Tweets Classification with BERT in the Field of Disaster Management

Guoqin Ma sebsk@stanford.edu

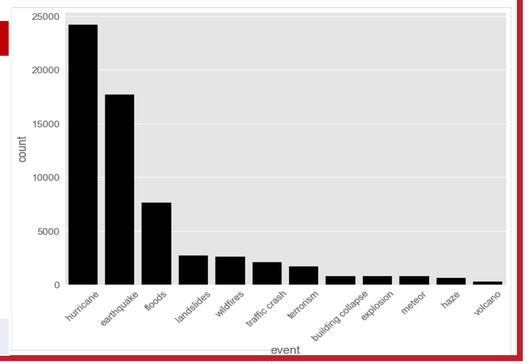
Motivation

- Social media provides vast and timely information from users during disasters
- The information is noisy on social media, not all proper to use for decision making
- ★ An accurate text classifier is required to perform filtering before analysis

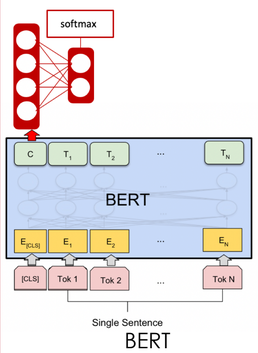
Tweet Dataset

The data are collected from various datasets on CrisisLex and CrisisNLP.
Train:dev:test = 66346:4000:4000

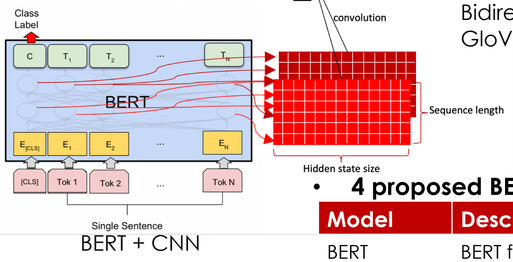
Label	Size
not related or not informative	25785
Other useful information	18877
donations and volunteering	8925
affected individuals	8009
sympathy and emotional support	5020
infrastructure and utilities damage	4559
caution and advice	3171
Sum	74346



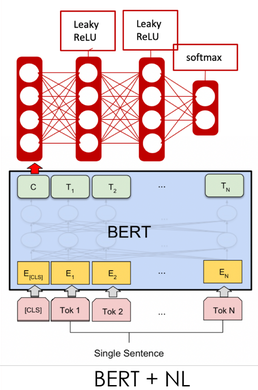
Approach



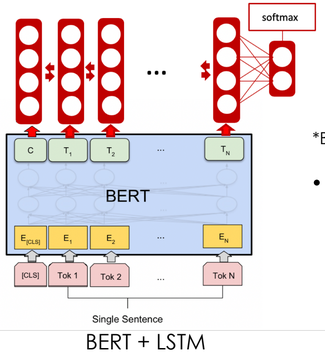
BERT



BERT + CNN



BERT + NL



BERT + LSTM

- **Evaluation Metrics**
  accuracy, Matthew coef, precision, recall, F1-score.
- **Baseline model**
  Bidirectional LSTM with GloVe Twitter embedding
- **4 proposed BERT-based model**

Model	Description
BERT	BERT for seq classification
BERT+NL	[CLS] final hidden state goes through 2 non-linear layers.
BERT+CNN	BERT with convolution over all the layers of transformers
BERT+LSTM	BERT final hidden state as input to Bi-LSTM

*$BERT_{base}$ is used in this project

- **Loss function for all models**

$$\mathcal{L} = \boldsymbol{w_i}(-\boldsymbol{x}_{\mathrm{i}} - \log\left(\sum_{\boldsymbol{k=1}}^{\boldsymbol{n_{class}}} \boldsymbol{e}^{\boldsymbol{x_k}}\right))$$

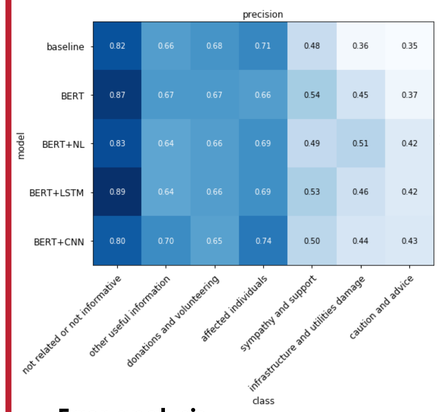
where $\boldsymbol{w_i} = \frac{\max_k class\_size_k}{class\_size_i}$,

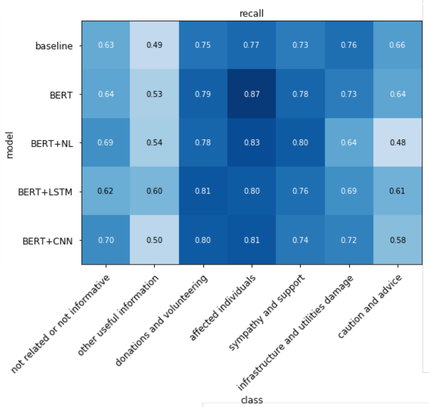
$\boldsymbol{x}_{\mathrm{i}}$ is the i-th entry of logit vector, with i as true label

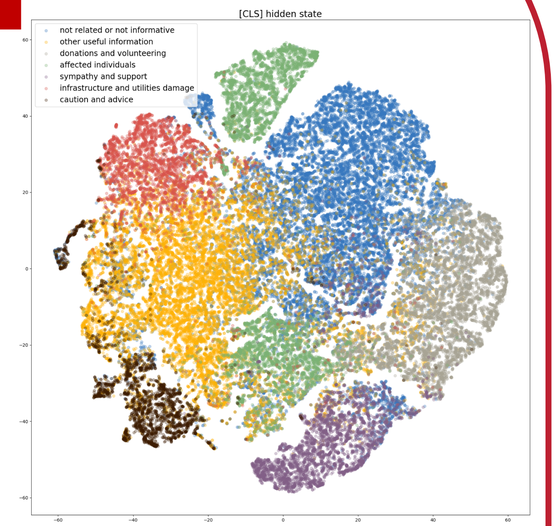
Deliverables

- **Performance**

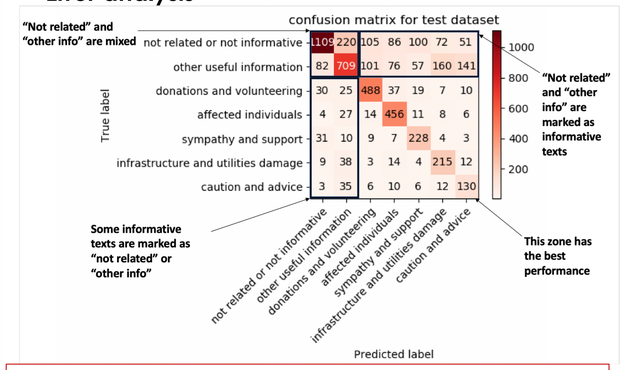
Model	Accuracy	Matthew coef	Macro precision (%)	Macro recall (%)	Macro F-1 (%)
Baseline	0.64	0.56	58.00	68.43	60.71
BERT	**0.67**	0.59	60.43	**71.14**	**64.00**
BERT+NL	**0.67**	0.59	60.57	68.00	63.14
BERT+LSTM	**0.67**	**0.60**	**61.29**	69.86	**64.00**
BERT+CNN	**0.67**	0.59	60.86	69.29	63.43

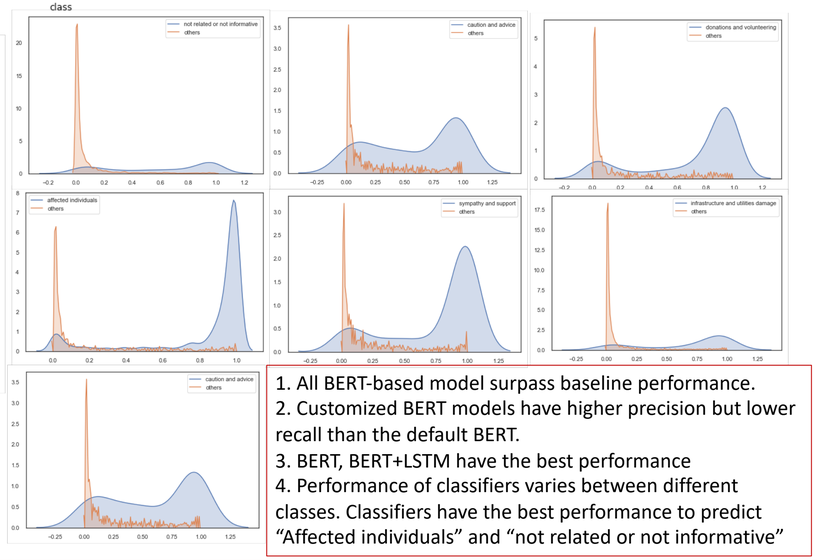






- **Error analysis**





Ambiguity and subjectivity in annotation:
2 examples of Tweets labeled as "not related or not informative":

- *"so heartbreaking , the people in tacloban was hit by Haiyan last year & now , devastated again by a new typhoon*
- *"Death toll rises to 2305 #NepalEarthquake"*

1. All BERT-based model surpass baseline performance.
2. Customized BERT models have higher precision but lower recall than the default BERT.
3. BERT, BERT+LSTM have the best performance
4. Performance of classifiers varies between different classes. Classifiers have the best performance to predict "Affected individuals" and "not related or not informative"

References

1. Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018.
2. Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. "What to Expect When the Unexpected Happens: Social Media Communications Across Crises Human Factors; Measurement". DOI: 10.1145/2675133.2675242. URL: http://dx.doi.org/10.1145/2675133.2675242.
3. 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.