# Classifying Relevant Social Media Posts During Disasters Using Ensemble of Domain-agnostic and Domain-specific Word Embeddings

**Ganesh Nalluru, Rahul Pandey, Hemant Purohit**
Volgenau School of Engineering, George Mason University
Fairfax, VA, 22030
{gn, rpandey4, hpurohit}@gmu.edu

## Abstract

The use of social media as a means of communication has significantly increased over recent years. There is a plethora of information flow over the different topics of discussion, which is widespread across different domains. The ease of information sharing has increased noisy data being induced along with the relevant data stream. Finding such relevant data is important, especially when we are dealing with a time-critical domain like disasters. It is also more important to filter the relevant data in a real-time setting to timely process and leverage the information for decision support.

However, the short text and sometimes ungrammatical nature of social media data challenge the extraction of contextual information cues, which could help differentiate relevant vs. non-relevant information. This paper presents a novel method to classify relevant social media posts during disaster events by ensembling the features of both domain-specific word embeddings as well as more generic domain-agnostic word embeddings. Therefore, we develop and evaluate a hybrid feature engineering framework for integrating diverse semantic representations using a combination of word embeddings to efficiently classify a relevant social media post. The application of the proposed classification framework could help in filtering public posts at large scale, given the growing usage of social media posts in recent years.

## Introduction

Social media has become one of the most effective ways of communication and information sharing during time-critical events like disasters. Due to its ease of availability, there is a vast amount of non-relevant posts such as jokes (Imran et al. 2015) shared in the name of the disaster that makes the process of finding relevant information (Purohit et al. 2013) a challenging task.

The current practice across various public services for information filtering is to primarily leverage human resources through keyword-based filtering, and for emergency domains, it is done by personnel of emergency management agencies like the Public Information Officers (PIOs). Their main task is to monitor the serviceable requests for help (Purohit et al. 2018) or any actionable information coming from the public to provide relevant intelligence inputs to the decision-makers (Hughes and Palen 2012). However, due to the burstiness of the incoming stream of social media data during the time of emergencies, it is really hard to filter relevant information given the limited number of emergency service personnel (Castillo 2016). Therefore, there is a need to automatically filter out relevant posts from the pile of noisy data coming from the unconventional information channel of social media in a real-time setting.

Our contribution: We provide a generalizable classification framework to classify relevant social media posts for emergency services. We introduce the framework in the Method section, including the description of domain-agnostic and domain-specific embeddings, learning models, and baselines. We then demonstrate the preliminary evaluation in the Experiments and Results section, by experimenting with the real-world Twitter data collected during three disaster events. Lastly, we conclude with lessons and future directions in the Conclusions section.

## Related Work

Prior research on using social media for disaster management focuses on issues such as high volume, variety, and velocity of the incoming streams of data (Castillo 2016). Researchers have also tried to combine different modalities to filter information from the social data (Nalluru, Pandey, and Purohit 2019; Alam, Ofli, and Imran 2018). The prior research on crisis informatics (Palen and Anderson 2016) has looked into a different perspective of computing, information, and social sciences when investigating the use of social media.

Understanding the text of social media has been challenging for various classification schemes. Researchers have utilized different rule-based features as well as more social media-centric features for information filtering (Agichtein et al. 2008; Mostafa 2013; Sriram et al. 2010). Moreover, there has been a rise in using word embeddings that captures more semantic relationships among texts for efficient short-text classification (Tang et al. 2014; Yang, Macdonald, and Ounis 2018).

Word embeddings can be categorized into two types. The first category is the domain-agnostic word embeddings that are trained on a much larger domain-independent text corpus,

e.g., word2vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and FastText (Bojanowski et al. 2017). These embeddings contain broader semantic information among the words and can capture better dependencies. However, they lack the information of Out-of-Vocabulary (OOV) words and their dependencies among other words. The other category of word embeddings is the domain-specific word embeddings, which are trained on domain-specific text corpora, e.g., CrisisNLP resource (Imran, Mitra, and Castillo 2016) for disaster domain. This word embedding representation captures more rigorous domain-specific dependencies and information of the words. However, since the data corpus on which they are trained on was small, they sometimes lack the generic information of the words.

Based on the above studies, we propose a generic classification framework for relevant information that exploits both domain-agnostic and domain-specific word embeddings for efficient classification of social media posts through a novel ensembling approach.

## Method

In this section, we describe in detail our proposed approach. We start with the different text features we have utilized. Next, we give a detailed description of the different learning models used. Finally, we explain our classification framework used for ensembling the different text features for efficient learning.

### Extracting Text Features

In this section, we describe different methods used to derive the text features. We have used three pre-trained embeddings for this, where two embeddings (FastText and GloVe) are trained on domain-agnostic general data in the English language while one embedding (Crisis) is trained on disaster domain-specific data.

**GloVe Embedding:** GloVe or Global Vector for word Representation (Pennington, Socher, and Manning 2014) learns the word embedding representation from word-word co-occurrence matrix. We have used the GloVe trained on 2 billion tweets with 27 billion tokens of 1.2 million vocabularies.

**FastText Embedding:** FastText (Bojanowski et al. 2017) uses Continuous-Bag-of-Words model with a position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives. FastText also represents a text by a low dimensional vector, which is obtained by summing vectors corresponding to the words appearing in the text. We have used FastText for the English language from word vectors for 157 languages[1].

**Crisis Embedding:** In crisis embedding (Imran, Mitra, and Castillo 2016), authors trained a Continuous-Bag-of-Words model with pre-trained word-embedding initialized at the beginning. They trained the model on a large crisis dataset from their AIDR platform (Imran et al. 2014).

[1]https://fasttext.cc/docs/en/crawl-vectors.html

### Model Building

Next, we discuss the deep learning sequence models used and advantages of using our proposed model.

Sequence models take outputs of previous layers into account and predictions are made based on the conditional probability of the next word given all the previous words.

RNN is commonly used sequence model where information from previous time points are used as input for the next point, and when predicting label $y$, the network uses information not only from the corresponding input $x$ but also from the previous layers. The drawbacks of using RNN is that it suffers from vanishing gradients problem, which cannot capture long-term dependencies, as it is difficult to update the weights of previous layers during backpropagation.

So, to tackle the vanishing gradient problem that a normal RNN model suffers, LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) models are used. These algorithms solve the vanishing gradient problem by implementing gates within their network, enabling them to capture much longer range dependencies and regulate the flow of information in a sequence.

**BI-LSTM and GRU Models:** LSTM has three gates that help to control the information flow in a cell, namely Forget gate, Input gate, and Output gate.

*Forget Gate:* Forget gate helps us to decide if information should be forgotten or kept. After getting the output of previous state, the current input is passed through the sigmoid function, which helps to squeeze values between 0 and 1, and based on this, the forget gate decides if information should be kept or forgotten.

*Input Gate:* Input gate helps us to update the cell state. The previous hidden state and current input is passed into a sigmoid function that helps us decide which word is important and then fed into a tanh function.

*Output gate:* The output gate decides the next hidden state by applying the sigmoid function and then multiplies the input with tanh to squeeze the values between -1 and 1.Finally,relevant information is sent by mutiplying it with sigmoid function.

LSTM does not take into account the words that come after the current word in the sequence and rather uses only the information from previous words in the sequence. Meanwhile, Bidirectional LSTM takes information from both left-to-right and right-to-left directions, henceforth covering both information for the previous words and future words into consideration at a particular timestep. For this reason, we have used Bi-LSTM model to get more contextual information.The GRU is a simpler version of LSTM since it does not have the cell state and only has two gates, a reset gate, and an update gate.

For training the above-selected model, first, we have used features from each embedding as baselines. Moreover, we have also utilized two different formats of combining the embeddings for better knowledge representation. For that, we have used two implementations to combine the above mentioned pre-trained word embeddings: Meta Embedding and our proposed ensemble of models with different embeddings.

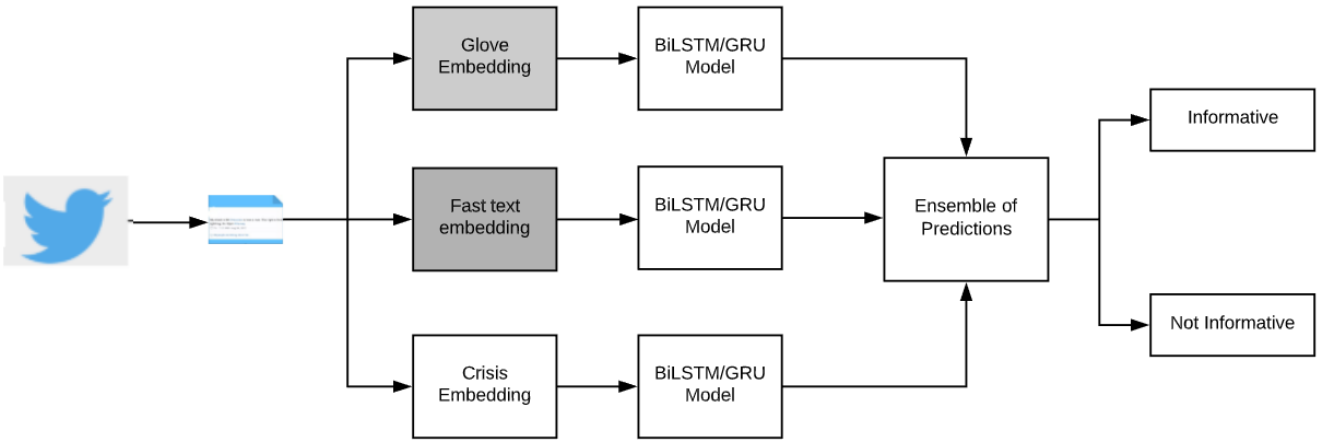


Figure 1: Proposed classification framework that exploits both ensemble of word embeddings trained on LSTM/GRU model.

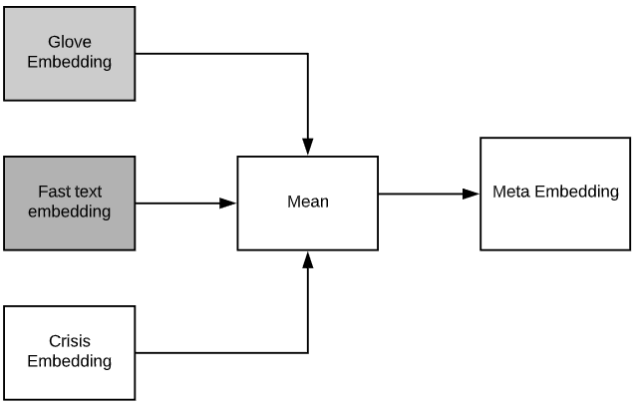


Figure 2: Meta Embedding. *(Average of all embeddings)*

**Meta Embedding:** Inspired from the meta embedding research from (Coates and Bollegala 2018), we take the average of all input embedding as the final representation for each word in the input tweet messages followed by training a Bi-LSTM or GRU model.

**Ensemble of model predictions from different embeddings (proposed):** In this method, we take the ensemble of all predictions from different models, each trained on different word embeddings, respectively. Each model is trained independently. We hypothesize that each feature set should learn on their own. Hence, for test data, if one classifier is uncertain about its predictions, the other classifiers act as a catalyst to get the correct final prediction. Each model was given different weights based on how the model performed on different word embeddings. For example,based on experiments for Hurricane Maria data with GRU model, predictions from fasttext and Glove based GRU models were given weights of 0.4 respectively, whereas crisis embedding based GRU model was given 0.2. Hence, to calculate final predictions $P_{final}$, we will use the equation 1 given below.

$$P_{final} = (0.4 \times P_{glove}) + (0.2 \times P_{crisis}) + (0.4 \times P_{fasttext}) \quad (1)$$

## Experiments and Results

For the validation of our hypothesis, we have done an extensive evaluation of our proposed approach against multiple baselines. We will explain our experimental setup and evaluation in this section.

### Dataset

We used the labeled dataset for three major disaster events of 2017 – Hurricane Maria, Hurricane Harvey, and Hurricane Irma provided by (Alam, Ofli, and Imran 2018). For our experimentation, our goal is to classify the informative tweets (relevant) vs non-informative tweets (non-relevant) using the proposed method of ensembling domain-agnostic and domain-specific embedding representations and compare its performance when only using single embedding representation. The size of the three datasets is as follows: Hurricane Maria has 2844 informative tweets and 1718 non-informative tweets, Hurricane Harvey has 3334 and 1109 tweets, and Hurricane Irma has 3564 and 957 tweets respectively. We conduct standard text pre-processing such as removing URLs, special characters, symbols like 'RT' or '@USER' tag, and stop-words as well as lowercasing the words.

### Classification Schemes

We employ various classification schemes for training with different features and learning algorithms as follows (*Lx* denotes learning model type, *Tx* text feature type, *M()* meta embeddings, and *E()* ensemble of the embeddings):

- **[*T1+L1*] Text embeddings (GloVe) + Bi-LSTM Model** This method uses GloVe embeddings as an input embedding to train a Bi-LSTM model. After the embedding layer,

Table 1: Comparison of accuracy for various training schemes. The results from the proposed ensemble of embeddings (scheme with *) outperform other schemes across all disaster events.

| Dataset | Input Features | [*M1*] Bi-LSTM | [*M2*] GRU |
|---|---|---|---|
| Hurricane Maria | [*T1*] GloVe | 78.64% | 76.67% |
| | [*T2*] FastText | **79.29%** | 75.57% |
| | [*T3*] Crisis | 76.34% | 76.01% |
| | [*M(T1+T2+T3)*] meta(GloVe + FastText + Crisis) | 78.09% | 76.01% |
| | [**E(T1+T2+T3)*] ensemble(GloVe + FastText + Crisis) | 79.18% | **77.21%** |
| Hurricane Harvey | [*T1*] GloVe | 85.93% | 84.13% |
| | [*T2*] FastText | 85.37% | 84.47% |
| | [*T3*] Crisis | 84.7% | 83.91% |
| | [*M(T1+T2+T3)*] meta(GloVe + FastText + Crisis) | 84.58% | 82.67% |
| | [**E(T1+T2+T3)*] ensemble(GloVe + FastText + Crisis) | **86.61%** | **84.58%** |
| Hurricane Irma | [*T1*] GloVe | 83.64% | 82.43% |
| | [*T2*] FastText | 83.31% | 82.65% |
| | [*T3*] Crisis | 82.76% | 82.32% |
| | [*M(T1+T2+T3)*] meta(GloVe + FastText + Crisis) | 81.87% | 82.09% |
| | [**E(T1+T2+T3)*] ensemble(GloVe + FastText + Crisis) | **83.75%** | **83.31%** |

we have a dropout of 30%. After that, we add a Bidirectional LSTM layer with 300 dimension output and 30% dropout for both input space and recurrent state. Then, we have two fully-connected layers of 1024 units with 80% dropout for generalization and a sigmoid layer of size 2 for prediction.Finally,model is fit with 100 epochs and early stopping.

- **[*T1+L2*] Text embeddings (GloVe) + GRU Model** This method uses same features of T1 and train on GRU model.
- **[*T2+L1*] Text Embeddings (FastText) + Bi-LSTM Model** This method uses FastText embeddings to generate text features (T2) and train that on Bi-LSTM.
- **[*T2+L2*] Text Embeddings (FastText) + GRU Model** This method uses same features of T2 and train that on GRU model.
- **[*T3+L1*] Text Embeddings (Crisis) + Bi-LSTM Model** This method uses the Crisis Embedding to generate text features (T3) and train that on Bi-LSTM.
- **[*T3+L2*] Text Embeddings (Crisis) + GRU Model** This method uses same features of T3 and train on GRU model.
- **[*M(T1+T2+T3) + L1*] Text Embeddings (Meta-embeddings of GloVe, FastText, and Crisis) + Bi-LSTM Model** This method uses the meta embedding of all three embedding as text features and train that on Bi-LSTM model.
- **[*M(T1+T2+T3) + L2*] Text Embeddings (Meta-embeddings of GloVe, FastText, and Crisis) + GRU Model** This method uses same features of T2, and train that on GRU model.
- *(Proposed)* **[*E*({*T1+T2+T3*} + *L1)*] Ensemble of all 3 classifiers of (GloVe + Bi-LSTM), (FastText + Bi-LSTM), and (Crisis + Bi-LSTM)** This method ensembles all three embeddings types and train a Bi-LSTM model.
- *(Proposed)* **[*E*({*T1+T2+T3*} + *L2)*] Ensemble of all 3 classifiers of (GloVe + GRU), (FastText + GRU), and (Crisis + GRU)** This method ensembles all three embedding types and train a GRU model.

## Results

For evaluation, we have reported an accuracy score on the 20% test data based on the 80-20% split strategy. Table 1 shows the results of our proposed method with the combinations of an ensemble of all embeddings and their comparison to the other baselines.

As seen in the table, our proposed approach have given a better score than the baselines in all the datasets. Our final proposed approach uses an ensemble of both domain-specific (crisis) embedding and more generalized (glove, fast-text) embedding; each being trained with Bidirectional LSTM and GRU network model. Hence, it is better able to capture the domain-specific as well as generic dependencies among the textual words for each class of data, i.e. *informative* vs *not-informative*. We can also observe that training a different model for each embedding is better than combining all embeddings first and then training a single model (meta embedding). The reason for that can be the difference in the feature space in which each embedding has been trained on. Hence, averaging the features may lose some information or add noise what one dimension is representing in each feature embedding. We can also see that all experiments with Bi-LSTM model are slightly better than their corresponding GRU model.

Furthermore, as we can infer from the results table, accuracy of models with single pre-trained embedding are almost the same, and a model's accuracy improves when we take the ensemble of different model predictions.

Also, the model with only domain-specific crisis embeddings perform worse on most of the cases when compared to Glove and FastText embeddings; but, it gives a boost to the other embeddings when combined. Thus, this finding suggests the usefulness of combining both domain-specific

and domain-agnostic representations to provide more context to the learning algorithm.

## Conclusion and Future Work

This paper presented a novel approach to classify relevant social media posts by ensembling domain-agnostic and domain-specific word embedding representations, which can capture different types of information in a given text and provide more context to the learning model. We have validated our approach with three disaster event datasets and achieved over 86% accuracy score. We have also observed better performance on Bi-LSTM based model compared to GRU based model for our use case. In light of this preliminary investigation, the application of the proposed method has the potential to help improve social media services at emergency response organizations.

In future work, we can generalize our ensemble approach more according to the training data. More specifically, since we have fixed the ratio of the prediction score coming from different embedding-based learning at the beginning in our proposed approach, we can also learn this ratio as a part of the training. We can also add more embeddings for enhancing the contextual representation of more information in the text.

## Acknowledgment

Authors thank US National Science Foundation (NSF) grants IIS-1657379 and IIS-1815459 for partial support.

## References

[Agichtein et al. 2008] Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proc. of the 2008 international conference on web search and data mining*, 183–194. ACM.

[Alam, Ofli, and Imran 2018] Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*.

[Bojanowski et al. 2017] Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

[Castillo 2016] Castillo, C. 2016. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

[Coates and Bollegala 2018] Coates, J., and Bollegala, D. 2018. Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 194–198. New Orleans, Louisiana: Association for Computational Linguistics.

[Hughes and Palen 2012] Hughes, A. L., and Palen, L. 2012. The evolving role of the public information officer: An examination of social media in emergency management. *Journal of Homeland Security and Emergency Management* 9(1).

[Imran et al. 2014] Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Vieweg, S. 2014. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, 159–162. ACM.

[Imran et al. 2015] Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47(4):67.

[Imran, Mitra, and Castillo 2016] Imran, M.; Mitra, P.; and Castillo, C. 2016. Twitter as a lifeline: human-annotated Twitter corpora for NLP of crisis-related messages. In *Proc. LREC*.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, 3111–3119.

[Mostafa 2013] Mostafa, M. M. 2013. More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications* 40(10):4241–4251.

[Nalluru, Pandey, and Purohit 2019] Nalluru, G.; Pandey, R.; and Purohit, H. 2019. Relevancy classification of multimodal social media streams for emergency services. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 121–125. IEEE.

[Palen and Anderson 2016] Palen, L., and Anderson, K. M. 2016. Crisis informatics – new data for extraordinary times. *Science* 353(6296):224–225.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

[Purohit et al. 2013] Purohit, H.; Castillo, C.; Diaz, F.; Sheth, A.; and Meier, P. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19(1).

[Purohit et al. 2018] Purohit, H.; Castillo, C.; Imran, M.; and Pandey, R. 2018. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In *2018 IEEE/ACM ASONAM*, 119–126. IEEE.

[Sriram et al. 2010] Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; and Demirbas, M. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 841–842. ACM.

[Tang et al. 2014] Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1555–1565.

[Yang, Macdonald, and Ounis 2018] Yang, X.; Macdonald, C.; and Ounis, I. 2018. Using word embeddings in twitter election classification. *Information Retrieval Journal* 21(2-3):183–207.