results indicate that the ranking using multiple features perform better (i.e., match better with the actual risk ranking) compared with using any single feature.

Event	Sentiment Ranking	Risk Expression Ranking	Influence Ranking	Predicted Risk Ranking	Actual Risk Ranking
Earthquake in Luding	3	4	1	Very high	1
Flush flood in Datong Mountain	1	1	3	Higher	2
Rainstorm in Fujian, Jiangxi and Hunan	2	3	4	Higher	3
Drought in Yangtze River	4	2	2	Neutral	4
Low temperature in South China	6	5	6	Low	5
Rainstorm in Liaoning	7	8	7	Lower	6
Typhoon Chaba	8	6	5	Lower	7
Rainstorm in Sichuan	5	7	8	Very low	8

Table 2.6 Comparison with Matrix Grading Risk Ranking.

2.4.2. Evaluation of the sentiment analysis method

As discussed, we use RoBERTa to compute the two-dimension sentiment for each post in our model. Several deep learning language models (LM) are chosen as the benchmarks for comparison. The first one is TextCNN, a convolutional neural network (CNN) for text classification. It consists of an embedding layer, a CNN layer, and a softmax layer. The embed layer translates each word of an input sentence into a continuous embedding vector, and the CNN layer conducts convolution operations with max-pooling (Gong & Ji, 2018). The softmax layer then calculates the output as the probability distribution over labels (Kim, 2014). Compared with traditional machine learning methods such as support vector machine (SVM), TextCNN achieves state-of-the-art performance in the sentiment classification task.

The second benchmark is TextRCNN, a variant of TextCNN which combines a bi-directional recurrent neural network (RNN) and a max-pooling layer from CNN. The design of TextRCNN could capture contextual information with the recurrent structure and constructs the text representation with a convolutional network (Lai, Xu, Liu, & Zhao, 2015). TextRCNN outperforms CNN and RNN in various NLP tasks like sentiment analysis.

The third benchmark is TextRNNAtt (TextRNN with attention), an RNN network with an attention layer. The attention mechanism calculates the weight of each word in an input sentence and constructs a sentence representation (Yang et al., 2016). TextRNN could capture important words and outperform traditional SVM, RNN, and CNN networks by combining an attention layer in various review and sentiment datasets.

The fourth benchmark is DPCNN (deep pyramid CNN), which alternates between a convolution block and a downsampling layer over and over, leading to a deep network in which internal data shrinks in a pyramid shape (Johnson & Zhang, 2017). The pyramid architecture increases the network depth with less computation complexity. DPCNN outperforms the previous best models, including TextRCNN, on six benchmark sentiment and topic classification datasets (Johnson & Zhang, 2017).

TextCNN, TextRCNN, TextRNNAtt and DPCNN are four widely used deep learning benchmarks. By comparing against such representative benchmarks, we could better evaluate the performance of LLM (i.e., RoBERTa) and provide practical implications in the risk-related sentiment analysis task.

In each evaluation, two models, one for valence and one for arousal, are trained separately. We randomly split the dataset into text and training datasets and repeated the process ten times for the significance test. We choose three commonly used metrics, namely mean squared error (MSE), root-mean-square error (RMSE), and mean absolute error (MAE), as evaluation metrics. Table 2.7 shows the comparison results of our proposed method and the benchmarks.

	Valence			**Arousal**		
	MSE	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**
TextCNN	1.409	1.187	0.940	2.010	1.418	1.140
TextRCNN	1.076	1.037	0.792	1.801	1.342	1.049
TextRNNAtt	0.872	0.934	0.735	1.981	1.407	1.104
DPCNN	0.911	0.954	0.739	1.764	1.328	1.062
RoBERTa (Ours)	**0.724**	**0.851**	**0.642**	**1.327**	**1.152**	**0.833**

Table 2.7 Sentiment Analysis Model Performance.

According to Table 2.7, the proposed LLM (i.e., RoBERTa) shows the lowest MAE on the test dataset. Compared with benchmarks such as TextCNN and TextRCNN, RoBERTa achieves significantly better results (p < 0.05) in both valence and arousal metrics. Another finding is that RNN-related methods (i.e., TextRCNN and TextRNNAtt) outperform CNN-related methods (i.e., TextCNN and DPCNN) in most metrics. The finding shows that capturing sequence information is critical in predicting risk-related features such as valance and arousal. LLM has a superior ability to complete risk-related sentiment classification tasks as a variant of sequence networks with deep transformer architectures.

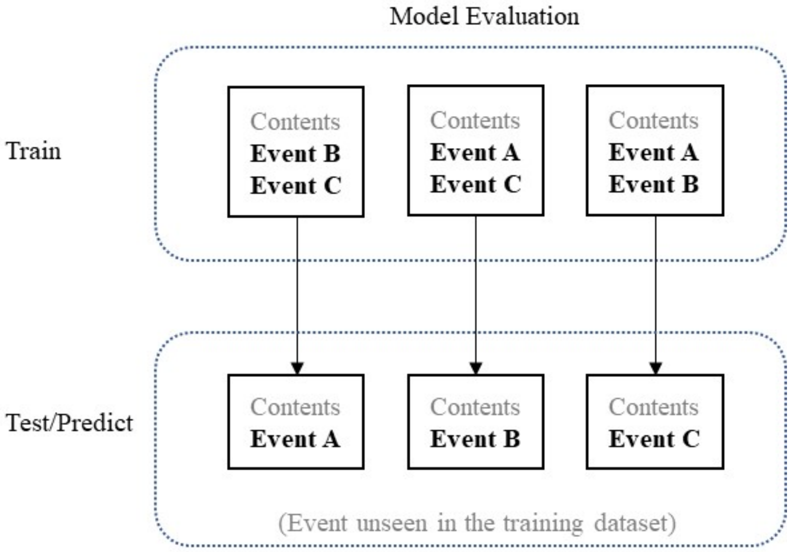

Figure 2.3. Cross-validation of language models and application.

As discussed in the previous section, the dataset used for pre-training the RoBERTa model for our context contains three events (i.e., A, B, and C). To evaluate the proposed LLM's ability to predict unseen events, we conduct a 3-fold cross-validation at the event level. In each fold, we treat data from two events as the training dataset and treat data from the remaining event as the testing dataset. For example, we train the proposed LLM on events B and C to predict the valance and arousal values for the posts in event A. The process is shown in Figure 2.3.

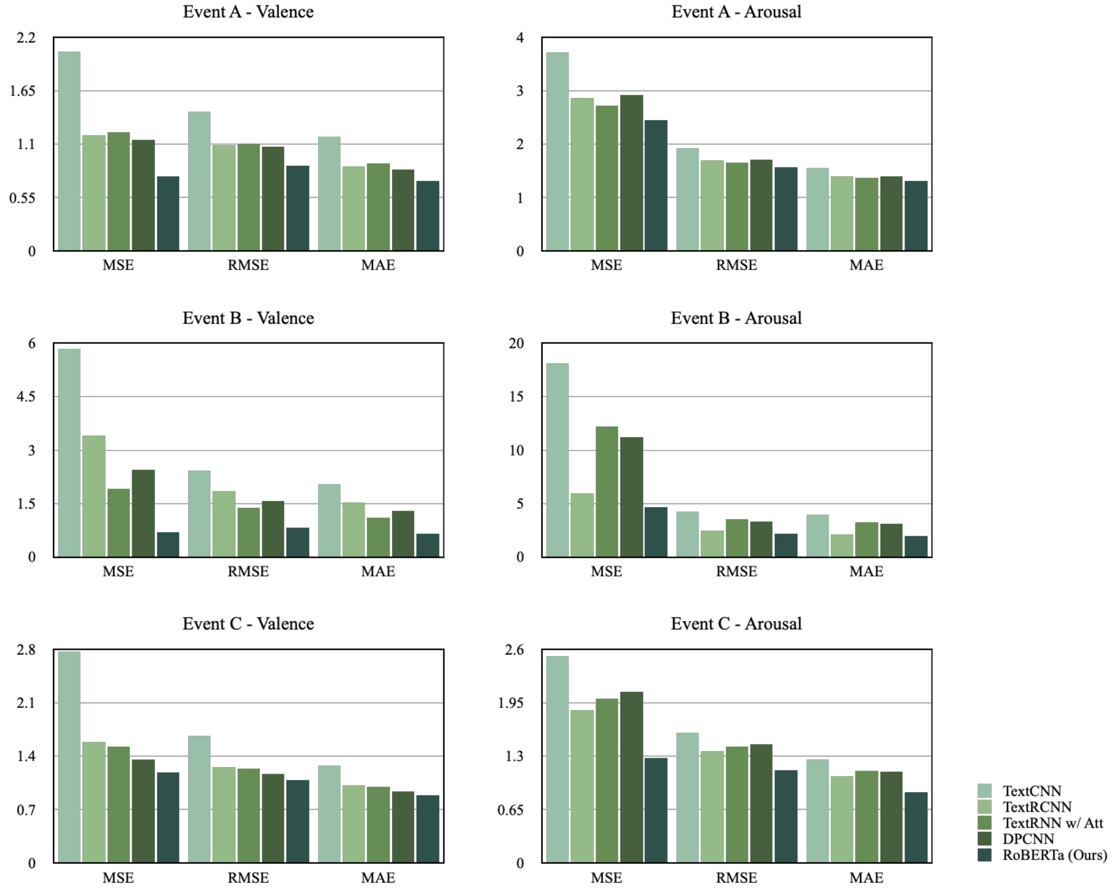

Figure 2.4. Cross-validation using unseen events.

Figure 2.4 shows the performance of LLM on predicting unseen event A, event B, and event C, along with the same benchmarks in the former experiment. The detailed performance results are provided in Table 2.8. The experiment results show that the proposed LLM still achieves better prediction performance than benchmarks. Moreover, when dealing with unseen events, the performance of several benchmarks is unstable. For example, for event B, the performance of TextCNN and TextRNNAtt in arousal are close to random guess (MSE is greater than 10). However, the prediction results of the proposed LLM are stable for all three events, demonstrating its ability to predict unseen events and its suitability for our risk assessment task.

Event A						
	Valence			Arousal		
Model	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**
TextCNN	2.048	1.431	1.174	3.709	1.926	1.554
TextRCNN	1.193	1.092	0.867	2.855	1.690	1.399
TextRNNAtt	1.218	1.104	0.901	2.716	1.648	1.360
DPCNN	1.143	1.069	0.838	2.921	1.709	1.396
RoBERTa	**0.769**	**0.877**	**0.717**	**2.453**	**1.566**	**1.308**

Event B						
	Valence			Arousal		
Model	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**
TextCNN	2.768	1.664	1.276	18.103	4.255	3.925
TextRCNN	1.589	1.260	1.015	5.940	2.437	2.125
TextRNNAtt	1.527	1.236	0.994	12.223	3.496	3.274
DPCNN	1.359	1.166	0.940	11.220	3.350	3.134
RoBERTa	**1.191**	**1.091**	**0.887**	**4.682**	**2.164**	**1.945**

Event C						
	Valence			Arousal		
Model	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**
TextCNN	5.841	2.417	2.041	2.513	1.585	1.258
TextRCNN	3.405	1.845	1.518	1.860	1.364	1.054
TextRNNAtt	1.908	1.381	1.098	1.999	1.414	1.123
DPCNN	2.440	1.562	1.297	2.079	1.442	1.110
RoBERTa	**0.694**	**0.833**	**0.644**	**1.279**	**1.131**	**0.861**

Table 2.8 Cross-validation using unseen events.

2.5. Discussions

2.5.1. Research Contributions

This research proposes a model to conduct risk assessment in the field of public safety based on social media information. There are four main research contributions. First, we are among the first to apply LLM on social media information under the context of disaster risk management. We demonstrate the superior performance of using LLM to analyze disaster-related data. By applying RoBERTa fine-tuned by a disaster-related dataset, the two-dimension disaster sentiments, valence and arousal, are added as an important measure in the risk ranking model.

Second, we utilize features that are calculated by various aggregated approaches including learning-based algorithms, lexicon-based computation and experience-based evaluation, in disaster management. The ranking comparison between the single feature ranking results and the aggregated results of combining LLM, lexicon-based text analysis, and expert knowledge shows that the proposed model outperforms single-feature rankings.

Third, we propose an effective risk assessment model, which is a novel methodology that could be applied to disaster risk ranking. This research has demonstrated the feasibility of using updated social media data to predict the range of consequences of disaster events. In addition, the validity of the matrix grading is illustrated. The proposed model of risk assessment for risk events is consistent with the actual losses caused by the corresponding disaster events. The result shows that the proposed model is effective and adequate for assessing the risks of disaster events through social media posts.

2.5.2. Practical Implications

Our research has important practical implications for government and policy makers. First, we demonstrate the advantages of utilizing social media data in disaster information management. Disaster risk ranking is determined by using creator information including verification status and number of followers, information related to the disaster post including retweet number and number of likes, and the text itself. Our model provides an early risk assessment for government to make more timely