Identifying Informative Twitter Users During Natural Disasters

CHLOE LARKIN*, Northeastern University, larkin.ch@husky.neu.edu
MAGNUS FRENNBERG*, Northeastern University, frennberg.m@husky.neu.edu
IAN MAGNUSSON*, Northeastern University, magnusson.i@husky.neu.edu

1 OBJECTIVES AND SIGNIFICANCE

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 [18]. 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 [7, 20]. Besides text, social media also contains images, and works like [11] use this data for tasks such as analyzing the extent and type of disaster-related infrastructural damage. Other works have combined images, video, and tweet text in multimodal approaches to disaster situational awareness [10, 13]

Building upon new work on the recently released CrisisMMD multimodal dataset[1], 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. Our model will identify highly informative users who can provide emergency services with valuable on-the-ground knowledge. Extracting natural language features from the text of tweets will be the subject of separate research, while this project will focus on extracting image features from a stream of user posts and fusing these with NLP features in a final user-informativeness prediction. The contributions we seek to make are:

(1) Achieving appreciable accuracy in classifying tweets about a separate natural disaster event in a held-out data set to demonstrate the generalizability of user history features,
(2) Creating a new output, a list of users that generate the highest number of informative tweets, whom emergency services may wish to contact for further information.

2 BACKGROUND

One study has already used the CrisisMMD dataset to classify whether individual posts are informative. [10]. "Informative" in this context is defined as a post that contains useful information for humanitarian aid. The study demonstrates the effectiveness of a multimodal approach with a combined image and text model improving on a purely text-based model to achieve accuracy rates between 79% and 85.6% and AUC between 84.5% and 90.7% for various disasters. However, 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 history information.

Our focus on user history draws inspiration from [21], 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

*All authors contributed equally to this research.

informative users in a disaster can nevertheless be framed similarly as a task of aggregating information about a user based on multiple posts. Thus, we seek to incorporate user history into our multi-modal analysis.

In the domain of disaster-related Twitter, [17] found that incorporating elements of user history through simple unigram features from the previous two tweets improved F1 scores for classification of relevant posts. A follow-up study by the same team [16] 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.

With regard to computer vision, much research has gone into learning temporal relationships between frames in video data. While not focused on users-level features, a recent work in multimodal classification of disaster related YouTube clips fuses temporal information from video, audio, and accompanying text descriptions, achieving an 8% increase in MAP likelihood over models from prior frameworks [13].

While frames in a video are far more related than a sequence of images posted over hours or days, work on classifying the contents of photo albums suggests that temporal relationships between distinct images still contain useful information. Such research seeks to label the events of a sequence of images, for instance identifying an album of pictures as a weekend ski trip. The authors of [15] found that using a skipping recurrent neural network (S-RNN) to learn visual-temporal sequences among images is an effective tactic to summarize groups of photos. Further work by [4] used convolutional neural networks (CNNs) to derive object and scene features from photographs, and then combined these extracted features in a probabilistic graphical model to predict album categories. While these works are not directly concerned with a disaster-related domain, they nevertheless support the notion that the combined features of a collection of user curated images can contain sequence and contextual information helpful for predicting a generalization about the group of inputs.

Time providing, 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. These attributes are can be ground-truthed by Twitter's geotagging and self-declared home location features. Work in computer vision [19] has demonstrated location prediction from images that sometimes exceeds human accuracy, especially when using recurrent neural networks to classify whole albums of photos. Location is even predictable with high accuracy from language features [9]. More generally language features have been shown to be used to predict whether users are participants in the events they describe [14]. We conjecture that users who post about a disaster at which they are present and engaged could perhaps have more relevant information to offer.

3 PROPOSED APPROACH

3.1 Dataset

We will use the CrisisMMD dataset described in [1], which includes 16,097 non-identical, human-annotated Twitter posts. Each tweet contains text and one or more images, with a total of 18,126 images among all tweets. These posts were 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. Tweets were selected based upon their use of keywords and hashtags surrounding the disaster, then filtered to posts with English-language text only, and finally organized in the dataset by the disaster event they were pulled from. Tweeted text and images were first given a high-level classification based on their informativeness for humanitarian aid efforts—either informative, not informative, or don't know or can't judge. The authors of the dataset describe their process of harvesting these images in [2]. Text and images labeled as informative were given further finer-grain, multi-class labels for different types of humanitarian categories. The following classes were selected by the authors based on research in humanitarian aid:

- infrastructure and utility damage
- vehicle damage
- rescue, volunteering, or donation effort
- injured or dead people
- affected individuals
- missing or found people
- other relevant information
- 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, images 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 informativeness prediction model. To derive user-level informativeness scores from the post-level labels, we will simply take the count of posts labeled as informative for each user. We anticipate that users' relative levels of post frequency, i.e. a prolific poster vs. a rare poster, may skew findings. Thus, we also plan to experiment with normalizing the count by dividing by the total number of posts per user. Each metric could have potentially different value for emergency services, and we base our decision on qualitative investigation of the top users by ground truth score for each metric as well as examining users with greatly differing scores under the different measurements.

Time permitting, we will incorporate consideration of finer-grain labels into our analysis. 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.

Through the use of Twitter APIs, we are also able to connect tweets to specific users based on the TweetIDs in our dataset. This allows us to connect tweets with the same author to form user histories. We hope to use this information to determine if a user is consistently tweeting relevant information. Finally, we plan to make use of timestamps and possibly even geolocation metadata in the dataset to expand the context window of our user histories.

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.

Finally in the event that the length of the user histories contained within the existing dataset is insufficient, we have the following fallback plan. We will use the Twitter API to extract more publicly available tweets from users who already appear in the dataset. Since the dataset is constructed from tweets using certain hashtags and keywords, there will likely be additional tweets within the same time period as the dataset for most of these users that were not originally captured. We will use these unlabeled tweets to perform semi-supervised learning.

3.2 Methodology

Our prospective approach will build upon prior work in multimodal prediction of disaster-tweet informativeness, while introducing the new element of user history into analysis. Here we first provide an overview of our model architecture, followed by detailed justifications for our choices. For feature extraction from the individual posts in each modality, we will follow [10] using TF-IDF and GloVe embeddings [12] for text features and ResNet [6] for image features. Our novel contribution will be to aggregate these post-level features into user history sequences. We will use these sequences to train recurrent neural networks for each modality that regress a user-level informativeness score derived from the number of relevant posts made by each user in the dataset. These recurrent networks will be used to extract user-level feature vectors based on the values of the second-to-last network layer for each user in the dataset. Finally following previous work that uses non-neural-network models for multimodal fusion [10, 13], the user-level features for each modality will be used to train an SVM or GBM which regresses a single user informativeness score.

We believe the simple feature extraction approach of [10] is supported by the success of similar methods in other work. Using word count features similar to TF-IDF, [17] found that feature selection, such as removing rare words and selecting words with highest pointwise mutual information, discovered effective n-gram features for classifying disaster relevant tweets. They also found that word embeddings had by far the greatest impact among features on F1 scores.

Meanwhile for image feature extraction [11] successfully apply VGG-16, a similar pre-trained CNN, to the task of identifying building damage in social media images of disasters.

For implementing our user-level feature extraction, we plan to use LSTMs and compare them against baseline feedforward networks for each modality.In [21], bi-directional LSTMs with bi-attention were found to be the most successful method for modeling sequences of social media posts in a user history. Meanwhile for sequences of images, [13] employ LSTMs to capture temporal relationships in disaster related videos. Nevertheless we are aware of some limitations of LSTMs—such as [15] finding that LSTMs overemphasize short-term connections to the detriment of longer distance interactions in the domain of generating storylines from photo albums, or [16] finding that LSTMs did not help to resolve long distance and irregular discourse dependencies for predicting evacuation behavior from the tweets of disaster victims. Thus we plan to investigate ways to mitigate these issues through the use of attention mechanisms and will also compare our approach to a baseline that simply fuses the whole user histories as input to a fully connected network. For text this can be done simply by treating the whole user history as one document, while for images we can take the mean of the feature vectors produced by ResNet on each image in a user history. This comparison will allow us to examine if the sequence as well as grouping of posts in a history provides important information for predicting user informativeness.

For fusing text and image modalities, our initial plan will be to follow [10] in using LightGBM, and time permitting we will also experiment with an SVM as indicated by other work. In [10] LightGBM [8] and XGBoost [5] are tested against a baseline logistic-regression model to make informativeness predictions on the CrisisMMD dataset and they find that LightGBM outperforms the other methods. Reimplementing this approach will not only be a proven approach on which to test our new user history based models, but also will allow us to replicate the previous work's post-level approach as a baseline to compare against. That is, we could predict if each post in a user history is informative and use that count to directly derive an informativeness score. If time remains, we would also like to implement an SVM approach as it is successfully used in [13] to fuse text, video, and audio modalities in another disaster use-case.

Note that we have decided on this architecture with separate stages of training over an end-to-end differentiable model in part to allow for incremental development. If problems arise in one stage of the project, this will allow us to trouble shoot independently and modify our plans more easily if changes to parts of the architecture are required. This will also break up training times, allowing for faster iteration if we need to train a model again after some changes.

3.3 Evaluation Strategy and Expected Outcomes

By incorporating user histories and multi-modal information into our framework, we expect to find greater accuracy in regressing ground truth user informativeness scores derived from counts of informative posts (see 3.1 Dataset). Having formulated our problem as a regression we will examine R^2 and possibly also median absolute error to investigate the effect of outliers. Additionally to examine the impact of user-history length on the success of our predictions we will examine how prediction error changes between users with few posts to users with more posts.

Using the proposed informativeness score, we will compare the performance of our model against several baselines. First, we will compare our full multimodal model against the performance of its single modality components. We also plan to implement a baseline feedforward approach to user-level feature extraction in each modality with which we can compare test performance for the modalities separately and together. Time permitting, we also hope to replicate the post-level informativeness classifier of [10] which we can compare against our model by using the post-level predictions across a test set as counts for an informativeness score. Finally, we hope to demonstrate the generalizability of our models on test data consisting of an entirely different natural disaster type (e.g., earthquake as opposed to hurricane) unseen in the training data.

4 ROLES

Every group member will play some part in all aspects of the project. We plan to remain in communication and collaboration during each step because we all wish to learn from each part of building the project. That being said, our tentative division of work is: Magnus will work with the data wrangling, focusing on cleaning, updating (if needed), and presenting the data in the format needed. Chloe has experience working in R and with integrating multi-modal datasets, so she will take the lead on said parts. Ian will be responsible for training the model, which includes the hardware aspects of running the training session as well as observing that everything works as intended. All additional work will be shared, including further research, designing algorithms, and writing the final report.

One component of analysis, extracting features from the text of each tweet, will be completed as a separate project for the course CS6120: Natural Language Processing (NLP), taught by Dr. Lu Wang (luwang@ccs.neu.edu). Ian, Chloe, and another graduate student in that class, Jiahui Zhang, will complete this NLP project. We have confirmed with Dr. Wang and Dr. Radivojac that this collaboration is permissible given the size of the project and the separability of these different parts. The result of that assignment will be a regression model that scores user informativeness based on user history of text only tweet data. For the present project, this NLP model will be used to generate user-level feature vectors (from the second to last layer of the network) which will be used as input to our multimodal fusion along with the image-based model we are designing for this assignment.

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] Siham Bacha, Mohand Said Allili, and Nadjia Benblidia. 2016. Event recognition in photo albums using probabilistic graphical models and feature relevance. *Journal of Visual Communication and Image Representation* 40 (2016), 546–558.

[5] 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

[6] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] 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

[8] 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

[9] 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

[10] 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

[11] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 569–576.

[12] 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

[13] Samira Pouyanfar, Tianyi Wang, and Shu-Ching Chen. 2019. Residual Attention-Based Fusion for Video Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[14] 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

[15] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision*. Springer, 71–88.

[16] 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

[17] 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

[18] 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

[19] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *European Conference on Computer Vision (ECCV)*.

[20] 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

[21] 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