CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding

Junhua Liu1,2, Trisha Singhal1, Lucienne T.M. Blessing1, Kristin L. Wood1, Kwan Hui Lim1

j@forth.ai, {trisha_singhal, lucienne_blessing, kristinwood, kwanhui_lim}@sutd.edu.sg

1Singapore University of Technology and Design
2ProQod Singapore

ABSTRACT

Classification of crisis events, such as natural disasters, terrorist attacks and pandemics, is a crucial task to create early signals and inform relevant parties for spontaneous actions to reduce overall damage. Despite the crises, such as natural disaster, that can be predicted by professional institutions, certain events are first signaled by everyday citizens, i.e., civilians, such as the recent COVID-19 pandemics. Social media platforms such as Twitter often expose firsthand signals on such crises through high volume information exchange. In the case of Twitter, this corresponds to on average over half a billion tweets posted daily. Prior works proposed various crisis embeddings and classification using conventional Machine Learning and Neural Network models. However, none of the works perform crisis embedding and classification using state of the art attention-based deep neural networks models, such as Transformers and document-level contextual embeddings. This work proposes CrisisBERT, an end-to-end transformer-based model for two crisis classification tasks, namely crisis detection and crisis recognition, which shows promising results across accuracy and f1 scores. The proposed model demonstrates superior robustness over various benchmarks, as it shows marginal performance compromise while extending from 6 to 36 events with only 51.4% additional data points. We also propose Crisis2Vec, an attention-based, document-level contextual embedding architecture for crisis embedding, which achieves better performance than conventional crisis embedding methods such as Word2Vec and GloVe. To the best of our knowledge, our works are first to propose using transformer-based crisis classification and document-level contextual crisis embedding in the literature.

Keywords Crisis Classification · Contextual Crisis Embedding · CrisisBERT · Crisis2Vec · Tweets

1 Introduction

Crisis-related events, such as earthquakes, hurricanes and train or airliner accidents, often stimulate a sudden surge of attention and actions from both media and the general public. Despite the fact that crises, such as natural disasters, can be predicted by professional institutions, certain events are first signaled by everyday citizens, i.e., civilians. For instance, the recent COVID-19 pandemics was first informed by general public in China via Weibo, a popular social media site, before pronouncements by government officials.

Social media sites have become centralized hubs that facilitate timely information exchange across government agencies, enterprises, working professionals and the general public. As one of the most popular social media sites, Twitter enables users to asynchronously communicate and exchange information with tweets, which are mini-blog posts limited to 280 characters. There are on average over half a billion tweets posted daily. Therefore, one can leverage on such high volume and frequent information exchange to expose firsthand signals on crisis-related events for early detection and warning systems to reduce overall damage and negative impacts.

Event detection from tweets has received significant attention in research in order to analyze crisis-related messages for better disaster management and increasing situational awareness. Several recent works studied various natural crisis events, such as hurricanes and earthquakes, and artificial disasters, such as terrorist attacks and explosions.
These works focus on binary classifications for various attributes of crisis, such as classifying source type, predicting relatedness between tweets and the crises, and assessing informativeness and applicability [6, 7, 8]. On the other hand, several works proposed multi-label classifiers on affected individuals, infrastructure, casualties, donations, caution, advice, etc. [9, 10]. Crisis recognition tasks are likewise conducted such as identifying crisis types, i.e., hurricanes, floods and fires [11, 12].

Machine Learning-based models are commonly introduced in performing the above mentioned tasks. Conventional linear models such as Logistic Regression, Naive Bayes and Support Vector Machine (SVM) are reported for automatic binary classification on informativeness [13] and relevancy [8], among others. These models were implemented with pre-trained word2vec embeddings [14]. Several unsupervised approaches are also proposed for classifying crisis-related events, such as the CLUSTOP algorithm utilizing Community Detection for automatic topic modelling [15]. A transfer-learning approach is also proposed [16], though its classification is only limited to two classes. The ability for cross-crisis evaluation remains questionable.

More recently, numerous works proposed Neural Networks (NN) models for crisis-related data detection and classification. For instance, ALRashdi and O’Keefe investigated two deep learning architectures, namely Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) using domain-specific and GloVe embeddings [17]. Nguyen et al. propose a CNN-based classifier with Word2Vec embedding pretrained on Google News [14] and domain-specific embeddings [18]. Lastly, parallel CNN architecture was proposed to detect disaster-related events using tweets [19, 20].

While prior works report remarkable performance on various crisis classification tasks using NN models and word embeddings, no studies are found to leverage the most recent Natural Language Understanding (NLU) techniques, such as attention-based deep classification models [21] and document-level contextual embeddings [22], which reportedly improve state-of-the-art performance for many challenging natural language problems from upstream tasks such as Named Entity Recognition and Part of Speech Tagging, to downstream tasks such as Machine Translation and Neural Conversation.

1.1 Main Contributions

This work focuses on deep attention-based classification models and document-level contextual representation models to address two important crisis classification tasks. We study recent NLU models and techniques that reportedly demonstrated drastic improvement on state-of-the-art and localize for domain-specific crisis related tasks.

Overall, our main contribution of this work includes:

- proposing CrisisBERT, an attention-based classifier that improves state-of-the-art performance for both crisis detection and recognition tasks;
- demonstrating superior robustness over various benchmarks, where extending CrisisBERT from 6 to 36 events with 51.4% of additional data points only results in marginal performance decline, while increasing crisis case classification by 500%;
- proposing Crisis2Vec, a document-level contextual embedding approach for crisis representation, and showing substantial improvement over conventional crisis embedding methods such as Word2vec and GloVe . . .

To the best of our knowledge, this work is the first to propose a transformer-based classifier for crisis classification tasks. We are also first to propose a document-level contextual crisis embedding approach.

2 Attention-based Neural Crisis Classifier

In this section, we discuss the recent works that propose various machine learning approaches for crisis classification tasks. While these works report substantial improvement in performance over prior works, none of the works use state of the art attention-based models, i.e., Transformers [21], to perform crisis classification tasks. We propose CrisisBERT, a transformer-based architecture that builds upon a Distilled BERT model, fine-tuned by large-scale hyper-parameter search.

2.1 Conventional Crisis classifiers

Various works propose linear classifiers for crisis-related events. For instance, Parilla-Ferrer et al. proposed an automatic binary classification, based on informative and uninformative tweets using Naive Bayes and Support Vector Machine (SVM) [13]. A SVM with pretrained word2vec embeddings approach was also proposed [14].

Besides linear models, recent works also propose deep learning based methods with different neural network architectures. For instance, ALRashdi and O’Keefe investigated Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural
Figure 1: Illustration of CrisisBERT and Crisis2Vec models. During the tokenization phase, word-level and subword-level embeddings are created to produce context embedding for each input word. The vectors are then prefixed with a class header (shown as [CLS] in the figure) and passed into a DistilBERT model. Since we are performing classification task, the CLS token vector, i.e. the first output vector, is then passed into a linear classifier for detection or recognition task, whereas the remainder of the output vectors are average-pooled to create Crisis2Vec embeddings.

Network (CNN) models using domain-specific and GloVe embeddings [23]. Nguyen et al. proposed a CNN model to classify tweets to get information types using Google News and domain-specific embeddings [18].

2.2 Transformer

In 2017, Vaswani et al. from Google introduced Transformer [21], a new category of deep learning models which are solely attention-based and without convolution and recurrent mechanisms. Later, Google proposed the Bidirectional Encoder Representations from Transformers (BERT) model [24] which drastically improved state-of-the-art performance for multiple challenging Natural Language Processing (NLP) tasks. Since then, multiple transformer-based models have been introduced, such as GPT [25] and XLNet [26], among others. Transformer-based models were also deployed to solve domain specific tasks, such as medical text inferencing [27] and occupational title embedding [28], and demonstrated remarkable performance.

The Bidirectional Encoder Representation of Transformer (BERT), for instance, is a multi-layer bidirectional Transformer encoder with attention mechanism [24]. The proposed BERT model has two variants, namely (a) BERT Base, which has 12 transformer layers, a hidden size of 768, 12 attention heads, and 110M total parameters; and (b) BERT Large, which has 24 transformer layers, a hidden size of 1024, 16 attention heads, and 340M total parameters. BERT is pre-trained with self-supervised approaches, i.e., Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

While Transformers such as BERT are reported to perform well in natural language processing, understanding and inference tasks, to the best of our knowledge, no prior works propose and examine the performance of transformer-based models for crisis classification.

2.3 CrisisBERT

In this work, we investigate the transformer approach for crisis classification tasks and propose CrisisBERT, a transformer-based classification model that surpasses conventional linear and deep learning models in performance and robustness. Figure [1] illustrates an overview of the proposed architecture. Overall, the architecture of CrisisBERT includes three layers, namely (1) Contextual Tokenization, (2) Transformer Language Model, and (3) Logistic Classifier.
Tokenization. DistilBERT inherits the embedding layer from BERT, where each word is tokenized contextually with subword-level word embeddings and positional encoders. Prefixed with a special token ([CLS]), the vector pairs are concatenated and passed into a DistilBERT LM for training.

| Hyper Parameter     | Final       | Search Space               |
|---------------------|-------------|----------------------------|
| Transformer         | BERT        | {BERT, XLNet, GPT2, RoBERTa}|
| Distillation        | True        | {True, False}              |
| Optimizer           | ADAMW       | {AdamW, Adam, SGD}         |
| Learning Rate       | 5e-5        | {5e-3, 2e-3, 5e-4, 2e-4, 5e-5, 2e-5} |
| Batch Size          | 32          | {16, 32, 64}               |

Table 1: Hyper-parameter search space and final values used

Transformer LM. We perform large-scale model selection and hyperparameter search to fin the best performing transformer LM for the crisis classification tasks. Several transformer models are investigated, including BERT, XLNet, GPT2 and RoBERTa. For each model, we conduct hyper-parameter search, where the search space includes variations of distillation [29], optimizers, learning rates, and batch sizes.

Table 1 shows the breakdown of the search space and the final hyper-parameters for CrisisBERT. Each set of parameters is randomly chosen and ran with 3 epochs and two trials. In total, we evaluate over 300 hyper-parameters sets using a Nvidia Titan-X (Pascal) for over 1,000 GPU hours.

Taking into consideration of performance and efficiency trade-off, we select the DistilBERT model for our Transformer LM layer. DistilBERT is a compressed version of BERT Base through Knowledge Distillation. With utilization of only 50% of the layers of BERT, DistilBERT performs 60% faster while preserving 97% of the capabilities in language understanding tasks. The optimal set of hyper-parameters for DistilBERT includes an AdamW [30] optimizer, and initial learning rate of 5e-5, and a batch size of 32.

Output Layer. The output layer of DistilBERT LM is a set of 768-d vectors led by the class header vector. Since we are conducting classification tasks, only the [CLS] token vector is used as the aggregate sequence representation for classification with a linear classifier. The remainder of the output vectors are processed into Crisis2Vec embeddings using Mean-Pooling operation.

2.4 Crisis2Vec

As discussed in Section 2.3, Crisis2Vec embedding is a byproduct of CrisisBERT, where the embeddings are constructed based on a pre-trained BERT model, and subsequently fine-tuned with three corpora of crisis-related tweets [6, 31, 32] to be domain-specific for crisis-related tweet representation.

Crisis2Vec leverages the advantages of Transformers, including (1) leveraging a self-attention mechanism to incorporate sentence-level context bidirectionally, (2) leveraging both word-level and positional information to create contextual representation of words, and (3) taking advantage of the pre-trained models on large relevant corpora.

To the best of our knowledge, we are the first who propose a document-level contextual embedding approach for crisis-related document representation. Upon convergence, we construct the fixed-length tweet vector using a MEAN-Pooling strategy [22], where we compute the mean of all output vectors, as illustrated in Algorithm 1.

Algorithm 1: Crisis2Vec with Mean Pooling

Input: Tweet as a sequence of Words

Output: Tweet_Embedding

Initialize Token_Vector;

for word in Tweet do
  Word-Token ← Embed(Word);
  Append Token_Vector ← Word-Token;
end
Output_Vectors ← CrisisBERT_LM(Token_Vector);
Tweet_Embedding ← Pool_Mean(Output_Vectors);
return Tweet_Embedding;
3 Crisis Classification

In this work, we conduct two crisis classification tasks, namely Crisis Detection and Crisis Recognition. We formulate the Crisis Detection task as a binary classification model that identifies if a tweet is relevant to a crisis-related event. The Crisis Recognition task on the other hand extends the problem into multi-class classification, where the output is a probability vector that indicates the likelihood of a tweet indicating specific events. Both tasks are modelled as Sequence Classification problems that are formally defined below.

3.1 Crisis Detection

We define the Crisis Detection task \( D = (S, \Phi) \), which is specified by \( S = \{ s_1, ..., s_n \} \) a finite sample space of tweets with size \( n \). Each sample \( s_i \) is a sequence of tokens at \( T \) time steps, i.e., \( s_i = \{ s_1^i, ..., s_T^i \} \). \( \Phi \) denotes the set of labels that has the same sequence as the sample set, \( \Phi = \{ \phi_1, ..., \phi_n \} \) and \( \phi_i \in \{ 0, 1 \} \) where \( \phi_i = 1 \) indicates that sample \( s_i \) is relevant to crisis, and \( \phi_i = 0 \) indicates otherwise. A deterministic classifier \( C_D : S \rightarrow \phi \) specifies the mapping from sample tweets to their flags.

Our objective is to train a crisis detector using the provided tweets and labels that minimizes the differences between predicted labels and true labels, i.e.,

\[
\min J_D(\Phi, C_D(S))
\]

where \( J_D \) denotes some cost function.

3.2 Crisis Recognition

Similarly, we define a Crisis Recognition task \( R = (S, L) \), where sample space \( S \) is identical to that in Crisis Detection. \( L \) denotes a sequence of multi-class labels that have the same sequence as \( S \), i.e., \( L = \{ l_1, ..., l_n \} \), and \( l_i \in \mathbb{R}^m \) for \( m \) number of classes. A deterministic classifier \( C_R : S \rightarrow L \) specifies the mapping from the sample tweets to the crisis classes.

The objective of the crisis classification tasks is to train a sequence classifier using the provided tweets and labels that minimizes the differences between predicted labels and true labels, i.e.,

\[
\min J_R(L, C_R(S))
\]

where \( J_R \) denotes some cost function for classifier \( C_R \).

| Label | Crisis event | # Data Points | Dataset | Label | Crisis event | # Data Points | Dataset |
|-------|-------------|---------------|---------|-------|-------------|---------------|---------|
| 1     | 2012_Sandy_Hurricane | 6318 | C6 | 2     | 2013_Alberta_Floods | 5189 | C6 |
| 3     | 2013_Boston_Bombings | 6577 | C6 | 4     | 2013_Oklahoma_Tornado | 4827 | C6 |
| 5     | 2013_Queensland_Floods | 6333 | C6 | 6     | 2013_West_Texas_Explosion | 6157 | C6 |
| 7     | sydneysiege | 837 | C8 | 8     | charliehebdo | 903 | C8 |
| 9     | ferguson | 859 | C8 | 10    | germanwings-crash | 248 | C8 |
| 11    | putinmissing | 59 | C8 | 12    | ottawashooting | 639 | C8 |
| 13    | ebola-essien | 21 | C8 | 14    | prince-toronto | 93 | C8 |
| 15    | 2012_Colorado_wildfires | 953 | C26 | 16    | 2012_Costa_Rica_earthquake | 909 | C26 |
| 17    | 2012_Guatemala_earthquake | 940 | C26 | 18    | 2012_Italy_earthquakes | 940 | C26 |
| 19    | 2012_Philippines_floods | 906 | C26 | 20    | 2012_Typhoon_Pablo | 907 | C26 |
| 21    | 2012_Venezuela_refinery | 939 | C26 | 22    | 2013_Australia_bushfire | 949 | C26 |
| 23    | 2013_Bohol_earthquake | 969 | C26 | 24    | 2013_Brazil_nightclub_fire | 952 | C26 |
| 25    | 2013_Colorado_floods | 925 | C26 | 26    | 2013_Glasgow_helicopter_crash | 918 | C26 |
| 27    | 2013_LA_airport_shootings | 912 | C26 | 28    | 2013_Lac_Megantic_train_crash | 966 | C26 |
| 29    | 2013_Manao_floods | 921 | C26 | 30    | 2013_NY_train_crash | 999 | C26 |
| 31    | 2013_Russia_meteor | 1133 | C26 | 32    | 2013_Sardinia_floods | 926 | C26 |
| 33    | 2013_Savar_buildingCollapse | 911 | C26 | 34    | 2013_Singapore_haze | 933 | C26 |
| 35    | 2013_Spain_train_crash | 991 | C26 | 36    | 2013_Typhoon_Yolanda | 940 | C26 |

Table 2: Classes Description

4 Experiments and Results

In this section, we discuss the experiments performed and their results in order to propose a highly effective and efficient approach for text classification.
4.1 Datasets

Three datasets of labelled crisis-related tweets \[6, 31, 32\] are used to conduct crisis classification tasks and evaluate the proposed methods against benchmarks. In total, these datasets consist of close to 8 million tweets, where overall 91.6k are labelled. These data sets are in the form of: (1) 60k labelled tweets on 6 crises \[6\], (2) 3.6k labelled tweets for 8 crises \[32\], and (3) 27.9k labelled tweets for 26 crises \[31\]. Table 2 describes more detail about each dataset and their respective classes.

For our experimental evaluation, the 91.6k labelled crisis-related tweets are organized into two datasets, annotated as C6 and C36. In particular, C6 consists of 60k tweets from 6 classes of crises, whereas C36 comprises all 91.6k tweets in 36 classes. Both datasets are split into training, validation and test sets that consist of 90%, 5% and 5% of the original sets, respectively. Table 3 describes the statistics of the split sets.

| Dataset | # Data Points | # Classes | Train | Test | Split |
|---------|---------------|-----------|-------|------|-------|
| C6      | 60k           | 6 + 1     | 54072 | 3004 | 3004  |
| C36     | 91.6k         | 36 + 1    | 82506 | 4584 | 4584  |

Table 3: Dataset Description

4.2 Proposed Models

CrisisBERT. We evaluate the performance of CrisisBERT against multiple benchmarks, which comprise recently proposed crisis classification models in the literature. These works include linear classifiers, such as Logistic Regression (LR), Support Vector Machine (SVM) and Naive Bayes \[33\], and non-linear neural networks, such as Convolutional Neural Network (CNN) \[20\] and Long Short-Term Memory \[34\].

Furthermore, we investigate the robustness of CrisisBERT for both detection and recognition tasks. This is achieved by extending the experiments from C6 to C36, which comprise 6 and 36 classes respectively, but with only 51.4% additional data points. We evaluate the robustness of the proposed models against benchmarks by observing the compromise in robustness performance, while realizing the drastically improved classification performance.

As described in Section 2.3, we use the optimal set of hyper-parameters for CrisisBERT in the experiments, which include the use of a BERT model with distillation (i.e. DistilBERT), an AdamW \[30\] optimizer, an initial learning rate of 5e-5, a batch size of 32, and a word dropout rate of 0.25.

Crisis2Vec. To evaluate Crisis2Vec, we choose the two classifiers with the aim to represent both traditional Machine Learning approaches and the NN approaches. The two selected models are: (1) a linear Logistic Regression model, denoted as \(LR_{c2v}\), and (2) a non-linear LSTM model, denoted as \(LSTM_{c2v}\). We evaluate the performance of Crisis2Vec with the two models by replacing the original embedding to Crisis2Vec, ceteris paribus.

4.3 Metrics

We use two common evaluation metrics, namely Accuracy and F1 score, which are functions of True-Positive (TP), False-Positive (FP), True-Negative (TN) and False-Negative (FN) predictions. Accuracy is calculated by:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

For a F1-score of multiple classes, we calculate the unweighted mean for each label, i.e., for \(n\) classes of labels as:

\[
F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]

where

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

and

\[
\text{Recall} = \frac{TP}{TP + FN}
\]
4.4 Benchmark Algorithms

We select and implement several crisis classifiers proposed in recent works to serve as benchmarks for evaluating our proposed methods. Concretely, we compare *CrisisBERT* with the following models:

- **LR<sub>W2V</sub>:** Logistic regression model with Word2Vec embedding pre-trained on Google News Corpus [33]
- **SVM<sub>W2V</sub>:** Support Vector Machine model with Word2Vec embedding pre-trained on Google News Corpus [33]
- **NB<sub>W2V</sub>:** Naive Bayes model assuming Gaussian distribution for features with Word2Vec embedding pre-trained on Google News Corpus [33]
- **CNN<sub>G</sub>:** Convolutional Neural Network model with 2 convolutional layers of 128 hidden units, kernel size of 3, pool size of 2, 250 filters, and GloVe for word embedding [20]
- **LSTM<sub>W2V</sub>:** Long Short-Term Memory model with 2 layers of 30 hidden states and a Word2Vec-based Crisis Embedding [34]

| Models            | Detection-C6 | Detection-C36 | Recognition-C6 | Recognition-C36 |
|-------------------|--------------|---------------|----------------|-----------------|
| **Proposed**      |              |               |                |                 |
| *CrisisBERT*      | 95.5         | 95.6          | 94.2           | 94.7            |
| LSTM<sub>W2V</sub> | 95.1         | 95.1          | 92.8           | 93.5            |
| LR<sub>W2V</sub>  | 93.2         | 93.2          | 89.0           | 90.1            |
| **Benchmark**     |              |               |                |                 |
| LR<sub>W2V</sub>  | [33]         | 91.5          | 91.6           | 85.3            |
| SVM<sub>W2V</sub> | [33]         | 91.3          | 91.4           | 85.0            |
| NB<sub>W2V</sub>  | [33]         | 86.8          | 86.8           | 82.4            |
| CNN<sub>G</sub>   | [20]         | 91.2          | 91.3           | 91.1            |
| LSTM<sub>W2V</sub>| [34]         | 91.7          | 91.7           | 88.0            |

Table 4: Experimental results of Crisis Classification tasks on C6 and C36 datasets for proposed models and benchmarks, where best performers are emphasized. Results show that *CrisisBERT* records highest performance across all tasks.

4.5 Results

Overall, the experimental results show that both proposed models achieve significant improvement on performance and robustness over benchmarks across all tasks. The experimental results for *CrisisBERT* and *Crisis2Vec* are tabulated in Table 4.

**Crisis Detection.** For Crisis Detection tasks on C6 dataset, *CrisisBERT* achieves a 95.5% F1-score and 95.6% Accuracy, which exceeds previous best model results, namely LSTM with pre-trained Word2Vec embeddings, by 3.8% and 3.8% respectively. In terms of embedding, LSTM with *Crisis2Vec* records 95.1% for both F1-score and Accuracy, which shows 3.4% improvement over LSTM with Word2Vec. Similarly, LR with *Crisis2Vec* records 93.2% for both F1-score and Accuracy, which shows 1.6% improvement over LR with Word2Vec.

For Crisis Detection tasks on C36 dataset, *CrisisBERT* achieves 94.2% F1-score and 94.7% Accuracy, which exceeds previous best model results, namely CNN with pre-trained GloVe embedding, by 2.7% and 2.1% respectively. In terms of embeddings, LSTM with *Crisis2Vec* records 92.8% F1-score and 93.5% Accuracy, which exceed LSTM with Word2Vec by 4.8% and 4.2%. Similarly, LR with *Crisis2Vec* records 89.0% F1-score and 90.1% Accuracy, which exceed LR with Word2Vec by 3.7% and 3.5% respectively.

**Crisis Recognition.** For Crisis Recognition tasks on C6 dataset, *CrisisBERT* achieves 98.7% F1-score and 98.6% Accuracy, which exceeds previous best model results, namely CNN with pre-trained GloVe embeddings, by 8.2% for both F1-score and Accuracy. In terms of embedding, LSTM with *Crisis2Vec* records 97.5% for both F1-score and Accuracy, which shows a significant improvement of 10.2% over LSTM with Word2Vec. Similarly, LR with *Crisis2Vec* records 93.6% F1-score and 93.7% Accuracy, which shows 6.3% and 5.2% improvement respectively over LR with Word2Vec.

For Crisis Recognition tasks on C36 dataset, *CrisisBERT* achieves 97.1% F1-score and 97.9% Accuracy, which significantly exceeds previous best model results, namely LR with pre-trained Word2Vec embedding, by 25.0% and 15.6% respectively. In terms of embeddings, LSTM with *Crisis2Vec* records 88.0% F1-score and 95.6% Accuracy, which significantly exceed LSTM with Word2Vec by 29.1% and 23.3%. Similarly, LR with *Crisis2Vec* records 85.1% F1-score and 90.9% Accuracy, which significantly exceed LR with Word2Vec by 13.0% and 8.6% respectively.
Robustness. Comparing Crisis Detection task between C6 and C36, CrisisBERT shows 1.7% and 1.3% decline for F1-score and Accuracy, which is much better than most benchmarks, i.e., from 1.7% to 6.3%, except CNN. However, when we compare the more challenging Crisis Recognition tasks between C6 and C36, the performance of CrisisBERT compromises marginally, i.e., 1.6% for F1-score and 0.7% for Accuracy. On the contrary, all benchmark models record significant decline, i.e., from 6.0% to 67.2%.

Discussion. Based on experimental results discussed above, we observe that: (1) CrisisBERT’s performance exceeds state-of-the-art performance for both detection and recognition tasks, with up to 8.2% and 25.0% respectively, (2) CrisisBERT demonstrates higher robustness with marginal decline for performance (i.e., less than 1.7% in F1-score and Accuracy), and (3) Crisis2Vec shows superior performance as compared to conventional Word2Vec embeddings, for both LR and LSTM models across all experiments.

5 Related Work

5.1 Crisis Classification

Event detection from tweets has received significant attention in research in order to analyze crisis-related messages for better disaster management and increasing situational awareness [4, 5, 2, 3]. Parilla-Ferrer et al. proposed automatic binary classification of informativeness using Naive Bayes and Support Vector Machine (SVM) [13]. Stowe et al. presented an annotation scheme for tweets to classify relevancy and six [8]. Furthermore, use of pre-trained word2vec reportedly improved SVM for Crisis classification [14]. Lim et al. proposed CLUSTOP algorithm utilizing the community detection approach for automatic topic modelling [15]. Pedrood et al. proposed to transfer-learn classification of one event to the other using a sparse coding model [16], though the scope was only limited to only two events, i.e. Hurricane Sandy (2012) and Supertyphoon Yolanda (2013).

A substantial number of works focusses on using Neural Networks (NN) with word embeddings for crisis-related data classification. Manna et al. [33] compared NN models with conventional ML classifiers [33]. ALRashdi and O’Keefe investigated and showed good performance for two deep learning architectures, namely Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) with domain-specific GloVe embeddings [17]. However, the study had yet to validate the relevance of model on a different crisis type. Nguyen et al. applied CNN to classify information types using Google News and domain-specific embeddings [18]. Kersten et al. [20] implemented a parallel CNN to detect two disasters, namely hurricanes and floods, which reported a F1-score of 0.83. The CNN architecture was proposed earlier by Kim et al. [19].

5.2 Crisis Embedding

Word-level Embeddings such as Word2Vec [35] and GloVe [23] are commonly used to form the basis of Crisis Embedding [18, 36] in various crisis classification works to improve model performance. For context, Word2Vec uses a Neural Network Language Model (NNLM) that is able to represent latent information on the word level. GloVe achieved better results with a simpler approach, constructing global vectors to represent contextual knowledge of the vocabulary.

More recently, a series of high quality embedding models, such as FastText [37] and Flair [38], are proposed and reported to have improved state of the art for multiple NLP tasks. Both word-level contextualization and character-level features are commonly used for these works. Pre-trained models on large corpora of news and tweets collections are also made publicly available to assist in downstream tasks. Furthermore, Transformer-based models are proposed to conduct sentence-level embedding tasks [22].

6 Conclusion

Social media such as Twitter has become a hub of crowd generated information for early crisis detection and recognition tasks. In this work, we present a transformer-based crisis classification model CrisisBERT, and a contextual crisis-related tweet embedding model Crisis2Vec. We examine the performance and robustness of the proposed models by conducting experiments with three datasets and two crisis classification tasks. Experimental results show that CrisisBERT improves state of the art for both detection and recognition class, and further demonstrates robustness by extending from 6 classes to 36 classes, with only 51.4% additional data points. Finally, our experiments with two classification models show that Crisis2Vec enhances classification performance as compared to Word2Vec embeddings, which is commonly used in prior works.
7 Acknowledgement

This research is funded in part by the Singapore University of Technology and Design under grant SRG-ISTD-2018-140.

References

[1] Internet Live Statistics. Twitter usage statistics. *Internet*, 2020.
[2] Naina Said, Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications*, 78(22):31267–31302, 2019.
[3] Luke S Snyder, Mortezza Karimzadeh, Christina Stober, and David S Ebert. Situational awareness enhanced through social media analytics: A survey of first responders. *arXiv preprint arXiv:1909.07316*, 2019.
[4] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015.
[5] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.
[6] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In * Eighth international AAAI conference on weblogs and social media*, 2014.
[7] Shanshan Zhang and Slobodan Vucetic. Semi-supervised discovery of informative tweets during the emerging disasters. *arXiv preprint arXiv:1610.03750*, 2016.
[8] Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, and Kenneth M Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The fourth international workshop on natural language processing for social media*, pages 1–6, 2016.
[9] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *Isgram*, 2013.
[10] Amanda L Hughes, Lise AA St. Denis, Leysia Palen, and Kenneth M Anderson. Online public communications by police & fire services during the 2012 hurricane sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1505–1514, 2014.
[11] Grégoire Burel, Hassan Saif, Miriam Fernandez, and Harith Alani. On semantics and deep learning for event detection in crisis situations. 2017.
[12] Justin Michael Crow. Verifying baselines for crisis event information classification on twitter. In *Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2020).* ISCRAM Digital Library, 2020.
[13] Beverly Estephany Parilla-Ferrer, PL Fernandez, and JT Ballena. Automatic classification of disaster-related tweets. In *Proc. International conference on Innovative Engineering Technologies (ICIET)*, volume 62, 2014.
[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
[15] Kwan Hui Lim, Shanika Karunasekera, and Aaron Harwood. Clustop: A clustering-based topic modelling algorithm for twitter using word networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 2009–2018. IEEE, 2017.
[16] Bahman Pedrood and Hemant Purohit. Mining help intent on twitter during disasters via transfer learning with sparse coding. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 141–153. Springer, 2018.
[17] Reem AL Rashdi and Simon O’Keefe. Deep learning and word embeddings for tweet classification for crisis response. *arXiv preprint arXiv:1903.11024*, 2019.
[18] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.
[19] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
[20] Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. Robust filtering of crisis-related tweets. In *ISCRAM 2019 Conference Proceedings-16th International Conference on Information Systems for Crisis Response and Management*, 2019.
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 2019.

[26] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764, 2019.

[27] Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kuo-Kai Shyu. Ncuee at mediqa 2019: Medical text inference using ensemble bert-bilstm-attention model. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 528–532, 2019.

[28] Junhua Liu, Yung Chuen Ng, Kristin L Wood, and Kwan Hui Lim. Ipod: An industrial and professional occupations dataset and its applications to occupational data mining and analysis. arXiv preprint arXiv:1910.10495, 2019.

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[31] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pages 994–1009, 2015.

[32] Arkaiz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one, 11(3), 2016.

[33] Sukanya Manna and Haruto Nakai. Effectiveness of word embeddings on classifiers: A case study with tweets. In 2019 IEEE 13th International Conference on Semantic Computing (ICSC), pages 158–161. IEEE, 2019.

[34] Abhinav Kumar, Jyoti Prakash Singh, Yogesh K Dwivedi, and Nripendra P Rana. A deep multi-modal neural network for informative twitter content classification during emergencies. Annals of Operations Research, pages 1–32, 2020.

[35] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155, 2003.

[36] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. Applications of online deep learning for crisis response using social media information. arXiv preprint arXiv:1610.01030, 2016.

[37] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[38] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649, 2018.