Identifying Informative Twitter Users During Natural Disasters

Chloe Larkin*
larkin.ch@husky.neu.edu
Khoury College of Computer Sciences, Northeastern University

Jiahui Zhang*
zhang.jiahu@husky.neu.edu
Khoury College of Computer Sciences, Northeastern University

Ian Magnusson*
magnusson.i@husky.neu.edu
Khoury College of Computer Sciences, Northeastern University

ABSTRACT

The use of natural language processing (NLP) models to classify informative social media posts can aid emergency services during disasters by providing timely situational awareness. We seek to build upon [3]'s achievement of classifying tweets from the CrisisMMD dataset of labeled tweets from natural disasters at an accuracy rate of 0.811 to 0.869 and AUC of 0.872 to 0.9 (using language features from individual tweets only). We apply novel MLP and LSTM models on the same dataset that incorporate unlabeled user history context. Our efforts so far have not surpassed previous accuracy and AUC, but we nevertheless offer these modest contributions to the field:

(1) Producing results not far behind the LightGBM approach used by [3] following an alternative approach with our MLP and LSTM architectures.
(2) In some cases our models that incorporate user history achieve higher accuracy data points with middle length user histories that exceeds the performance of our non-history baseline. While our history models do not consistently achieve a higher accuracy and AUC than non-history models across the complete test sets, we believe that this increased performance on history-rich users indicates possible room for further improvement meriting further investigation.

1 PROBLEM DESCRIPTION

Task

In a natural disaster, emergency services have the urgent tasks of searching for people who need help, assessing infrastructural damage, and coordinating volunteer efforts. Enhancing the situational awareness needed for these tasks through information from social media has long been a subject of research [6]. Public platforms such as Twitter enable victims, volunteers, and other informed parties outside of organized institutions to communicate on-the-ground needs and knowledge in real time.

However, identifying informative content amidst an outpouring of media attention, sympathies, and even political statements about the event requires more sophisticated filtering than selecting relevant hashtags (e.g., "#Hurricane-Maria"). To this end, many works have sought to filter posts for relevancy and type of disaster-related information using language features within individual microblog posts [2, 7]. A multimodal human-annotated dataset of tweets surrounding natural disasters, the CrisisMMD dataset, has recently been made publicly available for use of researchers seeking to train filtering models for tweet relevancy. Building upon recent work training relevancy classification models on labeled tweets in the CrisisMMD dataset [3], we examine the effect of expanding the window of both image and text features to include a user's whole history of posts during the disaster. Extracting image features from tweets will be the subject of separate research, whereas this project focuses exclusively on natural language features.

Input and Output

We have operationalized the problem as follows: A target tweet and the history of other posts by the author of that tweet are the **inputs** to our models. These produce a binary classification of the target tweet, indicating whether the tweet is informative to emergency services as an **output**.

Examples

Table 1 contains examples of informative and non-informative tweets from the CrisisMMD dataset. An informative tweet contains information pertinent to humanitarian aid efforts: cautions, advice, and warnings, injured, dead, or affected people, rescue, volunteering, or donation request or effort, damaged houses, damaged roads, damaged buildings; flooded houses, flooded streets; blocked roads, blocked bridges, blocked pathways; any built structure affected by earthquake, fire, heavy rain, strong winds, gust, etc., and disaster area maps [1].

2 RELATED WORK

Prior works have classified the relevancy of tweets surrounding natural disasters with varying strategies. This paper primarily builds upon the findings of [3], who used the CrisisMMD dataset to classify approximately 16,000 tweets surrounding natural disaster events as either informative or

*All authors contributed equally to this research.

Tweet text	Informativeness Rating
@Fox26Houston Flooding in Bellaire. #HurricaneHarvey https://t.co/vNA7TNaPqK	Informative
Harvey destroyed Dairy Treet. I'm officially pissed https://t.co/7biA2Xj9vL	Informative
When hurricane Harvey is about to hit but you remembered you left your DVD copy of Shrek 3 at home	Not informative
When we get back to SCHS after Harvey hits https://t.co/kHMnnURUAA	Not informative

Table 1: Examples of informative and non-informative tweets in the dataset.

not informative. They extracted TF-IDF and GloVe word embedding feature vectors from each tweet in the CrisisMMD dataset in to achieve an overall classification accuracy rate of between 79% and 85.6% per natural disaster (they also achieve still higher accuracies by incorporating image features).

While [3] solely used features extracted from labeled tweets when developing their classification model, a literature search of classification schemes in social media data suggests that including additional context about each tweet's author can improve model performance. For example, [8] found that including features that described users' prior posts improved binary predictions of whether a user would remain active in a Twitter or Reddit conversation. They found that F1 scores increased as longer user histories were used. They further noted that user history is more easily extractable than data regarding a user's network, and that user history may be a widely accessible set of features in social media research.

More specifically in our task's domain of disaster-related tweets, [5] found that including unigram features from a user's prior two tweets improved accuracy during a relevancy classification task. A follow-up study by the same team [4] noted difficulty incorporating a wider context window with recurrent neural network approaches. They speculate that, although manual inspection of the data shows that context from past posts is frequently crucial to understanding a particular post, useful context does not occur at regular locations within a user history—making it difficult for models to learn what parts of a history to consider. However [4] are identifying preparation and evacuation behaviors, which we believe may be a problem with much more complex contextual structure than informativeness classification.

3 METHODOLOGY

Gathering and Preprocessing the Data

We used the same dataset utilized by [3], the CrisisMMD dataset, which contained tweets surrounding seven natural disasters in 2017: Hurricane Maria, Hurricane Harvey, Hurricane Irma, California wildfires, Sri Lanka flooding, the 2017 Iraq/Iran earthquake, and the 2017 Mexico earthquake. Inspection of the CrisisMMD dataset found 16,097 tweets which were each represented by a JSON file that included the text of the tweet, at least one image, timestamps, and additional metadata. The dataset also included a key of classifications in which each tweet was assigned one of the following labels: "informative," "not informative," and "don't know / can't tell." To clean the data, we removed tweets assigned a "don't know / can't tell" classification from our dataset; this reduced the dataset by seven tweets to 16,090. Three major hurricanes (Harvey, Irma, Maria) that are inspecting by this paper have in total 9,742 informative tweets and 3,784 non-informative ones.

Further inspection of labeled dataset found that the mean user history length, or number of tweets per user, in the dataset was lower than 1.25, with a median length of 1. Our judgment was that this was too short a user history length to meaningfully evaluate the effect of considering user histories in our models, particularly given that over half of the tweets' users had no other tweets recorded in the dataset. To extend user histories, we used the Tweepy Twitter API to extract the full information of tweets for each author in the dataset from a list of unlabeled tweets IDs also included in the CrisisMMD dataset. The unlabeled portion of the dataset contains all of tweets originally retrieved by the dataset's authors using hashtag and keyword searches within a window of time around the disaster. The labeled dataset was sampled from this much larger collection of tweets. We utilize this unlabeled set of tweets rather than retrieving user histories directly from users timelines because the latter approach requires use of Twitter's paid historical tweet functionality. Nevertheless our new approach increased the size of the dataset from 16,090 to 145,253 tweets across all events, and increased the mean and median user history lengths to 12.88 and 3 respectively. Table 3 summarizes the resulting increases in user history length from incorporating these unlabeled tweets.

Following the collection of additional data, we initiated preprocessing of tweets' raw text into feature vectors. All tweet text was transformed to lowercase and stopwords were removed. We follow [3] in using TF-IDF and GloVe word embeddings features in order to measure our results against theirs. We use GloVe word embeddings to generate a representation of a tweet by taking the mean of all the embeddings of the words in the tweet. Additionally they use SVD on these embeddings to ensure that neither feature dominates due to quantity of features. Thus TF-IDF is truncated to the 200 most significant components of an SVD that is fit on all labeled and unlabeled tweets occurring in a given training set, matching the dimensionality of the 200 dimensional GloVe embeddings trained on 27 billion tweets that we also follow [3] in using. Finally the post-SVD TF-IDF and GloVe vectors are simply concatenated to form a feature vector for a given tweet.

Models

To best investigate the effect of incorporating user histories into tweet classification, we compare (1) our primary LSTM model, which classifies a target tweet based on the sequence of tweets by its author up to and including the target tweet, against (2) a baseline that only includes features from the target tweet and (3) another baseline that incorporates user history but ignores the order of the tweets.

(1) The LSTM model takes inputs from chronologically sorting the user histories, truncating the history at the target tweet, and extracting the TF-IDF and GloVe feature vectors as previously described. Drawing inspiration from the success of [8] in using bi-directional LSTMs with bi-attention for modeling sequential social media histories, we decided to implement an LSTM as our main model. We implement both bi-directional and uni-directional variants and treat the directionality as a hyperparamiter in our optimization search (described in the next section). Due to time constraints we do not implement a bi-attention mechanism.
(2) Our simplest baseline mirrors the approach of [3] by considering only the TF-IDF and GloVe feature vectors for a particular target tweet. We classify using a multilayer perceptron (MLP) model following inspiration from the finding of [4] that a multilayer perceptron was a stronger classifier of natural disaster tweets than an SVM baseline.
(3) Our more complex baseline concatenates the TF-IDF and GloVe feature vectors of the target tweet with mean TF-IDF and GloVe feature vectors derived from the full user history of the author of the target tweet. These mean vectors are formed by first combining all tweet text in the user history of the author being considered into one document. Then for TF-IDF we vectorize the whole combined document, or for GloVe we take the mean embedding of all words in the combined document. finally we apply SVD to TF-IDF vectors to truncate as described above. Finally the combined target and history feature vectors are the input for classification by a MLP model. This aggregated user history feature allows us to measure the impact of simply including user history features *en masse* in comparison to learning from the order of those features as our LSTM model attempts.

4 EXPERIMENTS

Cross-Validation

Following [3] we break each disaster each disaster event in to an 80/20 split for training and evaluation. We further split the 80% partition with ten-fold cross-validation. We devised a strategy that guaranteed an approximately equal number of labeled tweets allocated to each fold while ensuring that by a given user were not separated between different folds by itteratively appending all tweets by a user until reaching the desired fold size.

We used our ten-fold cross validation to conduct a randomized hyperparameter search, wherein sets of hyperparameters were sampled uniformly within specified ranges for each hyperparameter (chosen to allow a breadth of possibilities while staying within our computational means). Each set of hyperparameters was used to create models on all 10 folds and the resulting accuracies on the validation sets were averaged. 200 sets were conducted for each model on each event, and the hyperparameter set achieving the highest average accuracy in each was selected for use in evaluation. The hyperparameters for our LSTM model were: number of hidden neurons (50-500), number of stacked LSTM layers (1-2), directionality, learning rate (.0001-.1), momentum (.1-1), epochs (1-4). The hyperparameters for our baseline MLPs were all the same with the same ranges except activation function (relu, sigmoid, or tanh) replaces directionality.

Evaluation Method

Due to our interest in comparing our model's success relative to the models created by [3], our primary evaluation metrics for both our baseline and LSTM models were accuracy and AUC. Further, we compared our model's average accuracy to the accuracy achieved by a naive classifier that makes randomly chooses positive with a probability equal to the proportion of positive samples. These evaluation methods were proper for calibrating system performance as they allowed for a direct comparison among our models and those of [3], despite differences in model structure. Additionally, the ease of calculating model accuracy allowed for efficient searches for the strongest model parameters.

Event	Informative Tweets	Non-Informative Tweets	Total Tweets	Words per Informative Tweet
Harvey	3334	1109	4443	13.59
Irma	3564	957	4521	14.7
Maria	2844	1718	4562	14.2
All Hurricanes	9742	3784	13,526	14.2

Table 2: Comparison of number of informative, non-informative, and total tweets across each hurricane event, as well as number of words per informative tweet.

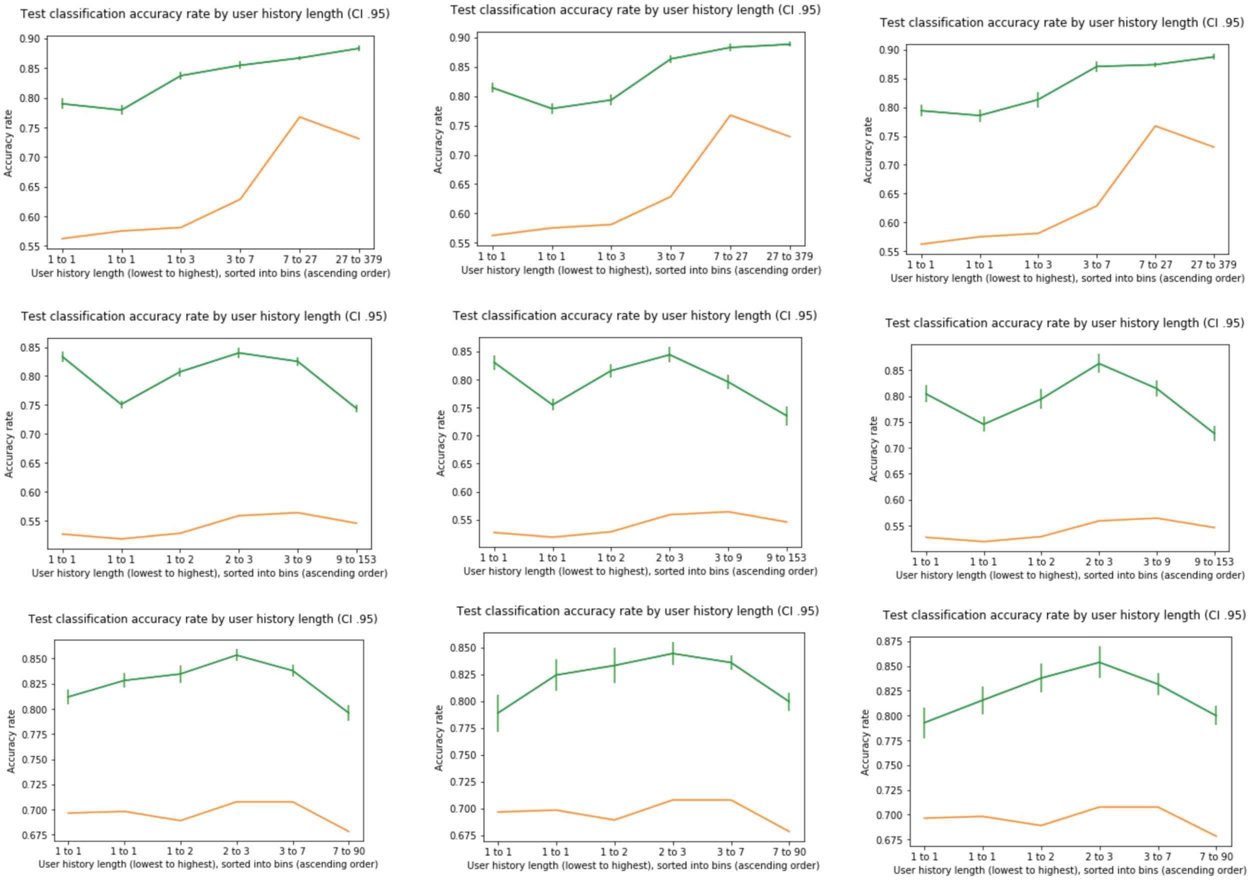

Figure 1: Comparison of model accuracy (green line) vs. naive classifier accuracy (orange line) by model. First row (left to right): Harvey LSTM, Harvey baseline with histories, Harvey baseline without histories; Second row (left to right): Maria LSTM, Maria baseline with histories, Maria baseline without histories; Third row (left to right): Irma LSTM, Irma baseline with histories, Irma baseline without histories. Error bars on the model accuracy lines represent 95% confidence intervals.

Model Performance

To varying degrees of success, our models approached the accuracy rates achieved by [3] as depicted by mean accuracy and AUC values over 100 trials in Table 2. Notably the event in consideration seems to make a large impact on the difficulty of accurate classification. Apparently our models do not learn any consistent improvement over the entire test sets from incorporating user histories.

To investigate further we examine the relationship of accuracy to the length of the user history available for the tweet being classified. Across all events, our second MLP baseline and our LSTM models demonstrated greater accuracy when classifying target tweets with a history length greater than one tweet. In Figure 1, we have separated accuracy rates for each model by user history length in a series of six bins from shortest to longest history, with each bin containing 1/6 of the data points in the test set. The blue line represents the mean accuracy rate achieved by our models across 100 runs, along with error bars indicating the 95% confidence interval. The incremental depiction of accuracy across user history lengths exposes improvements in model accuracy with increased user history length, with two of the three events achieving their peak classification accuracy at a history length of three images.

However further investigation revealed a relationship between user history length and the percent of samples that are labeled positive in the dataset. Particularly in the Harvey data, as lengths of a users history increase more and more of tweets by those users are labeled postive. Our datasets already have a bias towards positive samples (roughly 70% across events) that we have elected not to remove as it correctly represents the proportion of informative tweets within the gathered data. But it is possible that by further increasing the number of positive samples, the bins with longer user histories are simple taking greater advantage of the bias built into the model. To visualize this effect we have superimposed (in orange) the performance of a naive classifier that randomly selects positive with a probability equal to the proportion of true positives in this bin. While this does not represent a realistic model because it has knowledge about the class proportions in a test set, it can help illuminate the magnitude of this possible effect on filtering towards the model's bias. Nevertheless we can see that our model still outperforms this baseline at all points and did not improve linearly with the naive classifier, a positive outcome that speaks to our model having insights beyond frequency of each class. Of particular note is the performance of our LSTM model on the Harvey data where it gains accuracy on medium length user histories before any significant increase in positive class proportion. Harvey is also our dataset with the most unlabeled tweets and longest mean user histories, so we believe this performance might best reflect possible outcomes on data with more fully available user histories.

5 FUTURE DIRECTIONS

These findings indicate that incorporating user histories into classification model is a promising direction for the CrisisMMD classification task. We foresee several future directions that could benefit the performance of the model.
A) Refine the dataset by balancing the ratio of informative and non-informative tweets, extending user histories, and eliminating outlier histories (history length > 30).
B) Experiment with adding an attention mechanism to our LSTM model to determine whether further layers increase model accuracy.
C) Run seven-fold cross-validation, where each fold represents a held out event from the set of seven natural disasters, to test the effectiveness of the model on an entirely held-out event.

Event	Accuracy	AUC
Harvey - Nalluru et al.	0.869	0.900
Harvey - LSTM	0.835	0.841
Harvey - With Histories	0.8366	0.8440
Harvey - Just Labeled	0.8372	0.8401
Maria - Nalluru et al.	0.811	0.8806
Maria - LSTM	0.8003	0.8424
Maria - With Histories	0.7959	0.8486
Maria - Just Labeled	0.7917	0.8464
Irma - Nalluru et al.	0.853	0.872
Irma - LSTM	0.8270	0.7756
Irma - With Histories	0.8210	0.7631
Irma - Just Labeled	0.8218	0.7572

Table 3: Comparison of accuracy rates and AUCs among [3]'s single modality language feature model, our LSTM model, our baseline model when including user histories, and our baseline model when including only the labeled tweets

In addition, our extracted features here will be incorporated into a multi-modal project for another course, CS6140, which will integrate both text and image feature vectors into one cohesive tweet classifier.

6 ACKNOWLEDGMENTS

We would like to thank Professor Wang and our TAs for their support and feedback throughout the semester.

7 CODE

Please find our code available at this repository:
https://github.com/MangeFre/IITUDND

8 DATA

Please find our data available here:
bit.ly/2DYO8J7

REFERENCES

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *CoRR* abs/1805.00713 (2018). arXiv:1805.00713 http://arxiv.org/abs/1805.00713

[2] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical Extraction of Disaster-relevant Information from Social Media. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13 Companion)*. ACM, New York, NY, USA, 1021–1024. https://doi.org/10.1145/2487788.2488109

[3] Ganesh Nalluru, Rahul Pandey, and Hemant Purohit. 2019. Relevancy Classification of Multimodal Social Media Streams for Emergency Services. In *IEEE International Conference on Smart Computing, SMARTCOMP 2019, Washington, DC, USA, June 12-15, 2019*. 121–125. https://doi.org/10.1109/SMARTCOMP.2019.00040

[4] Kevin Stowe, Jennings Anderson, Martha Palmer, Leysia Palen, and Ken Anderson. 2018. Improving Classification of Twitter Behavior

During Hurricane Events. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Melbourne, Australia, 67–75. https://doi.org/10.18653/v1/W18-3512

[5] Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and Categorizing Disaster-Related Tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Austin, TX, USA, 1–6. https://doi.org/10.18653/v1/W16-6201

[6] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1079–1088. https://doi.org/10.1145/1753326.1753486

[7] Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark Cameron. 2012. ESA: Emergency Situation Awareness via Microbloggers. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2701–2703. https://doi.org/10.1145/2396761.2398732

[8] Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019. Joint Effects of Context and User History for Predicting Online Conversation Re-entries. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2809–2818. https://doi.org/10.18653/v1/P19-1270