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

1 INTRODUCTION

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 much research [13]. 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 among the outpouring of media attention, sympathies, and even political statements about the event requires more sophisticated filtering than selecting relevant hashtags (e.g., "#HurricaneMaria"). Many works in Natural Language Processing (NLP) literature have sought to filter posts by classification schema for relevancy and type of disaster-related information using language features within individual microblog posts [5, 14].

To advance the state of the art in classifying relevancy of natural disaster tweets, we examine the effect of including context beyond the words of a single post. We propose a multimodal model trained on a human-annotated dataset of tweets with images. The computer vision component will be the subject of separate research, while this project will focus on the incorporation of user context from tweet history. The contributions we seek to make are:

(1) Achieving an appreciable accuracy rate in classifying tweets about an entirely different held-out natural disaster to demonstrate the generalizability of user context features,
(2) Creating a new output, a list of users that generate the highest number of informative tweets, whom emergency services may wish to reach out to for further information.

2 RELATED WORK

Our focus on user context draws inspiration from [15], who successfully incorporate user history from Twitter and Reddit posts to predict when users will stay involved in a conversation. In particular, they point out the convenience of user history as a feature that is easily extracted compared to complex features like social network information, allowing this approach to be more easily generalized to a variety of applications. Moreover, they found that F1 scores increased as longer user histories were used. While conversation is a different domain than ours, we believe that identifying informative users in a disaster could be framed as a task of predicting when a user will post another relevant tweet.

In the domain of disaster-related Twitter, [12] found that incorporating simple unigram features from the previous two tweets improved F1 scores for classification of relevant posts. A follow-up study by the same team [11] noted difficulty incorporating a wider context window in neural network approaches and speculate that issues arise because helpful context does not occur in any predictable location within a tweet history. We believe user history might be more successfully applied to predicting user-level, rather than tweet-level, relevancy, because the over-all informativeness of a user will not be as dependent on specific pieces of discourse knowledge needed to understand particular posts.

Another type of context that may be fruitful to examine is location, both of the user at the time of a tweet's posting and user-level "home" location. Research has shown that these attributes, which can be ground-truthed by Twitters geotagging and self-declared home location features, are predictable with high accuracy from language features alone [7]. More generally language features have been shown to be used to predict whether users are participants in the events they describe [10]. Users who post about a disaster at which they are present and engaged could perhaps have more relevant information to offer.

Though the dataset we will use has only been recently released, one study has already used it to investigate classification of relevancy. In [8] a combined image and language approach achieves accuracy rates between 79% and 85.6% and AUC between 84.5% and 90.7% for various disasters. This study does not consider context beyond the frame of a single post, and so can serve excellently as a baseline to measure the impact of adding user context information.

*All authors contributed equally to this research.

3 DATASET

We will use the CrisisMMD dataset described in [1], which includes 16,097 non-identical, human-annotated tweets published during seven natural disasters in 2017: Hurricanes Irma, Harvey, and Maria, as well as California wildfires, the Mexico earthquake, the Iraq-Iran earthquake, and Sri Lanka floods. English-language tweets were selected based upon their use of keywords and hashtags surrounding the disaster, and are organized in the dataset by the disaster event they were pulled from. Tweets were first given a high-level classification based on their informativeness—either informative, not informative, or don't know or can't judge. Tweets labeled as informative were given further finer-grain multiclass labels for different types of humanitarian categories. The following classes were selected by the authors based on research in humanitarian aid:

(1) infrastructure and utility damage
(2) vehicle damage
(3) rescue, volunteering, or donation effort
(4) injured or dead people
(5) affected individuals
(6) missing or found people
(7) other relevant information
(8) not relevant or can't judge

The other relevant information category was inspected by [3] and described to contain other information that may useful to responders. Finally, tweets labeled as infrastructural and utility damage are further ranked with a scale of damage severity (severe, mild, little or no damage, don't know or can't judge).

The high-level informativeness labels should be quite suitable for our proposed relevance prediction model. Representation among finer-grained labels is somewhat skewed, with some categories having few instances. These subcategories could still be used, along with negative instances of other subcategories, to train binary classifiers. Finally, we plan to make use of user identities, timestamp, and potentially geolocation metadata in the dataset to arrange user histories that can provide a wider contextual window.

Each tweet in the CrisisMMD dataset is paired with one or more images the user attached to their post, which bear the same classification labels for the informativeness and humanitarian category of the image. The authors of the dataset describe their process of harvesting these images in [2]. Notably, the labels of a tweet may be different from the labels of its associated images. We consider the possibility of creating a relevancy superscore, which reports a tweet is relevant if either its text or image is informative.

4 METHODOLOGY

Our prospective approach will build upon prior work in disaster-tweet relevance classification while introducing the new element of user history into analysis. First, we will implement models to predict relevancy at the tweet-level following existing work. The authors of [12] found that support vector machine (SVM) analysis performed best in their comparison of three supervised learning models (SVM, maximum entropy models, and naive Bayes) for classifying individual tweets using hand crafted features. Following these insights, we will explore a hand-crafted-feature baseline that iterates from simple models like Naive Bayes up to an SVM model to determine whether similar accuracy can be replicated on our dataset.

For n-gram features, [12] found that feature selection such as removing rare words and selecting words with highest pointwise mutual information improved over using all features; thus, we plan to follow this approach in out baseline model. Among the features [12] tried, word embeddings were found to have by far the greatest impact on F1 scores. While they trained their own embeddings, [8] successfully incorporated pre-trained GloVe embeddings [9] in their work the crisisMMD dataset. Thus we will further study the literature of word embeddings and plan to include this feature in our models.

Subsequent study using two deep learning approaches, Multi-Layered Preceptron (MLP) and Convolutional Neural Network (CNN), for the same classification task was conducted by the authors of [11]. They found that though both models show appreciable improvements over the SVM baseline, CNN does not improve over the basic MLP. If we do pursue a deep-learning approach, we will most likely opt for MLP.

In the second stage of our research, we will explore user context in the task of identifying the most informative users. A simple baseline for predicting whether a user will post future relevant tweets would be to assume that any user who has already posted a relevant tweet will post another one. Another baseline would be to find the fraction of relevant tweets in the user's history, and simply use this value as a probability for a relevant tweet at any future time.

For more sophisticated modeling, we plan to experiment with the following approaches. In [15], bi-attention was found to be most successful method for modeling the interaction of user history and local information from specific conversations. Individual posts were first encoded before being passed to the interaction model; the authors found bi-directional LSTMs to be the most effective method for encoding. Due to their multimodal approach, [8] face a similar task of modeling the joint interaction of dissimilar features. They use gradient boosting decision tree analysis on the

CrisisMMD dataset to combine language and image features. They test both LightGBM [6] and XGBoost [4] against a baseline logistic-regression model and find that LightGBM outperforms XGBoost. They also employ singular valued decomposition (SVD) to normalize dimensionality across feature vectors from different modalities. We will explore each of these methods as promising tools for dynamically integrating tweets, user history, and even images in analysis.

5 EVALUATION

We will take several steps to evaluate our models, first on tweet-level relevance classification and then on user-level informativeness ranking. We will evaluate binary label classifications with typical confusion matrix assessments, and will report area under the curve (AUC) metrics for any probabilistic results of our analysis. To assess skew in model accuracy among classes (i.e., disaster and humanitarian categories), we will compare microaveraging and macroaveraging accuracy results. We will then compare the accuracy rates of our classification models to those in the literature.

In our secondary analysis of user informativeness, we will derive ground truth by a sum of each user's number of informative tweets. We will explore methods to normalize this value to avoid bias towards users with many or few posts. Finally, we will compare the accuracy of any multimodal image and tweet modeling to lower-level analyses and to the success of the LightGBM multimodal model in [8].

We also hope to demonstrate the generalizability of our models on test data consisting an entirely different natural disaster type (e.g., earthquake as opposed to hurricane) unseen in the training data.

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] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Processing Social Media Images by Combining Human and Machine Computing during Crises. *International Journal of Human–Computer Interaction* 34, 4 (2018), 311–327. https://doi.org/10.1080/10447318.2018.1427831 arXiv:https://doi.org/10.1080/10447318.2018.1427831

[3] Firoj Alam, Ferda Ofli, Muhammad Imran, and Michaël Aupetit. 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018.* http://idl.iscram.org/files/firojalam/2018/1579_FirojAlam_etal2018.pdf

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[5] 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

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3146–3154. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[7] S. E. Middleton, L. Middleton, and S. Modafferi. 2014. Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems* 29, 2 (Mar 2014), 9–17. https://doi.org/10.1109/MIS.2013.126

[8] 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

[9] 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

[10] Krishna Chaitanya Sanagavarapu, Alakananda Vempala, and Eduardo Blanco. 2017. Determining Whether and When People Participate in the Events They Tweet About. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 641–646. https://doi.org/10.18653/v1/P17-2101

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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