Social Media Multimodal Information Analysis based on the BiLSTM-Attention-CNN-XGBoost Ensemble Neural Network

Ling Jixian[1], An Gang[2], Su Zhihao[3], Song Xiaoqiang[4]
School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia[1]
School of Languages, Literacies and Translation Universiti Sains Malaysia, Penang, Malaysia[2]
Computer Science College, Guangzhou College of Applied Science and Technology, Guangzhou, China[3]
Ningxia University, Ningxia China[4]

***Abstract*—Social media users internalise information in a multimodal context. Social media functions as a primary information source for disaster situational awareness encompassing texts, photographs, videos, and other multimodal information widely used in emergency management. Applying ensemble learning to social media sentiment analysis has garnered much scholarly attention, albeit with limited research on rescue and its sub-domain, which is characterised as a major complexity. A multimodal information categorisation model based on hierarchical feature extraction was proposed in this study. The information of multiple modes is first mapped to a unified text vector space in modelling the semantic content at the sentence and multimodal information levels in the multimodal information. Multiple deep learning (DL) models were subsequently applied to model the semantic content at the aforementioned levels. This study offers a BiLSTM-Attention-CNN-XGBOOST ensemble neural network model to acquire extensive multimodal information characteristics. Based on the empirical outcomes, this method precisely extracted multimodal information features with an accuracy exceeding 85% and 95% for Chinese- and English-language datasets, respectively.**

Keywords—Natural language processing; social media sentiment analysis; multimodal information processing; ensemble neural network; emergency management

I. Introduction

Social media platforms involving Weibo, WeChat, Facebook, Instagram, and Twitter have evolved into (1) key real-time information acquisition channels in disaster management and (2) information communication tools in physical and virtual contexts following the advent of the mobile Internet [1-3]. A vast amount of rescue and assistance information is promptly shared on digital platforms. Users post many real-time words, photographs, and videos about casualties, facility damage, and emergency assistance on social media [4]. In disaster management, efficiently identifying complex multimodal information on social media and emergency information has garnered much attention during emergencies. Such multimodal information proves crucial in managing emergencies [5]. Natural language processing (NLP) and computer vision technologies based on machine learning (ML) and DL significantly improved their performance and effect in multimodal data-processing tasks owing to breakthroughs in technologies (artificial intelligence and big data), which primarily catalysed multimodal information analysis in social media [6]. Research on social media multimodal information analysis in the context of emergencies is a relatively novel study topic that has garnered much scholarly interest and developed a theoretical and methodological system [7]. Current studies on social media multimodal information in emergencies provide theoretical and methodological support for social media multimodal information analysis for emergency management [8].

Before the study model development, specific limitations in using Twitter to monitor situational awareness in a disaster event were identified in multimodalities and social media network types. For example, Twitter users only represented some residents, as not everyone uses Twitter. Generally, some disadvantaged (low-income, low-education, and elderly) groups without the devices or motivation to access social media applications are less inclined to publish or receive disaster-relevant information [9]. Current works only utilised Facebook and Twitter data regarding the perception and humanitarian challenges identified on Twitter. Current research combining images and text to classify social media emergency management information has focused on English datasets and Twitter platforms [10], with less research on cross-social media platforms and cross-language datasets. In addition, the classification accuracy of current research needs to be improved [11].

This study employed data from WeiBo, WeChat, Facebook, and Twitter. The dataset includes English and Chinese. Recent Studies highlight the need for a multimodal and complex information-processing model that gathers, models, and employs feedback information in the social media interaction with users and continually iterates and optimises the information [12]. Such an approach could passively or actively detect errors, implement online learning and dynamic updating mechanisms based on the faults, and construct a set of self-learning frameworks to alleviate information processing issues in emergency management.

The main contribution of this paper is to use the Bi-LSTSM-Attention-CNN-XGBOOST ensemble neural network model to resolve the intricate processing of multimodal information in social media. The model uses recurrent neural networks to improve the adaptive and migration capabilities of the model. Multimodal information obtained from different platforms is first mapped to a unified text vector space to solve the cross-platform cross-language problem. Various DL models are used to model the semantic content at the sentence and multimodal information levels. This model is divided into six layers: (1) Embedding layer: maps words into low-dimensional vectors; (2) BiLSTM layer: generates deeper semantic vector to represent each word with BiLSTM network; (3) Attention layer: achieves the attention weight for all word; (4) Pooling layer: engage k-max pooling to extract the top 'k' words, which are ranked in step 3; (5) CNN layer: input to the CNN network in performing convolutional operations and extracting features through first k-word vectors; (6) X GBOOST layer: the feature vectors extracted by the final CNN are classified by the X GBOOST integrated classifier.

This paper covers six sections. In Section II related work is presented. Section III covers the method while Sections IV and V explain results and discussions, while Section VI concludes the overall contributions and suggested directions for future work

II. Related Work

Given the advancements in mobile internet, social media platforms (WeiBo, WeChat, and Twitter) have gained prominence in real-time information acquisition channels for disaster management and function as information communication tools in physical and virtual worlds. Only 2% of the extracted works directly encompassed the term 'multimodal' in the title or abstract. All the studies were published in the past five years. Notably, social media data on multiple modalities were analysed. A total of 216 articles potentially incorporating the idea of multimodal fusion accounted for 27% of all the documents. The most common publications involved information analysis of text and image modalities.

A study by [13] identified temporal changes in the social media sites extensively utilised for disaster recovery, their usage patterns by catastrophe type, and the geographic areas other studies have emphasised. Empirical outcomes on social media use in multiple disaster recovery aspects, such as (1) financial support and donations, (2) solidarity and social cohesion, (3) infrastructural services and post-disaster reconstruction, (4) socioeconomic and physical well-being, (5) information support, (6) mental health and emotional support, and (7) business and economic activities were also highlighted. Twitter users from Australia and beyond have aided bushfire victims through positive messaging regarding contributions, relief assistance, news updates, and animal welfare. Research by [14] examined how Twitter was utilised to publicise blaze recovery in Australia. Concerns about climate emergency and the perceived lack of political action entailed the primary sources of undesirable attitudes. In [15], researchers proposed fine-tunned BERT for multimodal analysis. While in [16], trustworthy summaries from crisis-related microblogs are predicted.

Modern deep learning methods were recommended by [17] to learn an integrated representation of social media data involving text and picture modalities. Specifically, a multimodal deep learning architecture with a modality-agnostic common representation was defined with convolutional neural networks. The proposed multimodal architecture outperformed the models developed using a single modality (text or image) based on studies that employed data from actual disasters. The model outperformed the image-only model by approximately 1% in all the informativeness task metrics and 2% in all humanitarian task measurements. Concerns about people's situational awareness during the occurrence of a natural disaster vary based on the stakeholder.

In regarding perception-, humanitarian-, and action-level situational awareness into account, [10] structured a Twitter-based analytic framework for damage estimation: the most prevalent use of Twitter in disaster management. Social media platforms could promptly ascertain damage to save people in danger, determine evacuation routes, and plan countermeasures for potential catastrophes. A study by [18] computed the extent to which volunteering content from one crisis is transferrable to another through language consistency analysis in volunteer- and donation-related social media material across 78 crises. Particular techniques were also provided to offer computational assistance in this emergency support role and establish semi-automated models in classifying social media information on volunteering and donating in the wake of a new crisis. Resultantly, the social media materials associated with volunteering and donations were sufficiently comparable between disasters and disaster types to justify transferring models across the disasters. These models were subsequently assessed with direct resampling procedures.

In comprehending the implications of social media to influence the relief and recovery process, the thematic and emotional nature of Twitter content discussing the 2019–2020 Australian bushfire disaster and its associated wildlife damage was discussed [19]. Thus, the value of social media (and Facebook in particular) has been demonstrated for knowledge sharing, volunteer coordination, fundraising, peer motivation, and accountability. Social media also makes it more challenging to distribute aid and coordinate relief efforts based on disinformation and duplication.

Following [20], past models' categorization of catastrophe signals primarily depends on unimodal techniques. Although specific approaches utilized multimodality to manage data, their accuracy could not be ascertained. A multimodal fusion strategy was developed to identify the relevant catastrophe photos from social media networks and merged the image and text information. The textual characteristics were subsequently elicited from social media with a FastText framework after the visual data were extracted through a deep learning technique. In classifying pertinent catastrophe photographs, a novel data fusion model combining linguistic and visual elements were structured. Well-executed experiments on the Crisis MMD dataset of actual disasters were also conducted.

An efficient DL algorithm was presented in [21] to manipulate multimodal information sources (words and photos) and disseminate helpful information during natural disasters. The programme divided the tweets into seven crucial and actionable groups, including reports of 'hurt or dead individuals' and 'infrastructure damage'. Based on research incorporating a benchmark dataset, integrating text and picture data from multiple sources proved more successful than employing data from a single source: one of the two in extracting pivotal information in crisis circumstances. Using visual and linguistic inputs, the recommended multimodal architecture in [22] classified damage-related postings using ResNet50 and BiLSTM recurrent neural networks using attention mechanisms. The MTLTS, the first end-to-end method for gathering reliable summaries from a substantial number of tweets on disasters, was introduced to supervise methodology and improve the applicability of solutions to unprecedented situations. This innovative approach involving the extractive document summarization technique summarised tweets on specified events. By concurrently learning the structural properties of information cascades and response stances, the credibility verifier is optimized with advanced components to detect rumours.

Information processing is crucial for large-scale data visualisation, which could be performed through data harmonisation [23] and implies the exact representation as multimodal data. Multimodal information-processing methods are currently categorised into classical and DL methods. Conventional methods primarily include support vector machines (SVM), naïve Bayes (NB), KNN, and LDA [24] topic model algorithms. Deep learning methods have recently undergone rapid development. For example, Mikolov et al. [25] suggested the Word2vec model for word quantisation in 2013, substituting the simple one-hot vectorisation method and resolving data sparsity issues. The CBOW and Skip-gram models in Word2vec were identified to train word vectors. Numerous optimisation methods were recommended for training details to enhance the training speed significantly. An attention mechanism method for deep learning in 2014 to extract essential information [26]. The RNN network-attention mechanism integration outperformed most single models within the image domain.

The author in [27] applied the long short-term memory (LSTM) network to the text classification field in the same year, thus mitigating the lack of contextual information in CNN. The training time is longer for long text datasets as LSTM takes the whole text as input without extracting vital information. A study by [28] suggested a recurrent convolutional neural network structure (RCNN) that connected BiLSTM and the maximum pooling layer. The BiLSTM network captured contextual information while the pooling layer extracted vital features. Hence BiLSTM is fully utilised with the integration of both structures. Extracting contextual information that only retains the pooling layer in the CNN part weakens the CNN's ability to extract features. Noda proposes a model for the multimodal robot in [29] and [30] proposes DeepCEP in 2019 with multimodal information streams and complex, spatial, and temporal dependencies.

The DL method of complex multimodal information based on the neural network has been widely acknowledged for its ability to simulate the cognitive function of the human brain. Regardless, this method (not a mathematical model based on the workings of the human brain) only denotes a formal mathematical description of the neuron structure and signal transmission method. It is also deemed challenging to eliminate the reliance on large-scale training samples. In 2020, Bejan described the MemoSys system submitted in Task 8 of SemEval 2020 to classify Internet memes sentiments [31]. In 2021, Zahera and colleagues recommended I-AID, a multimodal approach to automatically categorize tweets into multi-label information types [15].

Reference [32] highlighted various multimodal learning challenges and ways to train deep networks to learn features for task management. Cross-modality feature learning was also demonstrated, which allows for better features for a single modality (video) when other modalities (audio and video) are present when the features are being learned. The means to train a classifier with audio-only data to be tested with video-only data and evaluated on a specific task using shared representations between modalities was also denoted. Research by [33] proposed a model and neural network that demonstrated how image-text modelling could mutually learn word representations and image attributes. This methodology provides sentence descriptions for images without templates, structured prediction, or syntax trees in contrast to many other existing techniques.

Essentially, [34] developed and assessed advanced neural network techniques that incorporate input from several modalities to investigate the impact of complex interactions between textual, visual, and metadata on project success prediction. The method requires data gathered from the pre-posting profile to enable pre-posting prediction. Following [35], a multi-label and multimodal framework using text, audio, and visual input modalities was offered to categorise unbalanced data and automatically create static and temporal features using spatiotemporal deep neural networks. A weighted multi-label classifier functioned to manage data using non-uniform distributions.

A revolutionary deep dual recurrent encoder model in [36] was proposed to comprehend speech data fully and concurrently manipulate text data and audio signals. As spoken and musical content constitutes emotional dialogue, this model employed RNNs to encode audio and text sequences information before integrating it to determine the emotion class. Additionally, [37] proposed a deep multimodal attentive fusion (DMAF), a model for image-text sentiment analysis, to manipulate the discriminative features and internal connection between visual and semantic content using a hybrid fusion framework for sentiment analysis. The multimodal information proves crucial in emergency management. In the ‘scenario-response’ approach, social media provides additional data sources for emergency management. The information content in multiple modalities is interrelated, albeit with contextual differences.

In a text containing events, the content, image, and video details genuinely depict the scenes of the event. Such aspects complement and confirm one another to reflect the development status of real-world events. It is deemed pivotal to abstract, generalize, and integrate the complex social media multi-modal information from different levels and aspects and summarize and extract more accurate and comprehensive information than a single modality for users’ holistic and in-depth understanding of real-time emergencies to reduce emergency management ambiguities. Information extraction from social media is implied as a binary text classification issue with the labels 'relevant' and 'irrelevant' [15]. Nevertheless, effective methods to connect relevant postings to finer-grained classifications remain scarce. Such fine-grained labelling proves useful for crisis responders who need promptly provide catastrophe reactions and update vital information. Mainly, categorizing tweets on disasters with multiple labels facilitates the rapid identification of tweets with helpful information. The model performed optimally when breaking down disaster-related tweets into specific information types.

In research of [12] proposes a multimodal deep-learning framework to identify damage-related information. This framework combines multiple pre-trained unimodal convolutional neural networks that independently extract features from raw text and images, followed by a final classifier that labels posts based on both modalities, ultimately achieving an accuracy of about 92.6% in the English dataset. An approach to classifying thematic social media by integrating visual and textual information through a fused CNN architecture is proposed [38]. Specifically, two CNN architectures are used, targeting visual and textual information of social media posts, respectively. The outputs of the two CNNs, i.e., the features extracted from the social media posts representing visual and textual features, are further connected to form a fused representation. Research in [5] proposed a multimodal approach to identify disaster-related information content from Twitter streams using text and images. The method is based on long and short-term memory and VGG-16 networks, which significantly improve performance. In addition, [39] investigated the extent to which integrating multiple modalities is essential for classifying crisis content. In particular, a multimodal learning pipeline was designed to fuse textual and visual inputs, utilise both, and then classify that content based on a specified task. The average F1 performance in two critical tasks (relevance and humanitarian category classification) was 88.31%. A Python-based data pipeline application, SMDRM, for processing social media data points was proposed in [7].

However, this current work mainly explores the classification of English datasets, and more research needs to be done on Chinese contingency management datasets in social media. Notwithstanding, most multimodal information processing algorithms in social media, which are almost global and static, did not learn from failures and user input in real-time. In research of [40] proposed a multimodal network (MDMN)-based approach for rumour detection, including a text feature extractor, a visual feature extractor, and a fusion classification network, applying a multitask sharing layer, a task-specific converter encoder, and a selection layer to improve the diversity and stability of the text. Domain adaptation involves training adaptive models to extract visual representations. Adaptive models can encode task data better than fine-tuned pre-trained models. Then, experiments with two fusion strategies to fuse multimodal representations of multimodal datasets collected from tweets and microblogs show that the proposed MDMN can outperform the baseline approach. The decision-level fusion strategy achieves more than 92% Recall. In classifying emergency management information in social media, it is nearly impossible for existing algorithms to imitate the intelligent behaviour of human interaction and lifetime learning. It highlights the need for a multimodal and complex information-processing model that gathers, models, and employs feedback information in the social media interaction with users and continually iterates and optimises the information. Such an approach could passively or actively detect errors, implement online learning and dynamic updating mechanisms based on the faults, and construct a set of self-learning frameworks to alleviate information processing issues in emergency management.

All related literature shows that social media sentiment analysis plays a vital role in the advanced technological world. However, it is also found that the existing DL models contribute a lot. However, no such model developed which cross-platform, cross-lingual, multimodal and used multimodal social media data with high accuracy.

III. METHOD

This study proposed the Bi-LSTSM-Attention-CNN-XGBOOST ensemble neural network model (see Fig. 1) to address the complex processing of multimodal information in social media. The model manipulated ensemble learning as presented in Fig. 2. The information of multiple modes was initially mapped to a unified text vector space. Various deep learning models were used to model the semantic content at sentence and multimodal information levels. The MPOD textual data were used for model training and testing. As the name implies multimodality, English- and Chinese-language corpora from different social media platforms were selected for the recommended model. This model is divided into six layers:

1) Embedding layer: the word embedding matrix is regarded as the neural network model parameters and input for the BiLSTM neural network model to encode a sentence.

2) BiLSTM layer: BiLSTM constitutes two independent LSTMs that could summarise information from the forward and backward directions of a sentence. Essentially, the information stemming from both directions could be merged. At each time 't', the forward LSTM computed the hidden vector 'fht' based on the past hidden vector 'fht−1' and the input word embedding 'xt'. The backward LSTM calculated the hidden vector 'bht' based on the opposite (past) hidden vector 'bht−1' and the input word embedding 'xt'. Subsequently, the forward hidden vector 'fht' and backward hidden vector 'bht' were merged into the final hidden vector of the BiLSTM model. In this model, the parameters of two opposing directions proved independent albeit sharing the same word embedding of a sentence to generate a deeper semantic vector and represent each word with the BiLSTM network.

3) Attention layer: the self-attention mechanism functions to extract the more important information by providing them with a higher weight to elevate their significance. This layer primarily aimed to achieve the attention weight for all words and their ranking.

4) Pooling layer: this layer accepts the temporal sequence output by the LSTM layer and performs temporal max-pooling with sole reference to the non-masked portion of the sequence.

The pooling layer converted the entire variable-length hidden vector sequence into a single hidden vector for its output to be fed into the dense layer. The aforementioned layer engages the k-max pooling to extract the top k-words, which are ranked in step 3.

5) CNN layer: CNN layers are added to the front end, followed by the LSTM layers with a dense layer on the output. This layer provides input to the CNN network to perform convolutional operations and extract the key features with first k-word vectors.

6) X GBOOST layer: the XGBoost method enhanced the features extraction mechanism for DL models. This layer generates the feature vectors that are extracted by the final CNN and classified by the XGBOOST integrated classifier.

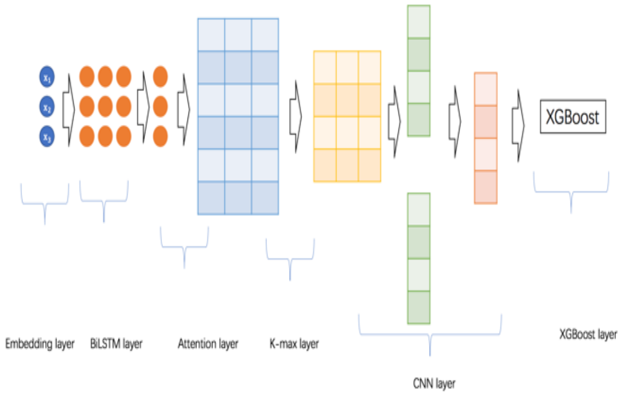

Fig. 1. The BiLSTM-attention-CNN-XGBOOST structure

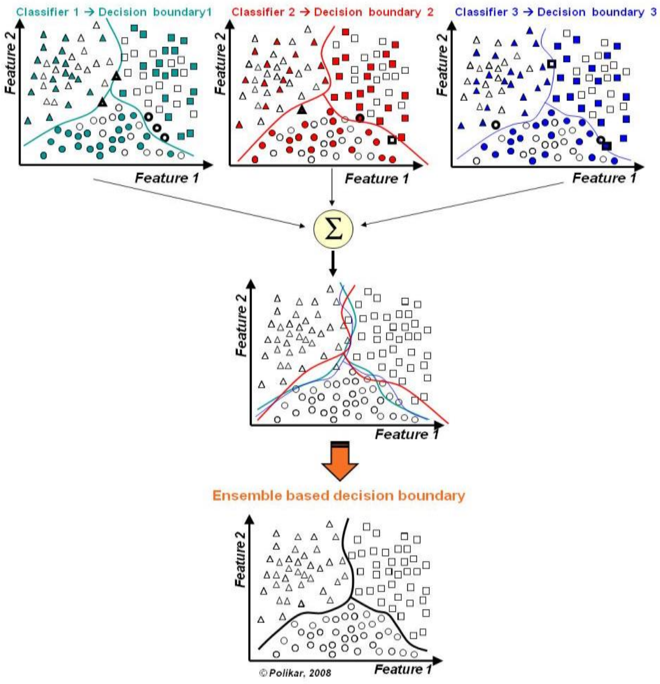
Classifier 1 → Decision boundary1
Classifier 2 → Decision boundary 2
Classifier 3 → Decision boundary 3
Feature 2
Feature 1
Σ
Ensemble based decision boundary
©Polikar, 2008

Fig. 2. The ensemble learning process [15]

A. Embedding Layer

Word embedding implies the representation method of word vectors in NLP. Although most conventional word vector representation methods utilise one-hot encoding, this method easily results in a high vector dimension. In the proposed study model, the Skip-gram model in Word2vec was utilised in word embedding to pre-train the words.

B. BiLSTM Layer

The LSTM layer utilises BiLSTM for deeper semantic vectors of words. This structure could combine the current word context to prevent future words in RNN from becoming more influential than their previous counterpart.

C. Attention Layer

Videos can be converted into text in the NLP domain. As the basic unit of text implies words, current models are primarily trained and operated on them. The attention mechanism suggested in this study mainly computed the attention of words in a given text. High levels of attention received by the word denoted its pivotal role in the task. Overall, extracting words with high attention could effectively classify multimodal information.

D. K-max Pooling Layer

The pooling layer in the NLP domain is primarily used to extract features and convert multimodal information of different lengths into fixed-length vectors. A common pooling operation denotes the max pooling operation, which retains one of the features. Meanwhile, the multimodal information ‘t’ typically denotes several words or their integration to express the meaning underlying the whole multimodal information. This model adopted k-max pooling, retained the first k-word vectors with larger attention scores, and obtained the combined feature vector ‘y’ of the first ‘k’ words. A larger weight implies higher levels of attention. Based on the outcomes derived from various experiments, the value of ‘k’ implied 8.

E. CNN Layer

This layer adopts the original Text CNN model for feature extraction. The first (network) layer takes the first k-word vectors retrieved by the k-max pooling layer as input. The second (convolution) layer performs convolution on input word vectors with several filters of varying sizes. The third (pooling) layer performs the highest number of pooling operations to create a novel feature vector.

F. XGBoost Layer

The BiLSTM-Attention-CNN-XGBoost ensemble neural network model proposed in this study implied the XGBoost ensemble classifier. If the length of a sentence is ‘n’, word embedding is performed through the pre-trained k-dimensional word vector with each word represented as a k-dimensional vector. Furthermore, the sentence could be converted into an n*k -dimensional data matrix as input. The convolutional layer employs 3*k, 4*k, and 5*k convolution kernels to extract features, pool the extracted features, and fully connect the features extracted by convolution kernels of varying sizes. Lastly, the multimodal information multi-classification processing is completed through the XGBoost classifier. In assuming that a sentence constitutes six words with each word denoting a four-dimensional word vector, the sentence could be characterised as a 6×4 data matrix as the convolution layer input through 3×4, 4× 4. The 5×4 convolution kernel performs feature extraction to obtain 4×1, 3×1, and 2×1 feature maps. A three-dimensional feature map is derived from maximum pooling while the XGBoost classifier is employed for classification processing.

IV. Experiment and Results

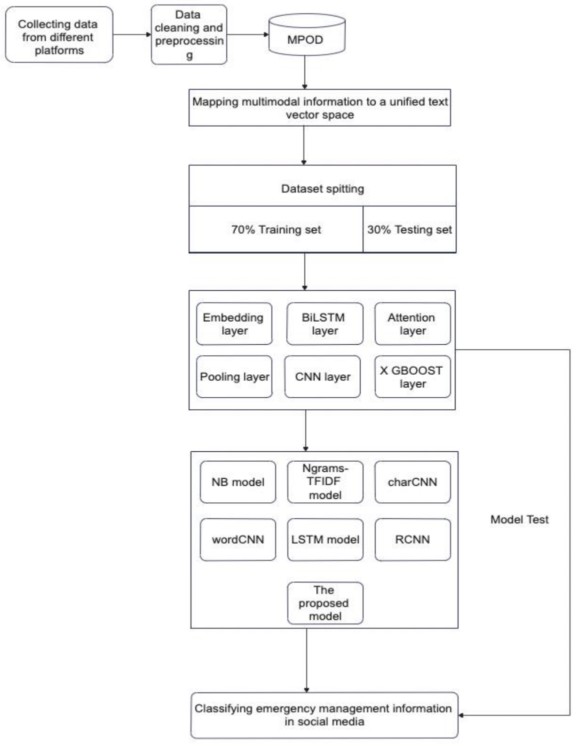

Fig. 3. The experiment framework

The experiment was conducted on ubuntu16.04 with Intel(R) Xeon(R) E5-265 as the CPU, 2.4GHz as the frequency, 32GB as the memory size, Python3.0 as the experimental programming language, and Tensorflow1.12.0 as the deep learning framework. The experimental dataset implied the news public opinion corpora in Weibo, Wechat, Twitter, and Facebook. As shown in Fig. 3, this self-collected dataset containing 6,700 text samples, 342 Image samples and 67 sounds samples was referred to as the multilingual public opinion dataset (MPOD). Notably, 70% of the samples were arbitrarily selected as a training set while the remaining 30% were chosen as a test set. The training set was manipulated to train the proposed model and evaluate it with the test set.

In Table I, the overall comparison with existing models implemented on the selected dataset are presented. The performance of the proposed model is more efficient and accurate then all with having 85 to 97%. The NB method performance varies from 78 to 93. Whereas N-gram with TF-IDF have 78-94, as well as charCNN have slightly incremental up to 95. The wordCNN have 95.66% and LSTM have 95.78% and last but not the least RCNN jumps to 96%.

TABLE I. The Comparison of Different Methods in MPOD

Method	**Chinese language set**		**English language set**	
	WeiBo	**WeChat**	**Twitter**	**Facebook**
NB model	84.12%	78.56%	93.21%	93.02%
Ngrams - TFIDF model	84.78%	78.94%	93.55%	94.67%
charCNN	85.23%	79.04%	93.67%	95.05%
wordCNN	84.97%	79.89%	94.87%	95.66%
LSTM model	85.05%	80.98%	95.67%	95.78%
RCNN	85.24%	81.45%	95.98%	96.02%
The proposed method	**87.64%**	**85.65%**	**98.34%**	**97.62%**

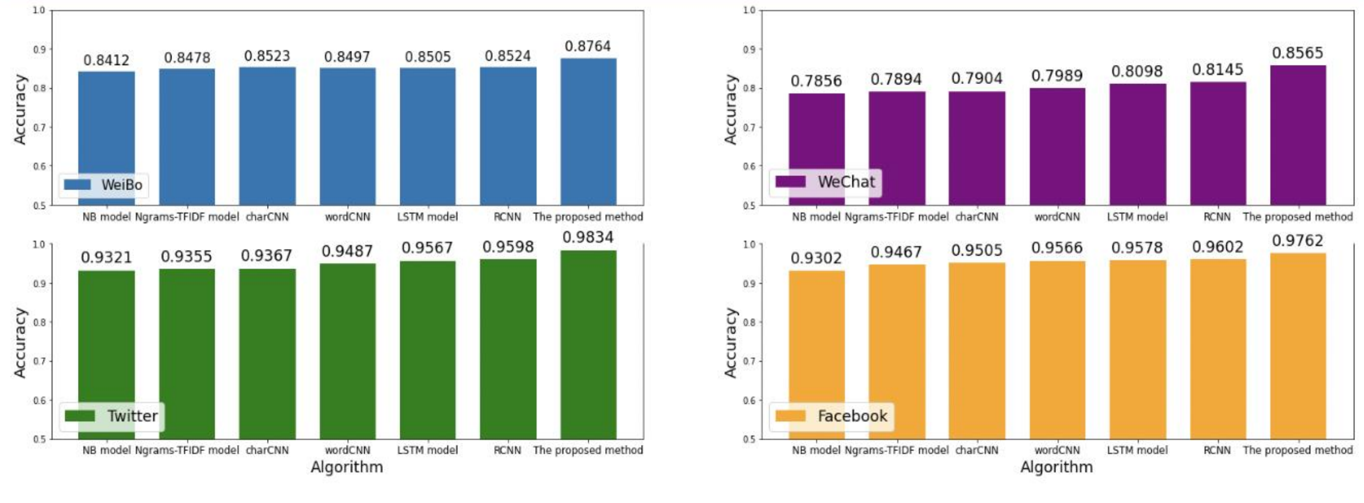

Fig. 4. The comparison of different methods in MPOD

As shown in Fig. 4, the BiLSTM-Attention-CNN-XGBOOST model proposed in this paper shows better accuracy than NB MODEL, Ngrams-TFIDF model, charCNN, wordCNN, LSTM model, RCNN model, regardless of Chinese or English data, whether from WeiBo, Twitter, Wechat or Facebook. The highest accuracy rate of Chinese language dataset is from WeiBo, reaching 87.64%, and the highest accuracy rate of English language dataset is from Twitter, reaching 98.34%.It can be clearly seen that the accuracy rate in the English data set is significantly higher than that in the Chinese data set by about 10%.In the English data set, Twitter performed better than Facebook, and in the Chinese data set, Weibo performed better than Wechat. Facebook and WeChat are biased towards acquaintance connections, while Twitter and Weibo are biased toward media connections.

V. Discussion

Six multimodal information classification methods were selected in this study to verify and discuss the model superiority. The conventional multimodal information classification method employed the NB and Ngrams-TFIDF models, while the deep learning method utilised the charCNN, wordCNN, LSTM, and RCNN models. Research by [32-37] similarly highlighted the multimodal model utilised in various textual, image and video datasets for different domains. Likewise, [15, 17, 20-22] employed multimodality to develop DL-based models in facilitating disaster management and recovery. Table I depicts the comparison of different classification algorithms and their effects on each model.

This study emphasised three multimodal information categories: (1) natural language, which can be both written and spoken; (2) visual signals that are typically represented by images or videos; (3) sounds that encode auditory and paralinguistic information, such as prosody and vocal expression signal. Following the data analysis in Table I, the classification effect of the study model on the same dataset is better for multimodal information. The two conventional methods, NB and Ngrams-TFIDF, depend on statistics, whereas charCNN, wordCNN, LSTM, and RCNN are classic neural network deep learning algorithms. Owing to the multi-layered nature of deep neural networks, a multimodal representation was constructed with several separate neural layers for each modality, followed by a hidden layer that projected the modalities into the joint space to employ neural networks.

The joint multimodal representation was conveyed through multiple hidden layers or directly used for prediction. In this study, the ensemble learning method served to extract the feature vector or latent semantic information of the multimodal counterpart by (1) studying the language cognitive mechanism of the brain and (2) analysing the relationship between cognitive mechanism and multimodal information computing methods. The four algorithms Bi-LSTM, attention mechanism, CNN, and XGBoost were integrated to extensively derive the advantages underlying each algorithm. For example, l-max pooling could reduce the number of model parameters that facilitates the mitigation of model overfitting issues. The attention mechanism could employ the human visual mechanism for intuitive interpretation, neural network interpretability, an in-depth understanding of the inner neural network mechanisms, and connect the neural network model structured based on the workings of the human brain. The weight-sharing network structure of CNN implies its mechanism.

The CNN resembles the biological neural network, thus reducing the complexity of the network model and number of weights. For example, XGBoost provides the model with a larger learning space and utilises feedback information for continuous model optimisation. As a result, the model classification accuracy in this study was 2% to 3% higher than that of the general method. This elevation resulted from using the k-max approach in the pooling layer in this study as the ‘k’ value substantially affected the experiment. Thus, several tests proved necessary to determine the most appropriate parameters and increase the experimental effect at the parameter adjustment stage of test training.

VI. Conclusion and Future Work

This study proposed a multimodal information classification method based on the hybrid BiLSTM-Attention-CNN-XGBoost neural network to classify multimodal information and maximise the benefits derived from each network model. Based on the experimental outcome, the classification accuracy was considerably improved upon adding the attention mechanism despite consuming much time as the experiment was performed on a single machine. The current study model would be tested on a distributed platform in the future to shorten the classification time. Based on the self-collected dataset called MPOD, the aforementioned hybrid neural network outperformed other models, such as CNN and Bi-LSTM, RCNN and Highways, LDA and SVM, attention mechanism, and GRU network. Further research would emphasise the multiple model fusion impacts on the experimental results with multimodal data.

References

[1] Gialampoukidis, Ilias, et al. "Multimodal Data Fusion of Social Media and Satellite Images for Emergency Response and Decision-Making." 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 2021.

[2] AlAbdulaali, A., Asif, A., Khatoon, S., & Alshamari, M.. Designing Multimodal Interactive Dashboard of Disaster Management Systems. Sensors, 22(11), 4292, 2022.

[3] Lin, J. Social Media Multi-modal Processing Mode for Emergency. In International Conference on Multi-modal Information Analytics (pp. 52-58). Springer, Cham, 2022.

[4] Dominguez-Péry, C., Tassabehji, R., Vuddaraju, L. N. R., & Duffour, V. K.. Improving emergency response operations in maritime accidents using social media with big data analytics: a case study of the MV Wakashio disaster. International Journal of Operations & Production Management, 2021.

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

[6] Zhang, C., Yang, Z., He, X., & Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3), 478-493, 2020.

[7] Lorini, V., Panizio, E., & Castillo, C. SMDRM: A Platform to Analyze Social Media for Disaster Risk Management in Near Real Time. In Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media. Retrieved from https://doi. org/10.36190, 2022.

[8] Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. Information Processing & Management, 57(5), 102261,2020.

[9] Ruiz Soler J. Twitter research for social scientists: A brief introduction to the benefits, limitations and tools for analysing Twitter data[J]. 2017.

[10] Zhai, W. A multi-level analytic framework for disaster situational awareness using Twitter data. Computational Urban Science, 2(1), 1-15, 2022.

[11] Zahera, H. M., Elgendy, I. A., Jalota, R., & Sherif, M. A. Fine-tuned BERT Model for Multi-Label Tweets Classification. In TREC pp. 1-7, 2019, November.

[12] Alam, Firoj, et al. "Deep learning benchmarks and datasets for social media image classification for disaster response." 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2020

[13] Ogie, R. I., James, S., Moore, A., Dilworth, T., Amirghasemi, M., & Whittaker, J. Social media use in disaster recovery: A systematic literature review. International Journal of Disaster Risk Reduction, 102783, 2022.

[14] Dahal, L., Idris, M. S., & Bravo, V. "It helped us, and it hurt us" The role of social media in shaping agency and action among youth in post-disaster Nepal. Journal of Contingencies and Crisis Management, 29(2), 217-225, 2021.

[15] Zahera, H. M., Jalota, R., Sherif, M. A., & Ngomo, A. C. N. I-AID: Identifying Actionable Information from Disaster-related Tweets. IEEE Access, 9, 118861-118870, 2021.

[16] Mukherjee, R., Vishnu, U., Peruri, H. C., Bhattacharya, S., Rudra, K., Goyal, P., & Ganguly, N. MTLTS: A Multi-Task Framework To Obtain Trustworthy Summaries From Crisis-Related Microblogs. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining ,pp. 755-763, 2022, February.

[17] Ofli F, Alam F, Imran M. Analysis of social media data using multimodal deep learning for disaster response[J]. arXiv preprint arXiv:2004.11838, 2020.

[18] Mousavi, P., & Buntain, C."Please Donate for the Affected": Supporting Emergency Managers in Finding Volunteers and Donations in Twitter Across Disasters. In Proceedings of the 19th ISCRAM Conference–Social Media for Crisis Management 605-622, 2022.

[19] Willson, G., Wilk, V., Sibson, R., & Morgan, A. Twitter content analysis of the Australian bushfires disaster 2019-2020: futures implications. Journal of Tourism Futures, 2021.

[20] Zou, Z., Gan, H., Huang, Q., Cai, T., & Cao, K. Disaster Image Classification by Fusing Multimodal Social Media Data. ISPRS International Journal of Geo-Information, 10(10), 636, 2021.

[21] Ahmad, Z., Jindal, R., Mukuntha, N. S., Ekbal, A., & Bhattachharyya, P. Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. Expert Systems with Applications, 195, 116626, 2022.

[22] Hossain, E., Hoque, M. M., Hoque, E., & Islam, M. S. A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts. IEEE Access, 10, 46538-46551, 2022.

[23] Kumar G, Basri S, Imam A A, et al. Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review[J]. Applied Sciences, 2021, 11(17): 8275.

[24] Wang Jinhua, Yu Hui, Chan Wen, et al Automatic text classification technology based on KNN + hierarchical SVM [J] Computer applications and software, 2016, 33 (2): 38-41

[25] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.

[26] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.

[27] ZhouCT , SunCL , LiuZY , etal . AC-LSTM neural network for text classification[EB]. arXiv:151.08630, 2015.

[28] Chen Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.

[29] Noda K, Arie H, Suga Y, et al. Multimodal integration learning of robot behavior using deep neural networks[J]. Robotics and Autonomous Systems, 2014, 62(6): 721-736.

[30] Tianwei Xing; Marc Roig Vilamala; Luis Garcia; Federico Cerutti; Lance M. Kaplan; Alun D. Preece; Mani B. Srivastava; DeepCEP: Deep Complex Event Processing Using Distributed Multimodal Information, 2019 IEEE INTERNATIONAL CONFERENCE ON SMART COMPUTING (SMARTCOMP), 2019. (IF: 3)

[31] Bejan I. MemoSYS at SemEval-2020 task 8: Multimodal emotion analysis in memes[C]//Proceedings of the Fourteenth Workshop on Semantic Evaluation. 2020: 1172-1178.

[32] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. Multimodal deep learning. ICML.2011.

[33] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models[C]. International conference on machine learning. PMLR, 2014: 595-603.

[34] Cheng C, Tan F, Hou X, et al. Success Prediction on Crowdfunding with Multimodal Deep Learning[C]. IJCAI. 2019: 2158-2164.

[35] Pouyanfar S, Wang T, Chen S C. A multi-label multimodal deep learning framework for imbalanced data classification[C]. 2019 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, 2019: 199-204.

[36] Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text[C]//2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018: 112-118.

[37] Huang F, Zhang X, Zhao Z, et al. Image–text sentiment analysis via deep multimodal attentive fusion[J]. Knowledge-Based Systems, 2019, 167: 26-37.

[38] Huang X, Li Z, Wang C, et al. Identifying disaster related social media for rapid response: a visual-textual fused CNN architecture[J]. International Journal of Digital Earth, 2019.

[39] Afyouni I, Al Aghbari Z, Razack R A. Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey[J]. Information Fusion, 2022, 79: 279-308.

[40] Zhou, H., Ma, T., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. . MDMN: Multi-task and Domain Adaptation based Multi-modal Network for early rumor detection. Expert Systems with Applications, 2022, 195, 116517.