Tweets Classification with BERT in the Field of Disaster Management

Guoqin Ma
Department of Civil Engineering
Stanford University
Stanford, CA 94305
sebsk@stanford.edu

Abstract

Crisis informatics focus on the contribution of user generated content (UGC) to disaster management. To leverage the social media data effectively, it is crucial to filter out noisy information from the large volume of data flow so that we could better estimate disaster damage with these data. Not satisfied with basic keyword-based filtration, many researchers turn to machine learning for solution. In this project, I apply deep learning techniques to address Tweets classification problem in disaster management field. The labels of Tweets reflect different types of disaster-related information, which have different potential usage in emergency response. In particular, BERT is used for transfer learning. The standard BERT architecture for classification and several other customized BERT architectures are trained to compare with the baseline bidirectional LSTM with pretrained Glove Twitter embeddings. Results show that BERT and BERT-based LSTM attain the best results, outperforming the baseline model by 3.29% on average in terms of F-1 score respectively. Ambiguity and subjectivity affect the performance of these models considerably. In some examples the models can surpass human performance.

1 Introduction

1.1 Crisis informatics

The rise of social media over the past 15 years has marked a shift in the potential of how information is collected and disseminated during natural disasters [17]. Researchers have leveraged social media to fulfill many tasks in disaster management, including but not limited to outbreak detection [6] [10], information retrieval [9], sentiment analysis [18] [10], evacuation behavior study [19] [4], hazard assessment [3] [7], and damage assessment [11] [16]. A foundation that guarantees success of all these tasks is an (or some) effective filtering technique(s) that could filter out noisy information carried by the data stream and cream off those messages containing rich information pertaining to disasters.

1.2 Related work on disaster text classification

Relevance classification is a widely used classification criterion in this field. Tweets are divided into 2 categories: on-topic or off-topic. The drawback of using this binary classification is that some messages may be on-topic but it conveys nothing useful for disaster response. 1 example is: *these hurricane sandy tweets are so hilarious tho*, which is an on-topic but not informative Tweet extracted from CrisisLexT6 Hurricane Sandy dataset.

Another classification criterion is informativeness-based. Representative datasets are CrisisLexT26 and some datasets on CrisisNLP. Minor difference in labeling may exist between these datasets, but generally the labels are: affected individuals, infrastructure and utilities damage, caution and advice, donation and volunteering, sympathy and emotional support, other useful information, not related or not informative, etc. This labeling could offer us more insights into the local situation when a disaster occurs. This labeling system is what I use in this project. Details on the meanings of labels could be found in the paper of Olteanu et al. [13].

Previous machine learning models include support vector machine [8], naive Bayes classifier [8] [12], random forest [8], LDA [16], CNN [1], etc. Since there is no gold-standard training dataset (different labeling systems, different text data, etc.), it is hard to compare across these models and there is no well-defined state-of-art model. Many previous papers train each disaster event separately (training data and test data are from one single event) and/or adopt a binary classification, which limit data size and practical usage of the trained model.

1.3 Achievement of this project

I aggregate all the datasets with the aforementioned multi-label annotation that I could find to use them as training data. I also set up a bidirectional LSTM with GloVe Twitter embeddings as baseline. The remaining task of this project is to make use of BERT to enhance the performance of classifier. The reason to choose BERT is that it has achieved state-of-art performance in many NLP tasks and it is open-sourced. I developed several BERT-based models and they all surpass the baseline performance. The fine-tuned approach in the paper of Devlin et al. [5] and BERT-based bidirectional LSTM achieved best performance, boosting up 3% in terms of accuracy, Matthew coef, F-1.

2 Approach

In this project, BERT models are built based on the `pytorch-pretrained-BERT` repository possessed by *huggingface* at `https://github.com/huggingface/pytorch-pretrained-BERT`. All the BERT models are builted upon BERT base uncased model, which has 12 transformer layers, 12 self-attention heads, and with a hidden size 768.

2.1 Text preprocessing

Texts are lowercased. Non-ascii letters, urls, @RT:[NAME], @[NAME] are removed. For BERT, an additional `[CLS]` token is inserted to the beginning of each text. Texts with length less than 4 are thrown away. No lemmatization is performed and no punctuation mark is removed since pre-trained embeddings are always used. No stop-word is removed for fluency purpose.

2.2 Baseline

The baseline model is a single-layer bidirectional LSTM neural network with hidden size 256. The model receives pre-trained GloVe Twitter 27B embeddings (200d) [15] as input. The stacked final hidden state of the sequence is linked to a fully connected layer to perform softmax.

2.3 BERT for sequence classification (default BERT)

This is the default BERT model for sequence classification. The final hidden state of the first word (`[CLS]`) from BERT is input to a fully connected layer to perform softmax [5]. This is a fine-tuning approach.

2.4 BERT + nonlinear layers (BERT+NL)

This model is developed from BERT default model for sequence classification. The final hidden state of the first word (`[CLS]`) from BERT is input to 3 fully connected layers. First two layers are with leaky ReLU activation (negative slope = 0.01) and the third layer is for performing softmax. This is a fine-tuning approach.

2.5 Customized BERT + LSTM (BERT+LSTM)

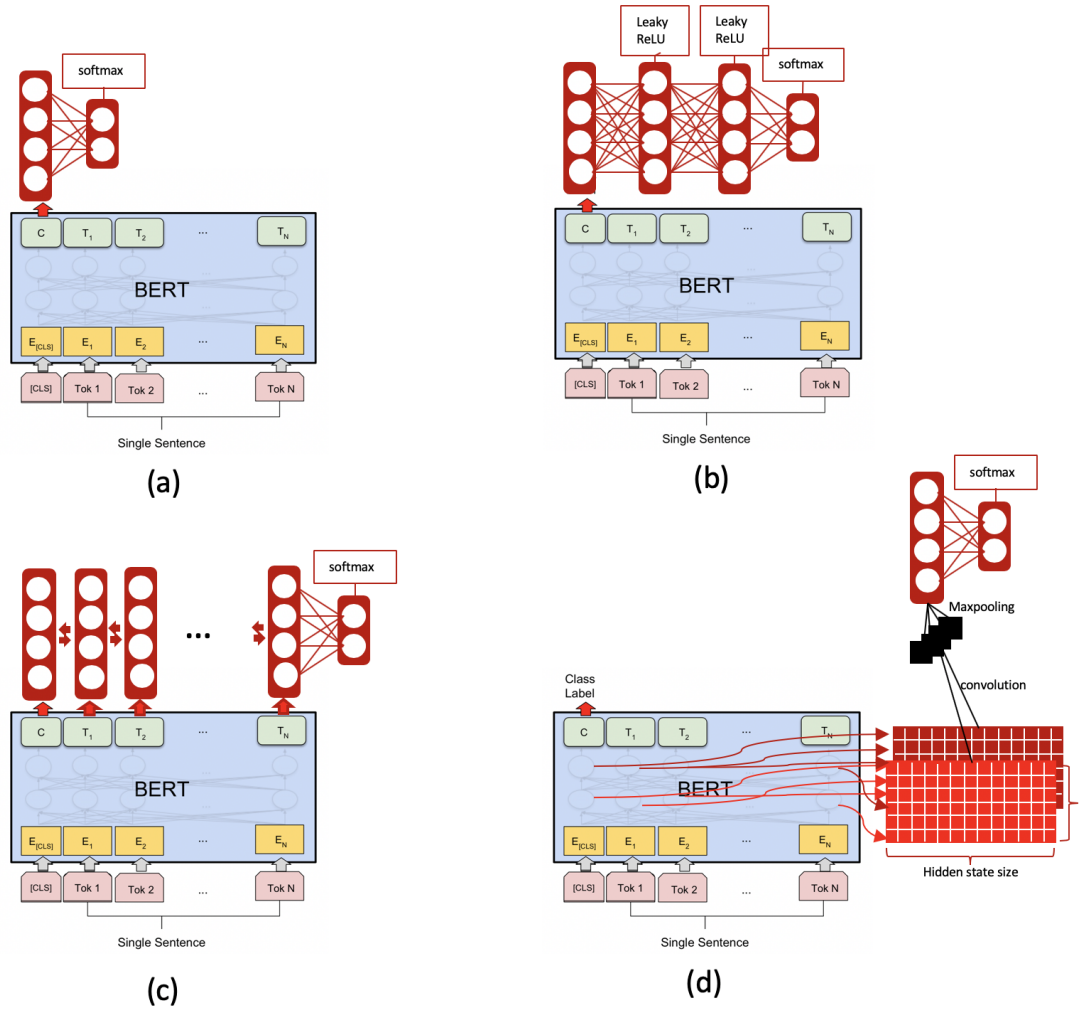
A bidirectional LSTM is stacked on top of BERT. The final hidden state of the all the words in the sequence from BERT is input to the bi-LSTM. The stacked final LSTM hidden state of the sequence is linked to a fully connected layer to perform softmax. To make it comparable with the baseline LSTM, the hidden size of the LSTM is also set to be 256. This is a feature-based approach.

2.6 Customized BERT + convolution (BERT+CNN)

This model utilizes hidden states from all the layers of BERT. For each layer, a convolution is performed with 16 filters of size (3, hidden size of BERT) on the hidden states of the sequence. The out channels are then concatenated together. For BERT base model, the total number of out channels is 192. Max value is taken from all the out channels to squeeze the convolution output into a 1 dimensional vector, which is then, again, linked to a fully connected layer to perform softmax. This is a feature-based approach.

The diagrams of the 4 BERT-based models are summarized in Fig. 1, modified from the diagram shown in the paper of Devlin et al. [5].

Figure 1: BERT-based model diagrams adapted from BERT paper: (a) default BERT, (b) BERT+NL, (c) BERT+LSTM, (d) BERT+CNN



3 Experiments

3.1 Data

The classification criterion is informativeness-based. Representative labeled datasets are CrisisLexT26 and some datasets on CrisisNLP. Minor difference in labeling may exist between these datasets, but generally the labels are: affected individuals, infrastructure and utilities damage, caution and advice, donation and volunteering, sympathy and emotional support, other useful information, not related or not informative, etc. These labels could offer us insights into the local situation when a disaster occurs. The aforementioned labeled datasets are compiled into a single large one. The sources of the datasets are described in Table 1. The number of labeled Tweets is 75800. After dropping out extremely short texts (length less than 4), the sample size is 74346. Most Tweets are in English, with some sparse exceptions. The distribution of disaster type is presented in Table 2. The distribution of labels is summarized in Table 3. The label distribution and disaster type distribution are both heavily skewed.

Table 1: Labeled Datasets

Collection	Dataset	Reference
CrisisLex	T26	[14]
CrisisNLP	#1	[8]
CrisisNLP	#2,#3	[9]
CrisisNLP	#6	[2]

Table 2: Labels

Label	Count
not related or not informative	25785
other useful information	18877
donations and volunteering	8925
affected individuals	8009
sympathy and emotional support	5020
infrastructure and utilities damage	4559
caution and advice	3171

Table 3: Disaster Types

Disaster type	Count
hurricane	30860
earthquake	20540
floods	8682
wildfires	3620
landslides	2598
traffic crash	2385
terrorism	1977
building collapse	945
meteor	915
explosion	907
haze	706
volcano	211

3.2 Evaluation method

5 metrics, namely accuracy, Matthews correlation coefficient, precision, recall, F1-score, are considered when evaluating a model. Accuracy, Matthews correlation coefficient, macro precision, macro recall, macro F-1 score are calculated from all the classes, while precision, recall, and F1-score, are calculated for each class.

3.3 Experimental details

Batch size is 32. Initial learning rate is set to 0.001 for non-BERT parameters and 0.00002 for BERT parameters, as suggested by [5]. Patience is 5, after reaching which learning rate decay by 50%. The maximum number of learning rate decay trials is 5 before earlystopping. Adam optimizer is used for all models' training. Max epoch is set to 100, but an earlystopping is expected before reaching 100 epochs. Cross entropy loss is the loss function, with $label_i$ weighted by $max_j(size(label_j))/size(label_i)$

Validation dataset and test dataset are of size 5000 respectively. The rest 64346 samples are used for training. Samples are shuffled between epochs during training.

3.4 Results

Table 4: Evaluation metrics

Model	Accuracy	Matthews coef	Macro precision	Macro recall	Macro F-1
baseline	0.64	0.56	58.00	68.43	60.71
default BERT	0.67	0.59	60.43	71.14	64.00
BERT+NL	0.67	0.59	60.57	68.00	63.14
BERT+LSTM	0.67	0.60	61.29	69.86	64.00
BERT+CNN	0.67	0.59	60.86	69.29	63.43

The evaluation metrics are summarized in Table 4. The accuracy score and Matthews of all BERT models are higher than the baseline for around 3%. Averagely, default BERT, BERT+NL, BERT+LSTM, BERT+CNN outperform baseline model by 3.3 (%), 2.4, 3.3, 2.7 respectively in terms of F-1 score; 2.4, 2.6, 3.3, 2.9 in terms of precision; 2.7, -0.4, 1.4, 0.9 in terms of recall. The 3 customized BERT models have better performance in precision but worse performance in recall than the default BERT.

Label-wise, "other useful information", "not related or not informative", "donation and volunteering", "affected individuals" have higher precision score, partially due to larger size because a clear descending transition can be seen from left to right. "donation and volunteering", "affected individuals", "sympathy and support", "infrastructure and utilities damage" have higher recall, also partially due to size effect. Overall, in terms of F-1 score, "other useful information", "donation and volunteering", "affected individuals" can be better classified. Detailed scores can be found in Appendix A. The divergent performance of models among different classes are discussed in the following section.

4 Error analysis

The confusion matrix of the test dataset from the default BERT model is shown in Fig. 2. The rest 4 models have the similar distribution of errors across the labels. The confusion matrix can be divided into four blocks. Mark texts from "not related or not informative" and "other useful information" as Group A and rest five labels as Group B. Many texts inside Group A are misclassified either into the other side or into Group B. A number of texts from Group B are also misclassified into Group A but are less mislabeled internally.

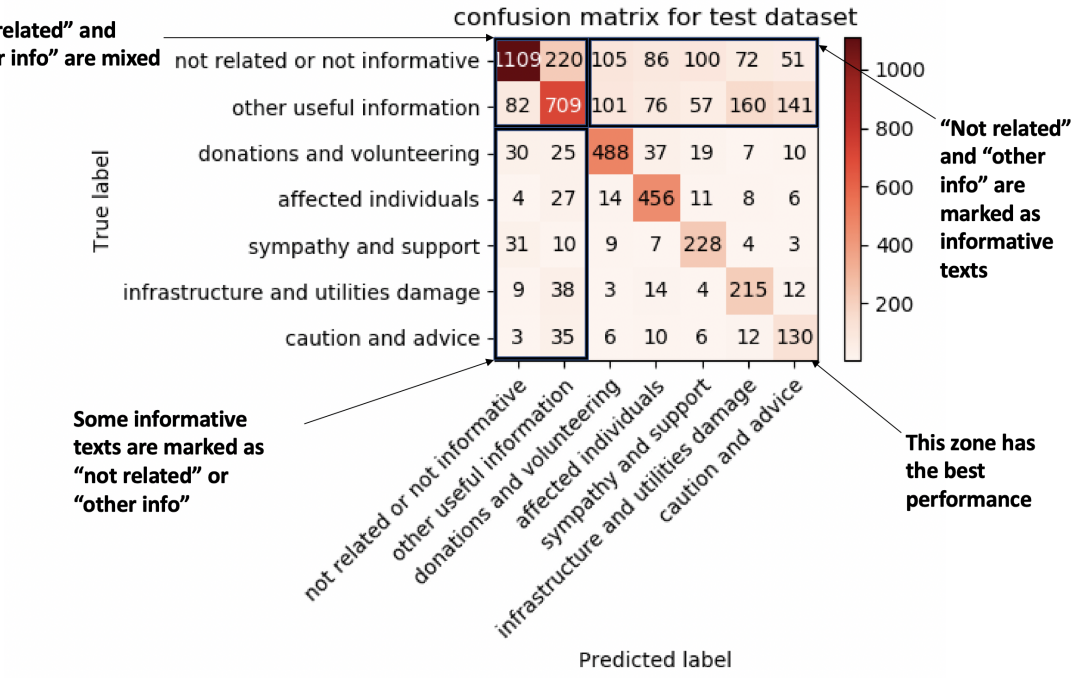
The reasons of the misclassification are summarized below, each reason followed by an example and a description to the example:

1. Ambiguity and subjectivity play an important role.

   Example: Short time lapse video from James Reynolds of EarthUncutTV as typhoon Ruby (Hagupit) approaches Calbayog City,... http://t.co/JiYbXJBzHu

   Description: This text is labeled as "not related or not informative" but is classified as "Other useful information". It does convey the on-site information on Typhoon development, agreeing with the description of "other useful information" in the paper of Olteau et al [13].

Figure 2: Confusion matrix of test dataset



2. Some need more contexts to fully understand, e.g. html.

   Example: Can u please take a note. @SushmaSwaraj @MEAIndia @PMOIndia @narendramodi @BJPRajnathSingh @HMOIndia @adgpi https://t.co/hP1GVWqi4n

   Description: This text is labeled as "Other useful information" but is classified as "not related or not informative". The link is about affected individuals by Nepal earthquake.

3. Non-ascii words and Emoji missing.

   Example: Now everyone is religious Ã83Â83Ã82Â¡Ã83Â82Ã82ÂÃ83Â82Ã82Â4 #prayforblahblah #harvey #irma #maria #mexico #hurricane #earthquake https://t.co/nzEbRWQG8W

   Description: This text is labeled as "not related or not informative" but is classified as "Other useful information". The Emoji "face with rolling eyes" is not proper represented in the original text data.

4. Semantic misconstrue (Sarcastic, metaphoric...)

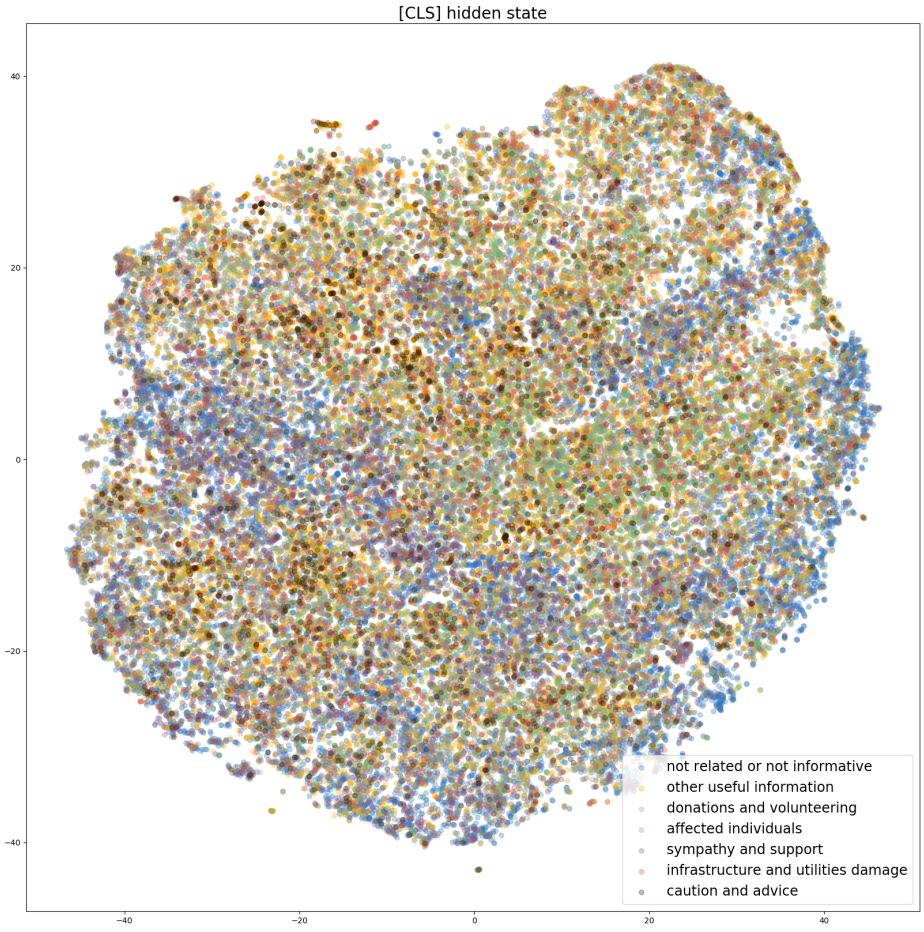
   Example: The Bay Area media is going a little over board with this Napa earthquake. Do we really need updates every 15 minutes?

   Description: This text is labeled as "not related or not informative" but is classified as "Other useful information". The author of this Tweet meant to be sarcastic.

5. Keyword influence or Misleading hashtag (#).

   Example: Skaters Make Best of Napa Earthquake by Shredding Buckled Streets http://t.co/TkuDa0FHN8 via @mashable

Figure 3: TSNE of [CLS] final hidden state



Description: This text is labeled as "not related or not informative" but is classified as "infrastructure and utilities damage". The classifier is confused by the word "shred", "buckle" and "street".

6. Some events may not really happen.

   Example: 38 cities in India fall in high risk earthquakes zones - India Today http://t.co/5PGlJvIKa9 via @indiatoday.Action needed to save the cities

   Description: This text is labeled as "not related or not informative" but is classified as "other useful information". The news describes the potential of earthquake but not real damage.

7. Message is too short

Example: Instant pond! #rubyPH http://t.co/XyCh1POC3l

Description: This text is labeled "other useful information" but is classified as "not related or not informative". The length of this message is perhaps too short to let the classifier make the decision.

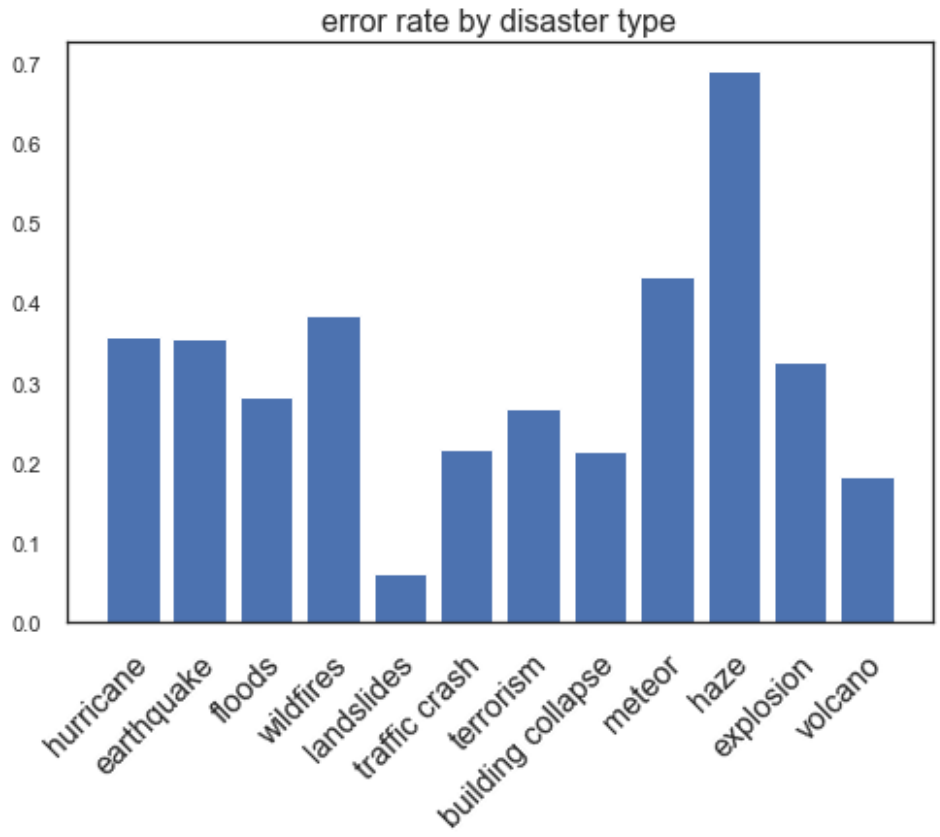
8. A lot of misclassified messages are news, which is less useful than Tweets from disaster eye-witnesses and the usage of which is debatable.

   Example: Today at a press conference in New York, Secretary of State John Kerry announced an additional $9m dollars of U.S. aid towards #Nepal. #cnn

   Description: This text is labeled as "not related or not informative" but is classified as "donation and volunteering".

Among the various reasons, ambiguity and subjectivity are the dominant reasons to explain the low accuracy of all the models when the author canvassed all the misclassified cases. In a lot of cases, it is hard to give the text full credit to any class. Figure 3 shows the TSNE projection of `[CLS]` final hidden state. It is hard to observe obvious clusters between classes. Distribution of predicted probabilities for each class can be found in Appendix B. "affected individuals" is the most separable label, which means texts from this class are least ambiguous, while "other useful information" is the least separable one, the probability distribution of which is nearly uniform between 0 and 1.

Figure 4: Error rate across different disaster types



The error rate (proportion of examples of the diagnal in the confusion matrix) across different disaster types are shown in Fig 4. Haze and meteor have the highest error rate because of the nature of the disaster and limited data size. Hurricane, earthquake, floods, wildfires have high error rate, potentially due to the aggregation process which may introduces additional noise to the labeled data.

5 Conclusions

Social media has been drawing attentions from researchers and practitioners in the field of disaster management. Accurate message classification is a necessary requirement to make decisions from the abundant but noisy user-generated data. BERT-based classifiers could attain better performance compared with the bi-LSTM baseline model. Some labels are better predictable than others. Ambiguity and subjectivity are a great obstacle to boost up performance of classifier. The quality of the data needs improvement.

6 Additional information

Mentor: Amita Kamath

7 Github repo

All codes are uploaded to the Github repository `https://github.com/sebsk/CS224N-Project`

8 Acknowledgement

I would like to thank Professor Manning and my mentor Amita for their guidance and advice to this work.

References

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran. "Graph Based Semi-supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets". In: (2018). DOI: arXiv:1805.06289v1. URL: http://arxiv.org/abs/1805.06289.

[2] Firoj Alam et al. "A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria". In: *the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Rochester, USA, 2018.

[3] L Burks, M Miller, and R Zadeh. "Rapid Estimate Of Ground Shaking Intensity By Combining Simple Earthquake Characteristics With Tweets". In: *Tenth U.S. National Conference on Earthquake Engineering*. Anchorage, AK, USA, 2014.

[4] Danaë Metaxa-Kakavouli, Paige Maas, and Daniel P. Aldrich. "How Social Ties Influence Hurricane Evacuation Behavior – Facebook Research". In: *the ACM on Human-Computer Interaction* 2 (2018). DOI: https://doi.org/10.1145/3274391. URL: https://research.fb.com/publications/how-social-ties-influence-hurricane-evacuation-behavior/.

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 2018.

[6] Paul S. Earle, Daniel C. Bowden, and Michelle Guy. "Twitter earthquake detection: Earthquake monitoring in a social world". In: *Annals of Geophysics* 54.6 (2011), pp. 708–715. ISSN: 15935213. DOI: 10.4401/ag-5364.

[7] Benjamin Herfort et al. "Twitter Analysis of River Elbe Flood". In: *International Conference on Information Systems for Crisis Response and Management*. Pennsylvania, USA, 2014. URL: http://www.producao.usp.br/handle/BDPI/46102.

[8] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *the 10th Language Resources and Evaluation Conference (LREC)*. 2016. ISBN: 9782951740891.

[9] Muhammad Imran et al. "Extracting Information Nuggets from Disaster-Related Messages in Social Media". In: *the 10th International ISCRAM Conference*. Ed. by T. Comes et al. Baden-Baden, Germany, 2013, p. 791. URL: http://www.crowdflower.com.

[10] Yury Kryvasheyeu et al. "Performance of social network sensors during Hurricane Sandy". In: *PLoS ONE* (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0117288.

[11] Yury Kryvasheyeu et al. "Rapid assessment of disaster damage using social media activity". In: *Science Advances* (2016). ISSN: 23752548. DOI: 10.1126/sciadv.1500779.

[12] Hongmin Li et al. "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26.1 (Mar. 2018), pp. 16–27. ISSN: 09660879. DOI: 10.1111/1468-5973.12194. URL: http://doi.wiley.com/10.1111/1468-5973.12194.

[13] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. "What to Expect When the Unexpected Happens: Social Media Communications Across Crises Human Factors; Measurement". In: (). DOI: 10.1145/2675133.2675242. URL: http://dx.doi.org/10.1145/2675133.2675242.

[14] Alexandra Olteanu et al. "CrisisLex : A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *the Eighth International AAAI Conference on Weblogs and Social Media*. 2014, pp. 376–385.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. ISBN: 9781937284961. DOI: 10.3115/v1/D14-1162.

[16] Bernd Resch, Florian Usländer, and Clemens Havas. "Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment". In: *Cartography and Geographic Information Science* 45.4 (2018), pp. 362–376. DOI: 10.1080/15230406.2017.1356242. URL: http://www.tandfonline.com/action/journalInformation?journalCode=tcag20.

[17] Christian Reuter and Marc-Andre Kaufhold. "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics". In: *Journal of Contingencies and Crisis Management* 26.1 (Mar. 2018), pp. 41–57. DOI: 10.1111/1468-5973.12196. URL: http://doi.wiley.com/10.1111/1468-5973.12196.

[18] Stefan Stieglitz et al. "Sense-making in social media during extreme events". In: *Journal of Contingencies and Crisis Management* 26.1 (Mar. 2018), pp. 4–15. ISSN: 14685973. DOI: 10.1111/1468-5973.12193. URL: http://doi.wiley.com/10.1111/1468-5973.12193.

[19] Kevin Stowe et al. "Improving Classification of Twitter Behavior During Hurricane Events". In: *the 6th International Workshop on Natural Language Processing for Social Media*. Melbourne, Australia, 2018, pp. 67–75. URL: http://aclweb.org/anthology/W18-3512.

A Precision, recall, and F-1 score for each class

Figure 5: precision

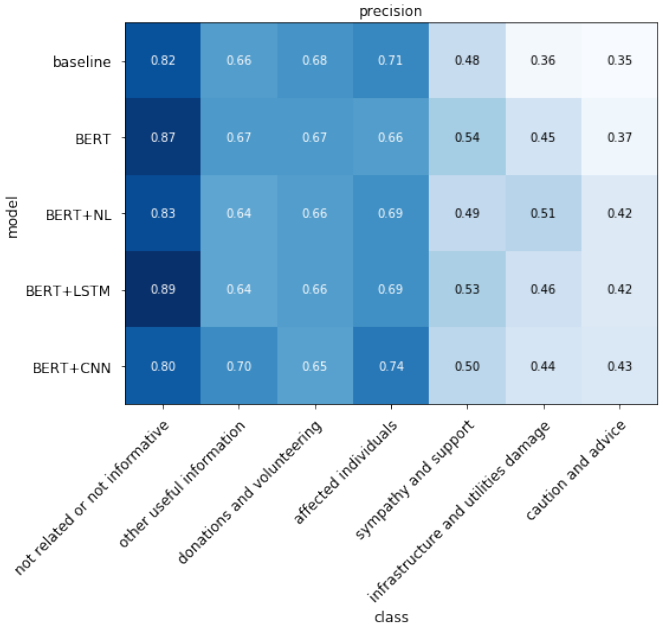


Figure 6: recall

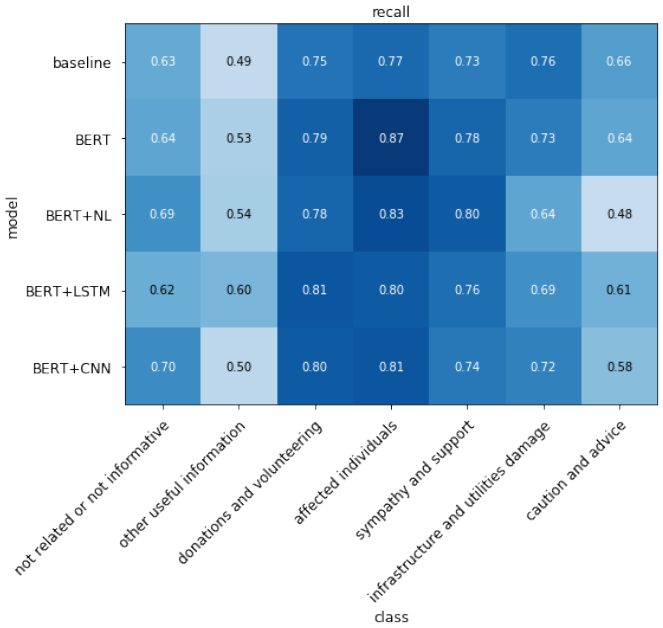
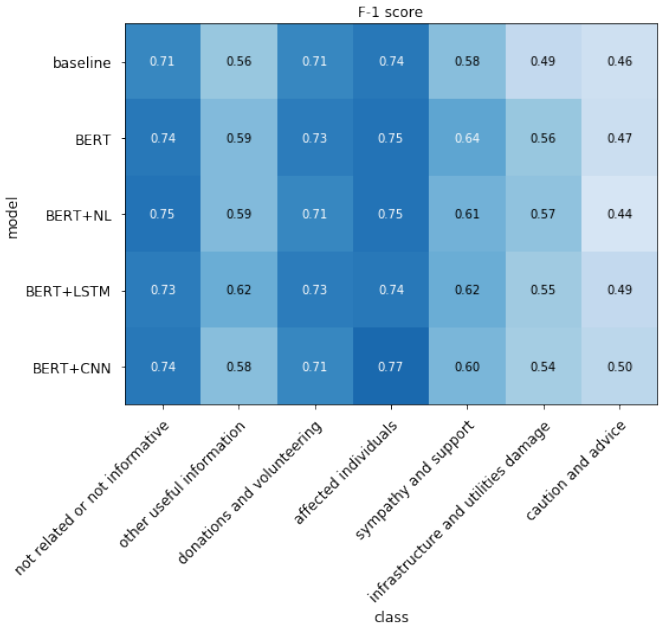


Figure 7: F-1 score



B Distribution predicted probabilities of each class

Figure 8: Not related or not informative

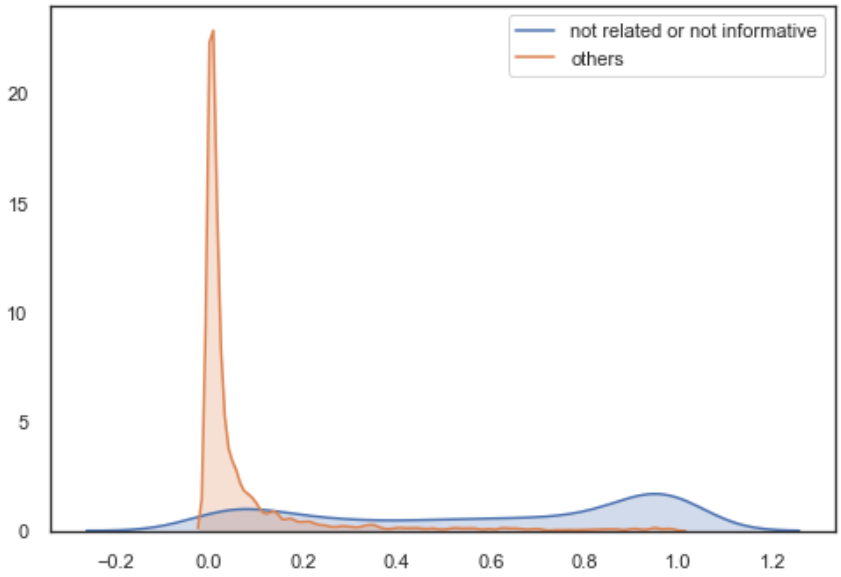


Figure 9: Other useful information

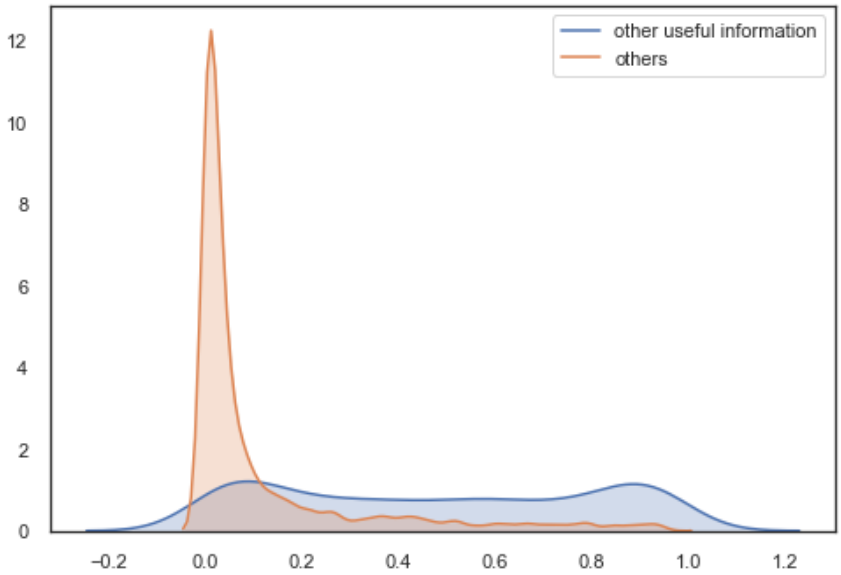


Figure 10: Donations and volunteering

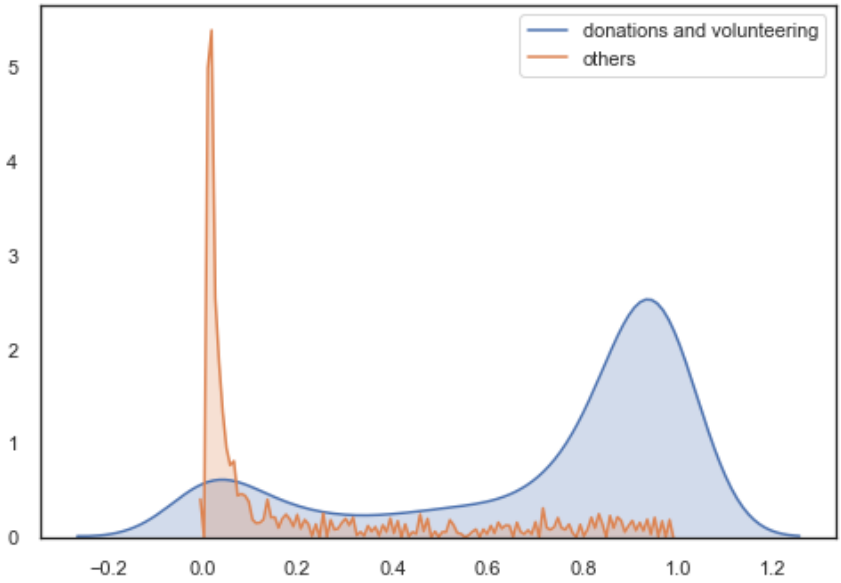


Figure 11: Affected individuals

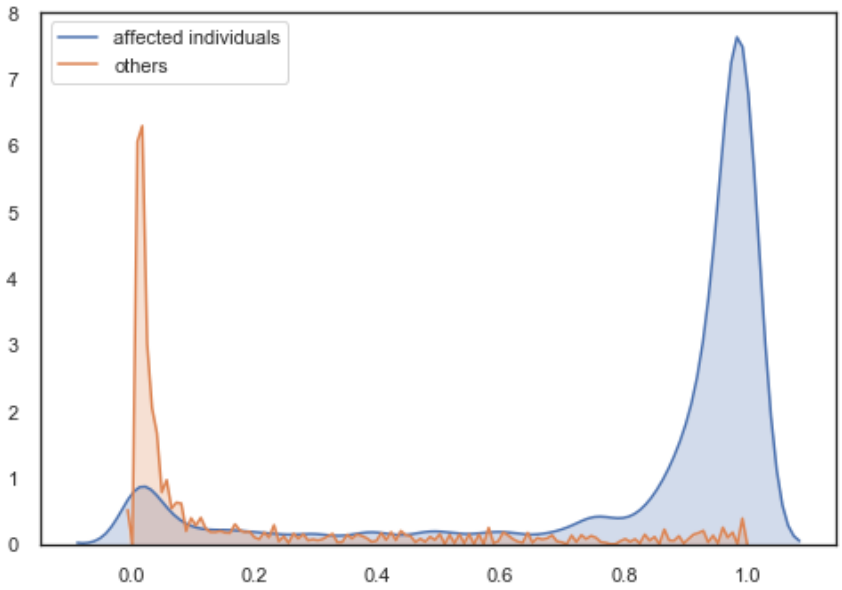


Figure 12: Sympathy and support

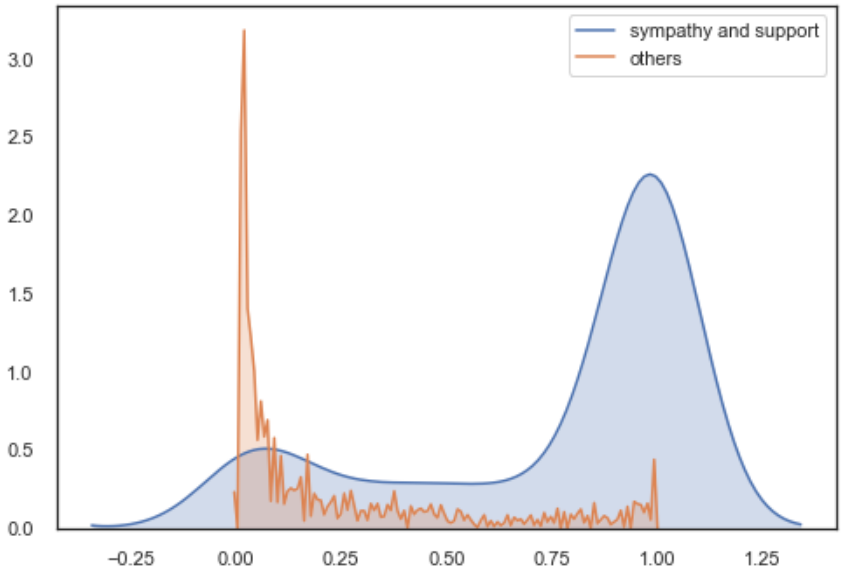


Figure 13: Caution and advice