Tweets Classification under Disastrous Events

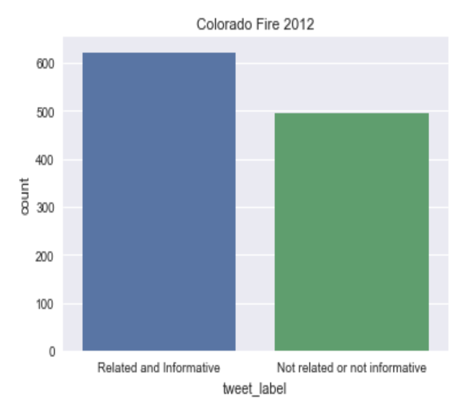
Guoqin Ma, Haotian Li {sebsk, htli}@Stanford.edu

Motivation

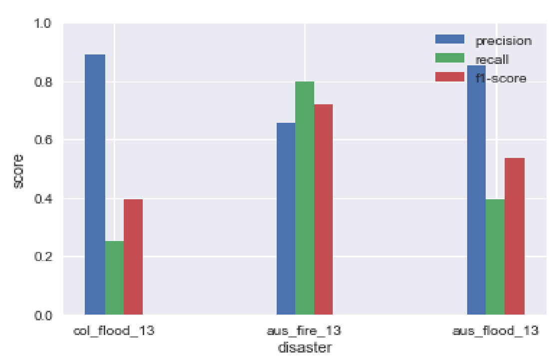
- Social media provides large volume and timely information during disasters
- Posts are not always useful
- Need to cope with big data in short time
- ★ An accurate and efficient text classifier is required

Source: The Denver Post

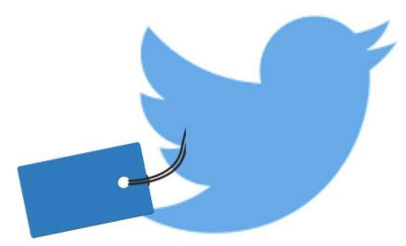
Challenge



Posts after keyword filtering are still noisy (so ML is needed)[1]



Data heterogeneity: disasters of each type have unique features[2]



Limited labeled data[2] making supervised learning difficult

Tweet Dataset

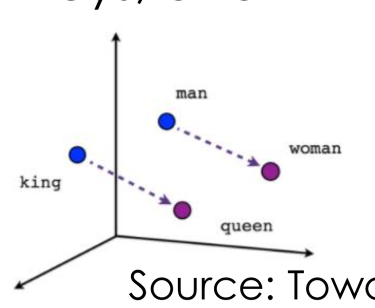
- 4 datasets picked from CrisisLexT26[3] 2012_Colorado_wildfires, 2013_Australia_bushfire, 2013_Colorado_floods, 2013_Queensland_floods
- Each contains ~1000 labeled tweets
- The tweets are filtered by keyword from tweets included in the 1% sample at the Internet Archive
- The **labels** are according to **informativeness**
- 2012_Colorado_wildfires is our training dataset. We would like to pretend that we only have limited data for 1 event and try to test capability of different models trained on it to predict other events.

Huge forest fire here in Colorado up in the mountains

So I set the world on fire! (whoops, sorry about that, Colorado)

Approach

- **Data pre-processing**
  - Remove government posts, emojis, website URLs, smileys, and stopwords, lemmatize words.
  - Bag of Word for Naïve Bayes classifier
  - Word Embedding (GloVe)[4] for neural network
- **Metrics**
  Precision, recall, F1-score, and accuracy
- **ML models**
  Multinomial Naïve Bayes, Convolutional neural network (CNN), Recurrent neural network (RNN) (LSTM specifically)
- **ML packages:** scikit-learn and Keras

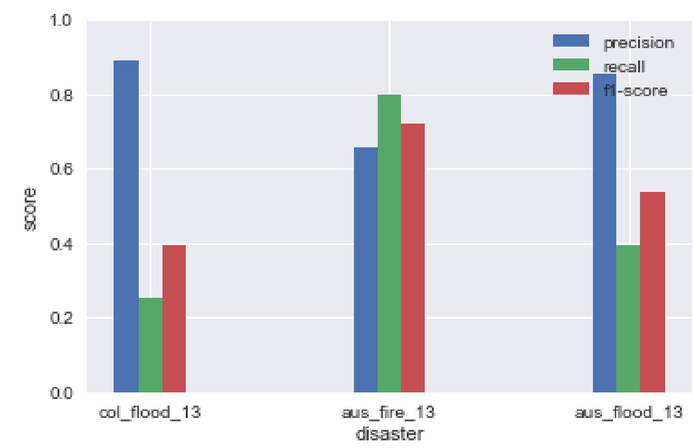


Source: Towards Data Science

Deliverables

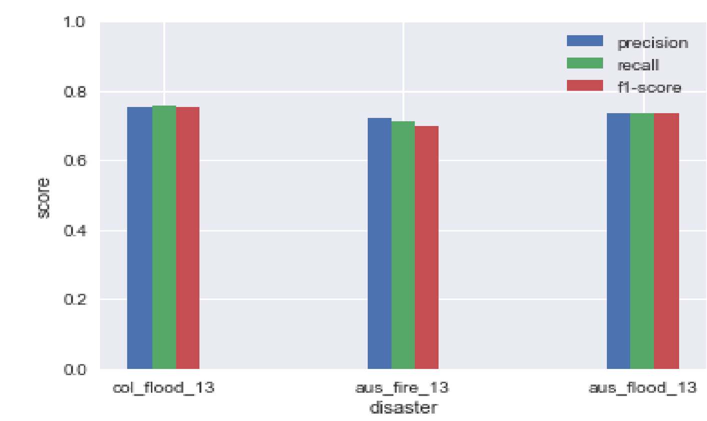
- **Comparison with Baseline**

Logistic Regression (Baseline)



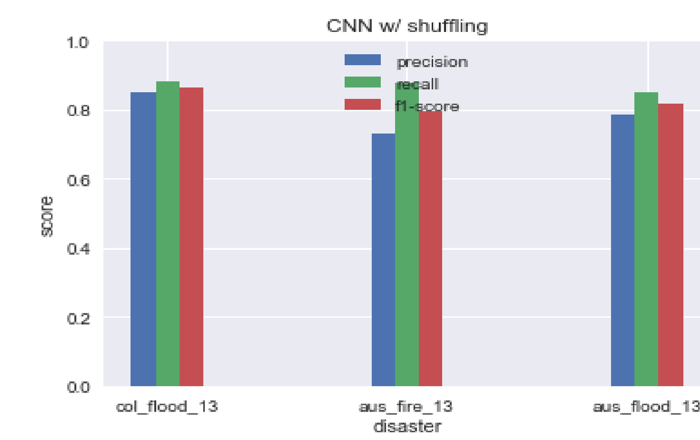
- Model learned from Colorado fire performs better to filter tweets in Australia fire
- The precisions of 2 flood events are high, while the recalls are unsatisfactory

Naïve Bayes



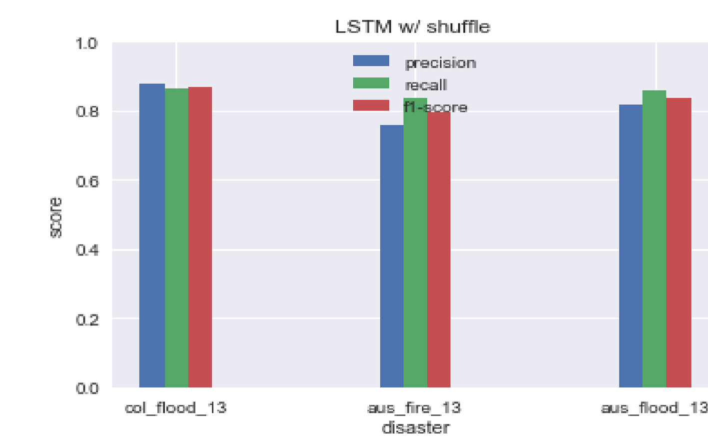
- Precision decreases slightly
- Recall and general accuracy improve remarkably
- Precision, recall and F1-score are quite close (stable)

CNN



- Faster in train and prediction than LSTM
- Results are stable in different events, no heavily affected by disaster type
- Easily overfitting, regularizer needed

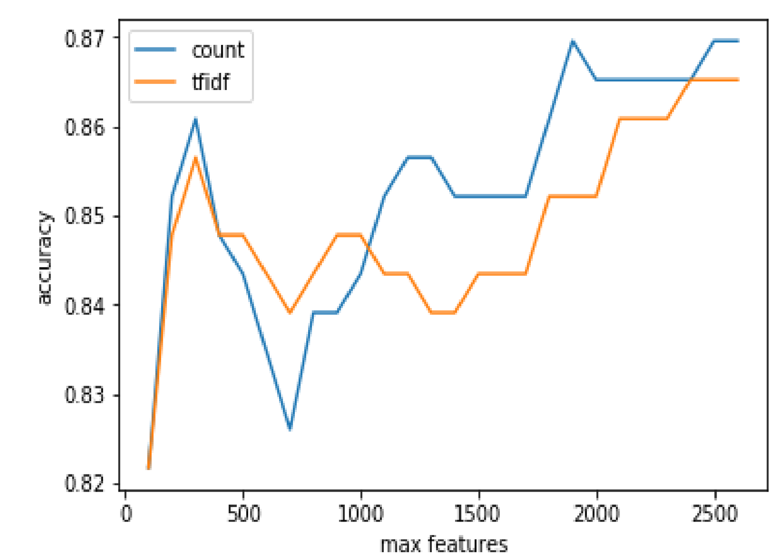
LSTM



- Perform well in different events, good incremental learning results, less sensitive to overfitting
- Run-time heavily depends on dropout and regularizer

- **Improvement of accuracy for each method**

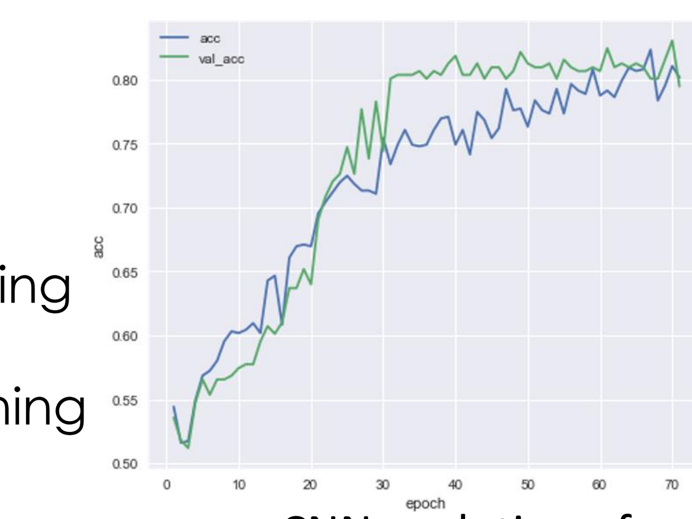
Naïve Bayes:
- Using Pre-processed texts
- Keep stopwords
- Limiting on the number of features could work for some dataset

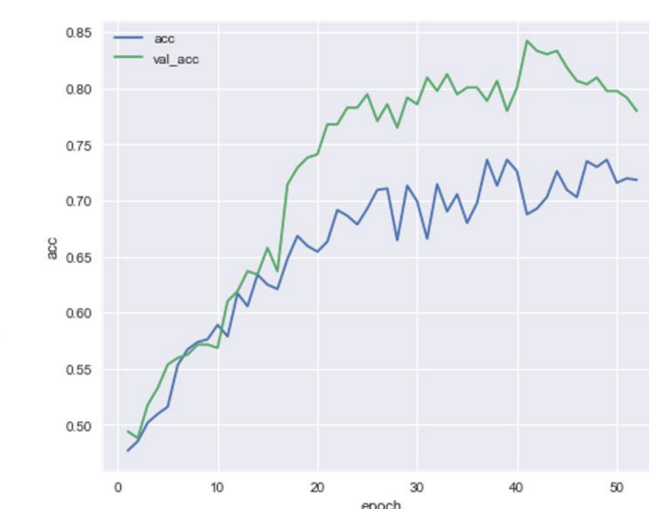


Neural Network:

Data Shuffling:
Reordering samples, re-spliting validation set during training can increase accuracy, fastening training time, averse overfitting

Incremental Learning:
Predict labels of new data, add the confident ones to train dataset to let the model adapt to the new event





CNN evolution of accuracy during training
Left: No shuffling Right: with shuffling

Event	Accuracy		
	before	unweighted	weighted
2013 Colorado flood	0.8059	0.8168	0.8190
2013 Australia fire	0.7618	0.7702	0.7692
2013 Australia flood	0.8004	0.8022	0.8076

LSTM with incremental learning

Future Work

1. Fine-tuning hyperparameters
2. Refine incremental learning method for both naïve Bayes classifier and neural network
3. Repeat some tests to prevent experiment contingency, perform statistical tests such as 2-sampled t-tests to justify our findings.
4. Find a way to shorten training time of LSTM without losing accuracy.
5. Include more types of event into our study.

Reference

1. Leykin, Dmitry, Mooli Lahad, and Limor Aharonson-Daniel. "Gauging Urban Resilience from Social Media." International Journal of Disaster Risk Reduction, April 2018.
2. Li, Hongmin, Doina Caragea, Cornelia Caragea, and Nic Herndon. "Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach." Journal of Contingencies and Crisis Management 26, no. 1 (March 2018): 16–27.
3. A. Olteanu, S. Vieweg, C. Castillo. 2015. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In Proceedings of the ACM 2015 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '15). ACM, Vancouver, BC, Canada.
4. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.