Progress Report: 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

The goal of this project is to design a model that uses natural language processing methods to accurately classify whether a tweet is informative regarding a natural disaster, in hopes that first responders to the disaster can use this model to identify tweets that can aid them in their response. Specifically, we hope to build upon [2]'s achievement of classifying such tweets with a 0.811 to 0.869 accuracy rate and 0.872 to 0.9 AUC (using language features only) by applying novel models to the CrisisMMD dataset of labeled tweets surrounding natural disasters. In this progress report, we will describe our data preprocessing methods, preliminary modeling, results, and next steps to improve upon our current best accuracy rates. Our efforts so far have yielded a baseline models with a accuracy rates of between 0.7984 and 0.8457 and an AUC of between 0.7701 and 0.8623 across different disasters.

1 CHANGES TO PROJECT

A central goal of ours is to incorporate user histories, or additional tweets published by a user within the same timeframe of their labeled tweet, into our models in hopes that this additional information will improve classification accuracy for each labeled tweet. We began this task by preparing data for hurricane events (Hurricanes Harvey, Irma, and Maria), which was the event group with the highest number of tweets in the dataset. Our measurement of user history lengths among tweets from the hurricane events found that the mean user history length was lower than 1.25 tweets for each event, and median and mode history lengths were 1 tweet. This indicated that the vast majority of users were only represented with one tweet in the dataset. To adapt to this limitation in the data so that we may still represent user histories in our model, we have extended our dataset to include additional unlabeled tweets pulled from Twitter for each user in the CrisisMMD dataset. These supplemental tweets extended our dataset from 16,090 labeled tweets to 145,253 total tweets across all events.

An additional modification to our strategy is that our objective function is no longer based on developing user relevance scores, but returns to the same binary classification of tweet informativeness used by [2] so that we can compare our accuracy rates to theirs more directly. However, our model will remain adaptable for building user relevance super-scores as it will output classifications of each tweet, which can be easily aggregated into a sum for each user.

Our initial proposal mentioned using geolocation to enhance our classification model. At this point we anticipate such an addition would require more time than we have for this project, but we are keeping this in mind as a stretch goal. Additionally, although the CrisisMMD dataset includes finer-grained labels such as "humanitarian category" which describes the type of humanitarian crisis mentioned in the tweet, we plan to focus our efforts solely on the binary informativeness classification for this project keeping our results comparable with [2].

2 DATA PREPROCESSING

Data Collection and Cleaning

The CrisisMMD dataset was available to us online for free, courtesy of [1]. The dataset included one .json file for each of seven natural disasters in 2017 (Hurricane Maria, Hurricane Harvey, Hurricane Irma, Iraq-Iran Earthquake, Mexico Earthquake, Sri Lanka Flooding, California Wildfires) which each contained metadata for all labeled tweets from that event. From these .json files we extracted tweet IDs, tweet text, and user IDs for each labeled tweet. The CrisisMMD dataset also included .tsv files matching each tweet ID to a binary, human-annotated "informative" / "uninformative" classification score. For data cleaning purposes, we excluded labeled tweets with the annotation "don't know / can't be sure" regarding informativeness. This removal reduced the dataset by only 6 tweets.

The CrisisMMD dataset also provides the raw list of tweet IDs that their labeled dataset is sampled from. These tweets were collected using keywords and geo-graphical information related to a given disaster event. In order to augment our user histories, we collect all tweets from this unlabeled data that are authored by users appearing in the labeled dataset. This approach collected 129,163 new unlabeled tweets using the "tweepy" twitter API to retrieve full tweet data. Across the tree hurricane events studied by [2], the mean user history length was 12.88, median length was 3, and mode length was 4.333 (see table 1).

*All authors contributed equally to this research.

Event	Mean	Median	Mode
Hurricane Harvey	22.60	4.0	2
Hurricane Irma	8.05	3.0	2
Hurricane Maria	7.99	2.0	1

Table 1: User history lengths by event after retrieving unlabeled data

To fully prevent data leaking, we split the dataset into train and test sets after collecting additional unlabeled tweets, but before beginning feature extraction. When separating the data, we ensured that the total number of labeled tweets in the test set reflected the proportion we sought (20%, to reproduce the approach used by [2]), but also that no user history was split between the train and test set. To that end, we randomly selected users to include in the test set, until the total number of labeled test tweets filled the test set with 20% of labeled tweets. The final results of this splitting process is two pairs of lists, one pair for training and one for testing, where A) the first list in each pair is all the labeled tweets by users in that set and B) the second list in each pair is a list user histories corresponding to the author aligned labeled tweet in the first list of the pair.

Finally we follow the procedure of [2] to preprocess all tweet text to lowercase and to remove stop words.

Feature Generation

Following the feature design of [2] we extract GloVe embedding [3] and TF-IDF features. We extend this approach to extracting features from our user histories as well.

First we extract mean GloVe embeddings for each labeled tweet and each user history to produce two feature matrices. We use follow [2] in using the 200 dimensional embeddings trained on Twitter data (27 billion tweets). The features for the labeled tweets are simply the sum of the GloVe embeddings of all the words in the preprocessed text of a given tweet, divided by the number of words. The features for the user histories treat all preprocessed text in all the tweets authored by a user as one concatenated document. The mean word embedding is then taken on this document following the same method as described for the labeled data.

Our next step was to extract information about relative document frequency of words in each tweet and user history. To standardize labeled tweets and their histories into appropriate dimensions for a multilayer perceptron, we fit a TF-IDF vectorizer on a corpus of every labeled and unlabeled tweet in the train set (excluding tweets from the test set). We then applied the vectorizer to create matrices of vectors for two feature types: A) the TF-IDF values of each labeled tweet in the train set and B) the TF-IDF values of the user history (as a single concatenated document) of the author of each labeled tweet. We ensured both matrices represented each labeled tweet in the same order with identical dimensions, where row i in matrix B represents the user history of the author of labeled tweet i in matrix A.

Following the construction of TF-IDF matrices for labeled tweets and user histories, we applied an SVD to reduce the dimensionality of each matrix. Our preliminary choice is to truncate to the 200 most important components, to have equal dimensionality with the GloVe features so as to avoid emphasising one or the other feature. Each resulting matrix from SVD was of shape (n labeled tweets in the train set $\times$ 200).

Finally we explored using this data unnormalized and normalized with min-max and Z-score normalization. We find that Z-score normalization consistently performs best, so we use this normalization scheme on the data from which we report our results.

3 BASELINE MODELING

The first step in our analysis has been to train a baseline model that compares the effectiveness of classifying informativeness using features solely from the tweet to be classified against an approach also incorporating user history features. As detailed in the feature extraction section, this first model ignores the sequence of tweets in a user history and instead takes mean TF-IDF and GloVe embeddings over the concatenation of all the text in a users history. This will allow us not only to see if this very basic approach to user history improves accuracy, but also to serve as a baseline for identifying the benefit of later incorporating sequence information into the features.

For the sake of simplicity and quick training, our baseline model is a simple multilayer perceptron of 600 neurons in each of two hidden layers (400 in the first layer, 200 in the second). At this point our aim is not so much to optimize performance as to compare performance between different disaster events and data with and without user history features, so we have not yet conducted a hyperparameter optimization. Likewise our learning rate (0.1), learning rate decay (0.1), and the number of epochs (2) were chosen to allow fast training and were selected to occur after the plateau in training loss observed through manual inspection.

Using this algorithm for producing models, we test the performance of our features (with and without user histories) by training 100 models on different random permutations of the same data. We take the the mean accuracy and AUC from these models to reduce the noise introduced by variations in outcomes of stochastic gradient decent optimization.

In our initial analysis, we train on each disaster event separately and only consider the three hurricanes in the dataset, so that our results will be comparable to those of [2].

Event	Accuracy	AUC
Harvey - Nalluru et al.	0.869	0.900
Harvey - With Histories	0.8457	0.8623
Harvey - Just Labeled	0.8414	0.8595
Maria - Nalluru et al.	0.811	0.8806
Maria - With Histories	0.7984	0.8430
Maria - Just Labeled	0.8048	0.8483
Irma - Nalluru et al.	0.853	0.872
Irma - With Histories	0.8190	0.7701
Irma - Just Labeled	0.8175	0.7783

Table 2: Comparison of accuracy rates and AUCs among [2]'s single modality language feature model, our model when including user histories, and our model when including only the labeled tweets

4 PRELIMINARY RESULTS

Our baseline model achieved comparable though slightly lower accuracies and AUC as those reported [2] in October 2019 (see table 2). Over 100 trials using history features, our model achieved a mean accuracy rate of between 0.7984 and 0.8457 and an AUC of between 0.7701 and 0.8623 across different disasters. While the inclusion of user history features increases accuracy and AUC on the Harvey dataset, this effect does not extend to the Irma and Maria datasets. One possible explanation is that our user histories are much longer on average for Harvey than the other two hurricanes, and thus contain more information.

The magnitude of the effect of adding user history features is also quite small, so it is difficult to determine if the impact is not just a result of noise from the stochastic gradient decent training process. A possibly relevant finding is that the increased accuracy observable on the Harvey dataset becomes smaller when normalization is not used (unnormalized Harvey accuracies : with histories = 0.8422, without = 0.8405). This might suggest that further optimization of our model creation algorithm could uncover a greater effect of the user history features.

Finally, our current use of user histories does not take into account the sequence of tweets; rather, it treats histories as a "bag of tweets" by taking the means of both GloVe and TF-IDF values for of all words in a combined document made by concatenating all tweets in a user history. This may be the reason why the effect of user histories remains small in our baseline model. Our future plans to implement a LSTMs based embedding of user histories will hopefully enhance the effectiveness of these histories in boosting classification accuracy by taking tweet chronology into consideration, as is discussed in the following section.

5 FUTURE WORK

Going forward, we have two major goals to complete: 1) achieving generalizable performance across disasters, 2) incorporating sequential information in user history features by training a LSTM sequence tagger.

First, we seek to train models that work across different events and event types. This will involve creating a new train/test split and extraction of features across all events in the CrisisMMD dataset rather than individual events. With this we seek to test if models trained on one event can achieve comparable performance on other events both within and across different types of disasters. This would be necessary for practical application of this technology, as training new models during a disaster event would be prohibitively slow.

Second, for improving our user history feature extraction above the baseline introduced in this report, we plan to use LSTMs. We follow [4] success with bi-directional LSTMs with bi-attention for modeling sequences of social media posts in a user history. We will work iteratively, by first implementing a plain LSTM approach, then a bi-directional LSTM, and finally incorporating bi-attention mechanisms (time permitting). Our LSTM will be trained as a sequence to sequence tagger of informative tweets from a user history sequence. We will then optimize on a loss function that only considers the predictions for labeled tweets within the user history, ignoring the predictions on the unlabeled tweets. We will evaluate our final model by classifying each tweet in a test set of labeled tweets using the user history up to and including that tweet as input and considering only the classification of the final tweet. This method will require us to update our baseline model by training it on truncated histories that only include tweets up until the labeled tweet.

One possible extension of our LSTM approach (requiring further research) would be to bootstrap our unlabeled data by using the LSTM tagger described above to tag all of our unlabeled data. Such a fully labeled dataset could be used to train a new LSTM tagger, this time without ignoring any of the predictions for any of the tweets in the sequence in our loss function. This may possibly increase the power of our parameter optimization by giving more information for the gradient decent on all tweets in the sequence. This idea requires further thought and we look forward to asking Prof. Wang about this possibility.

A few additional considerations may help us improve our models: Given that the baseline informativeness rate in the dataset is approximately 70% across events, we will seek to address concerns regarding bias and high minimum accuracy when evaluating our models. Additionally we will pursue a more sophisticated approach to selecting hyperparameters. We will experiment with truncating the top components in our SVDs to improve the generalizability of our models. We

will also conduct proper 10-fold cross-validation for hyper-parameter optimization in our MLP and LSTM models.

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] 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

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[4] 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