeric, punctuation, and source-based attributes. Second, we

Table 10. Emotion proportion of different information categories

Information type	Neutral	Positive	Negative
All samples(12600)	4,032	6,048	2,520
Informative sample(2863)	1988	693	182
Non-Informative sample(9373)	2044	5,355	2338

Figure 6. Situation awareness classification results based on our framework and SVM (a) FSVM with Mahalanobis distance kernel, (b) SVM

introduce a category-based LDA2vec model to extract topic features, suitable for unbalanced data sets. Finally, we propose an FSVM model based on the Mahalanobis distance kernel as a classifier to identify informative information. To evaluate our framework, we leverage two data sets and compare our model with several baselines. Experimental results demonstrate a significant improvement in the classification accuracy of informative and non-informative information, achieving an F1-score of 81% on balanced data sets and 91% on unbalanced data sets. Additionally, an analysis of the emotional tendencies of informative information reveals clear emotional clustering, indicating a distinct classification effect. Future work could involve expanding the framework to handle more complex and nuanced information, such as identifying misinformation or disinformation related to public health emergencies. Additionally, incorporating other types of data sources, such as images, videos, and audio clips, could provide richer context and improve the accuracy of information classification, especially in social media platforms where multimedia content is prevalent.

ACKNOWLEDGMENT

This paper is supported by the Development Foundation of Shanghai University of Finance and Economics Zhejiang College (No.2022FZJJ03), and the Fundamental Research Funds for the Central Universities (Grant Numbers 2023110139 and 2023110121).

REFERENCES

Alessa, A., & Faezipour, M. (2018). A review of influenza detection and prediction through social networking sites. *Theoretical Biology & Medical Modelling*, *15*(1), 1–27. doi:10.1186/s12976-017-0074-5 PMID:29386017

An, L., Yu, C., Lin, X., Du, T., Zhou, L., & Li, G. (2018). Topical evolution patterns and temporal trends of microblogs on public health emergencies: An exploratory study of Ebola on Twitter and Weibo. *Online Information Review*, *42*(6), 821–846. doi:10.1108/OIR-04-2016-0100

Bouzidi, Z., Amad, M., & Boudries, A. (2022). Enhancing warning, situational awareness, assessment and education in managing emergency: Case study of COVID-19. *SN Computer Science*, *3*(6), 454. doi:10.1007/s42979-022-01351-2 PMID:36035507

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS One*, *8*(12), e83672. doi:10.1371/journal.pone.0083672 PMID:24349542

Bruns, A. (2020). Crisis communication. In *The media and communications in Australia* (pp. 351–355). Routledge. doi:10.4324/9781003118084-31

Byrd, K., Mansurov, A., & Baysal, O. (2016, May). Mining Twitter data for influenza detection and surveillance. In *Proceedings of the international workshop on software engineering in healthcare systems* (pp. 43-49). ACM. doi:10.1145/2897683.2897693

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on Twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675-684). ACM. doi:10.1145/1963405.1963500

Chouigui, A., Khiroun, O. B., & Elayeb, B. (2018). A TF-IDF and co-occurrence based approach for events extraction from arabic news corpus. In *Natural language processing and information systems: 23rd international conference on applications of natural language to information systems, NLDB 2018, Paris, France, June 13-15, 2018* [Springer International Publishing.]. *Proceedings*, *23*, 272–280.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. doi:10.1518/001872095779049543

Freitas, J., & Ji, H. (2016, November). Identifying news from tweets. In *Proceedings of the first workshop on NLP and computational social science* (pp. 11-16). ACM. doi:10.18653/v1/W16-5602

Fu, X., Wang, Y., Li, M., Dou, M., Qiao, M., & Hu, K. (2020). Community evolutional network for situation awareness using social media. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 39225–39240. doi:10.1109/ACCESS.2020.2976108

Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, *56*(3), 1146–1165. doi:10.1016/j.ipm.2018.03.001

Hornmoen, H., Backholm, K., Frey, E., Ottosen, R., Reimerth, G., & Steensen, S. (2017). Key communicators' perspectives on the use of social media in risks and crises.

Huang, Q., Cervone, G., Xig, D., & Chang, C. (2015, November). DisasterMapper: A CyberGIS framework for disaster management using social media data. In *Proceedings of the 4th international ACM SIGSPATIAL workshop on analytics for big geospatial data* (pp. 1-6). ACM. doi:10.1145/2835185.2835189

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, *4*(3), 1549–1568. doi:10.3390/ijgi4031549

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), 1–38. doi:10.1145/2771588

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014, April). AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd international conference on world wide web* (pp. 159-162). ACM. doi:10.1145/2567948.2577034

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013a, May). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on world wide web* (pp. 1021-1024). ACM. doi:10.1145/2487788.2488109

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013b). Extracting information nuggets from disaster-related messages in social media. *Iscram*, *201*(3), 791–801.

Khatua, A., Khatua, A., & Cambria, E. (2019). A tale of two epidemics: Contextual word2vec for classifying Twitter streams during outbreaks. *Information Processing & Management*, *56*(1), 247–257. doi:10.1016/j.ipm.2018.10.010

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), e1500779. doi:10.1126/sciadv.1500779 PMID:27034978

Kumar, A., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research*, 1–32.

Lee, K., Agrawal, A., & Choudhary, A. (2013, August). Real-time disease surveillance using Twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1474-1477). ACM. doi:10.1145/2487575.2487709

Li, T.-H., Wang, Z., Lu, W., Zhang, Q., & Li, D.-F. (2022). Electronic health records based reinforcement learning for treatment optimizing. *Information Systems*, *104*, 101878. doi:10.1016/j.is.2021.101878

Liang, K. R., & Li, D.-F. (2020). A biobjective biform game approach to optimizing strategies in bilateral link network formation. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *52*(3), 1653–1662. doi:10.1109/TSMC.2020.3034480

Liu, X., Agarwal, S., Ding, C., & Yu, Q. (2016, June). An LDA-SVM active learning framework for web service classification. In 2016 IEEE international conference on web services (ICWS) (pp. 49-56). IEEE. doi:10.1109/ICWS.2016.16

Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., & da Silva Sousa, L. (2016). Tracking dengue epidemics using Twitter content classification and topic modelling. In *Current trends in web engineering: ICWE 2016 international workshops, DUI, TELERISE, SoWeMine, and Liquid Web, Lugano, Switzerland, June 6-9, 2016. Revised selected papers 16* (pp. 80-92). Springer International Publishing. doi:10.1007/978-3-319-46963-8_7

Nan, J.-X., Wei, L.-X., Li, D.-F., & Zhang, M.-F. (2024). A preemptive goal programming for multi-objective cooperative games: An application to multi-objective linear production. *International Transactions in Operational Research*. 31: 2427-2445

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014, May). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 376-385). IEEE. doi:10.1609/icwsm.v8i1.14538

Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011, March). Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. *In Proceedings of the ACM 2011 conference on computer supported cooperative work* (pp. 25-34). ACM. doi:10.1145/1958824.1958830

Rudra, K., Ganguly, N., Goyal, P., & Ghosh, S. (2018). Extracting and summarizing situational information from the Twitter social media during disasters. [TWEB]. *ACM Transactions on the Web*, *12*(3), 1–35. doi:10.1145/3178541

Salton, G., & Yu, C. T. (1973). On the construction of effective vocabularies for information retrieval. *SIGPLAN Notices*, *10*(1), 48–60. doi:10.1145/951787.951766

Scheele, C., Yu, M., & Huang, Q. (2021). Geographic context-aware text mining: Enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth*, *14*(11), 1721–1743. doi:10.1080/17538947.2021.1968048

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in Twitter to improve information filtering. *In Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 841-842). ACM. doi:10.1145/1835449.1835643

Unankard, S., Li, X., & Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web (Bussum)*, *18*(5), 1393–1417. doi:10.1007/s11280-014-0291-3

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. *In Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079-1088). ACM. doi:10.1145/1753326.1753486

Wang, Y., Li, X., & Mo, D. Y. (2020a, December). Personal health mention identification from tweets using convolutional neural network. *In 2020 IEEE international conference on industrial engineering and engineering management (IEEM)* (pp. 650-654). IEEE.

Wang, Y., Ruan, S., Wang, T., & Qiao, M. (2020b). Rapid estimation of an earthquake impact area using a spatial logistic growth model based on social media data. In *Social sensing and big data computing for disaster management* (pp. 68–87). Routledge. doi:10.4324/9781003106494-5

Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49–72. doi:10.1080/13658816.2017.1367003

Wiedmann, J. (2017, February). Joint learning of structural and textual features for web scale event extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). doi:10.1609/aaai.v31i1.10524

Xia, S., Zheng, S., Wang, G., Gao, X., & Wang, B. (2021). Granular ball sampling for noisy label classification or imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*. PMID:34460405

Yu, G. F., Li, D.-F., Liang, D. C., & Li, G. X. (2021). An intuitionistic fuzzy multi-objective goal programming approach to portfolio selection. *International Journal of Information Technology & Decision Making*, *20*(05), 1477–1497. doi:10.1142/S0219622021500395