Deep Neural Networks versus Naïve Bayes Classifiers for Identifying Informative Tweets during Disasters

Venkata Kishore Neppalli
University of North Texas
kishoreneppalli@gmail.com

Cornelia Caragea
Kansas State University
ccaragea@ksu.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

ABSTRACT

Traditional machine learning techniques have shown promising results in automating the process of identifying useful information in crisis-related data posted through micro-blogging services such as Twitter. More recently, deep learning techniques have also shown promise in the area of disaster response. In this paper, we focus on understanding the effectiveness of deep neural networks by comparison with the effectiveness of standard classifiers that use carefully engineered features. Specifically, we design various feature sets (based on tweet content, user details and polarity clues) and use these feature sets individually or in various combinations, with Naïve Bayes classifiers. Furthermore, we develop neural models based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) with handcrafted architectures. We compare the two types of approaches in the context of identifying informative tweets posted during disasters, and show that the deep neural networks, in particular the CNN networks, are more effective for the task considered.

Keywords

Machine Learning, Deep Learning, Naïve Bayes, Feature Engineering, Identifying Informative Tweets.

INTRODUCTION

In recent years, micro-blogging services such as Twitter have emerged as effective tools for broadcasting messages worldwide during disaster events. For example, over 20 million tweets were posted in less than a week during Hurricane Sandy, whereas five thousand tweets were posted every second during the earthquake and subsequent tsunami in Japan in 2011, resulting in 1.5 million tweets every 5 minutes. A quarter of Americans looked for information about the Boston bombings, and followed the updates regarding the hunt for bombers on social media such as Facebook and Twitter (Bullock et al. 2012). With millions of messages posted through micro-blogging services during disaster events, it has become imperative to identify valuable information that can help the emergency responders to develop efficient relief efforts and aid victims (Watson et al. 2017). Manually sifting through voluminous streaming data to filter useful information in real time is inherently impossible (Meier 2013).

Traditional machine learning techniques have shown promising results in automating the process of identifying useful, relevant and trustworthy information in big crisis data (Qadir et al. 2016), However, these techniques usually require carefully engineered features to represent the data (e.g., tweets), and produce accurate classifiers (Imran, Castillo, et al. 2015). Such features include standard bag-of-word features, part-of-speech tags, or other more sophisticated content-based features. Generally, the more comprehensive the set of features used and the more it captures the domain knowledge, the more accurate the classifier that it produces. Thus, accurate classifiers are produced at the cost of constructing (often expensive) features.

No.	Tweet	Class
1.	*"Boston police: At least 3 people killed in Boston marathon bombing http://t.co/duv2iiwmrz #bostonmarathon"*	Informative
2.	*"RT abcnews: Breaking: Flood maps for Brisbane River are now available http://t.co/2ExK39rY #bigwet"*	Informative
3.	*"#walangpasok because of a holiday in the philippines. hoping that the weather is much better. #floodph #maringph"*	Non-informative
4.	*"@user really sad about the tragedy:(,may they rest in peace #santamaria"*	Non-informative

Table 1. Examples of tweets from disaster events with class labels.

More recently, researchers have started to investigate the use of deep learning in the area of disaster response (Nguyen et al. 2017; C. Caragea, Silvescu, et al. 2016; Ben Lazreg et al. 2016). Deep learning approaches (LeCun et al. 2015) have significantly improved the performance on many tasks, such as speech and visual object recognition (LeCun et al. 2015), and tweet classification (Yuan et al. 2016), among others. They can make use of pretrained distributed representations (Mikolov et al. 2013), and have the ability to identify higher-level hierarchical data representations, eliminating the need for carefully engineered features (LeCun et al. 2015), at the cost of engineering the network architecture and learning the network weights based on larger amounts of labeled data than are generally needed by traditional machine learning algorithms.

However, in disaster events, which unfold rapidly, the amounts of *labeled data* are limited. Hence, one research question that can be raised is: *Given a limited amount of labeled crisis-related data, how does the effectiveness of deep learning models that use pretrained distributed representations, compare with that of standard machine learning classifiers that use carefully designed feature representations?* In particular, we compare Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) trained on distributed representations with Naïve Bayes classifiers trained on handcrafted features. These features include standard bag-of-word features, content-based features (e.g., URLs, hashtags, emoticons, slang), user-based features (e.g., number of friends and number of followers) and polarity-based features (e.g., positive and negative words).

We address the research question above by training classifiers for identifying informative tweets during disasters using the CrisisLexT26 dataset constructed by Olteanu et al. (2014). This dataset comprises of tweets from 26 disasters that occurred during 2012 and 2013. "Informative" tweets are defined as tweets that provide valuable information to anyone in the scene of a disaster (as a victim, supporter, responder, etc.), and non-informative tweets are defined as tweets which do not convey any useful content in the scene of a disaster. Examples of informative and non-informative tweets are shown in Table 1.

In summary, our contributions are:

- We engineer a set of features that combines "bag-of-word" features with features extracted from tweet content, user details, and polarity clues, and use those features with Naïve Bayes, as an effective representative of standard machine learning classifiers for text data. We show experimentally that models trained using the combination of all features perform better than models trained on each feature type independently (i.e., either "bag-of-word" features or features extracted from the tweet content, etc.).
- We engineer and train deep neural network models, specifically CNN and RNN models, with various architectures. We show that the deep neural network models are more effective in identifying informative tweets for both natural and non-natural disasters, as compared to the Naïve Bayes models.
- Furthermore, we show that the deep neural network models generalize better than Naïve Bayes across disasters of different nature (e.g., natural and non-natural disasters). We achieve this by comparing models trained on natural disasters and evaluated on non-natural disasters, and vice-versa.

Our study provides additional evidence to the claim that proper analysis of streaming data may lead to applications able to help not only the disaster response providers to allocate resources more efficiently, but also the victims who are in need and seeking support, by filtering out necessary informative messages for immediate availability. The problem of finding relevant information as a disaster evolves faces many challenges since extracting informative tweets and filtering out those that are non-informative has to be done in real time and with high accuracy. In particular, deep learning networks, such as CNN and RNN networks, can address this task more effectively than standard machine learning methods that require more extensive feature engineer.

The remaining of this paper is organized as follows: We first discuss the related work and challenges of using microblogging data in disaster events. Next, we describe the Twitter dataset in the DATASET section. We present the features used for identifying informative tweets from a stream of messages, and the deep neural networks used in the APPROACHES section. We discuss the results of our experiments in the RESULTS section. Finally, we conclude the paper and provide several directions for future work.

BACKGROUND AND RELATED WORK

Castillo et al. (2011) developed automatic machine learning techniques to assess the credibility of tweets related to specific topics or events (however, not restricted to emergency events). Using a comprehensive set of features extracted from user's posting behavior and tweet text and context (specifically, message-based, user-based, topic-based, and propagation-based), newsworthy events were classified as credible or not credible.

Subsets of the features introduced by Castillo et al. (2011), together with other specific features, have been used with traditional machine learning approaches for disaster tweet classification, as described in the following surveys (Imran, Castillo, et al. 2015; Beigi et al. 2016). For example, C. Caragea, Squicciarini, et al. (2014) used features such as unigrams, polarity clues, emoticons, punctuation and acronyms to identify the sentiment of tweets posted during Hurricane Sandy. Verma et al. (2011) used unigrams and bigrams, together with part-of-speech (POS) tags, subjectivity of tweet, tone of tweet (personal/impersonal), etc., to identify tweets that contain information useful for situational awareness. Rudra et al. (2015) used features such as count of personal pronouns, count of wh-words, fraction of subjective words, to classify tweets as situational or non-situational. Sen et al. (2015) have also focused on identifying situational awareness tweets using low level syntactic features, including POS tags, tweet formality, word subjectivity, emoticons and exclamations, personal pronouns, etc. Imran, Elbassuoni, et al. (2013) classified tweets according to several situational awareness categories, such as caution and advice, casualties and damage, donations, people missing, etc. using unigrams, bigrams, POS tags, presence of URL/hashtags, tweet length, etc. Yin et al. (2012) used similar features to identify tweets related to a disaster, identify disaster type and assess the impact of a disaster. Neppalli et al. (2017) used features such as presence of hashtags, number of hashtags, tweet length to predict tweet retweetability. Other works (C. Caragea, McNeese, et al. 2011; Ashktorab et al. 2014; Imran, Chawla, et al. 2016; Huang and Xiao 2015) have simply used the standard bag-of-word (BOW) features based on unigrams and/or bigrams, with accurate results.

More recently, research has started to investigate the use of deep learning (LeCun et al. 2015) in the area of disaster response (C. Caragea, Silvescu, et al. 2016; Nguyen et al. 2017; Ben Lazreg et al. 2016). For example, C. Caragea, Silvescu, et al. (2016) used convolutional neural networks (CNN) to identify informative messages in data from flooding disasters and reported significant improvements in performance over Support Vector Machines and fully connected Artificial Neural Networks. Similarly, Nguyen et al. (2017) used CNNs on situational awareness crisis data and noted improvements over traditional algorithms. Ben Lazreg et al. (2016) used a different deep learning approach, specifically a Long Short-Term Memory (LSTM) network, to learn a model from crisis tweets and used this model to generate snippets of information summarizing the tweets.

Inspired by the above-mentioned works, we compare several sets of features extracted from the tweet content, user details and polarity clues, to understand what features are the best-performing ones for the problem of identifying informative tweets posted during a crisis event. Moreover, we compare traditional machine learning algorithms, specifically Naïve Bayes classifiers, that use the most comprehensive set of features for identifying informative tweets, with deep learning networks that use pre-trained embeddings of the tweet words to understand what type of approach gives better results. Despite that the deep learning approaches often provide improved results over traditional models, they are not interpretable, whereas our handcrafted features could yield interpretable models. We focus on Naïve Bayes as a traditional machine learning algorithm, given that Naïve Bayes has no parameters that need to be tuned, and has shown good performance in the context of disaster tweet classification, when compared with other machine learning algorithms, such as Support Vector Machines (SVM) or Random Forests (RF) - see, for example, (Li, D. Caragea, C. Caragea, and Herndon 2018; Li, D. Caragea, and C. Caragea 2017). We evaluate the performance of the models trained on natural disasters (e.g., earthquakes and floods), as well as non-natural disasters (e.g., fire accidents and train crashes). Furthermore, we investigate what type of models, Naïve Bayes classifiers or deep neural networks, have the capability to generalize from one type of disaster to another (e.g., from natural to non-natural disasters).

DATASET

We used the CrisisLexT26 dataset constructed by (Olteanu et al. 2014). This collection contains tweets from 26 disasters that occurred during 2012 and 2013, which are manually annotated by crowd-sourced workers with the

No.	Non-natural Disasters	Natural Disasters
1.	Colorado Wildfires (2012)	Costa Rica Earthquake (2012)
2.	Australia Bushfires (2012)	Gautemala Earthquake (2012)
3.	Venezuela Refinery (2012)	Italy Earthquake (2012)
4.	Boston Marathon Bombings (2013)	Bohol Earthquake (2013)
5.	Brazil Night Club Fire (2013)	Typhoon Pablo (2012)
6.	Glasgow Helicopter Crash (2013)	Typhoon Yolanda (2013)
7.	LA Airport Shootings (2013)	Alberta Floods (2013)
8.	Lac Megantic Train Crash (2013)	Colorado Floods (2013)
9.	NY Train Crash (2013)	Philipinnes Floods (2012)
10.	Savar Building Collapse (2013)	Manila Floods (2013)
11	Spain Train Crash (2013)	Queensland Floods (2013)
12.	Singapore Haze (2013)	Sardinia Floods (2013)
13.	West Texas Explosion (2013)	Russia Meteor (2013)

Table 2. Summary of disasters used in the experiments.

following informativeness labels: "*related and informative*", "*related - but not informative*", "*not related*", and "*not applicable*." There are about 1000 tweets manually annotated in each of the 26 disasters. Among the 26 disasters, there are 13 non-natural disasters and 13 natural disasters. We classify as a natural disaster any catastrophic event caused by nature, or a natural process of the earth (e.g., cyclones, earthquakes, floods). We classify as a non-natural disaster any event caused by human actions, which may be intentional actions (e.g., gun shootings, bombings), or unintentional actions that lead to technological failures (e.g., industrial accidents, train crash). Table 2 shows the 26 disasters used in this study and their types. Using the language attribute available from Twitter, we found that some of the tweets in the dataset are in languages other than English. Furthermore, some tweets are labeled as *not-related* or *not applicable* to the disaster in question. We removed from the dataset the not-related and non-English tweets, and were left with 7200 tweets related to non-natural disasters and 6480 tweets related to natural disasters.

APPROACHES TO INFORMATIONAL TWEET CLASSIFICATION

Informational Tweet Classification as Supervised Learning

We formulate the problem of informational tweet classification as a supervised learning problem. Specifically, given an independent and identically distributed (iid) set of labeled tweets posted during various disasters, a hypothesis class (e.g., Naïve Bayes classifiers), and a performance criterion (e.g., accuracy), a supervised learning algorithm identifies a hypothesis (i.e., a classifier) that optimizes the given performance criterion. The identified classifier is used to classify tweets as informative or non-informative during the testing phase.

Feature Engineering for Traditional Learning Algorithms

Next, we describe the features that we use as input to the Naïve Bayes classifier. We divide them into four sets, namely bag-of-words (BOW), tweet content features (TC), user-based features (UF) and polarity-based (PF) features.

Bag-of-Words (BOW)

A vocabulary is first constructed, which contains all unique words from the collection of training tweets. Using this vocabulary, we use a binary representation to represent tweets as vectors. Specifically, we assign 1 for each word in the vocabulary if the word is found in the tweet, otherwise we assign 0.

Tweet Content Features (TC)

We designed these features based on the content of a tweet, as follows:

- Presence of URLs: True if at least one URL is present in the tweet, false otherwise.
- Presence of hashtags: True if at least one "#hashtag' is present in the tweet, false otherwise.
- Hashtag count: Number of occurrences of hashtags in the tweet.
- Presence of emoticons: True if at least one emoticon is present in the tweet, false otherwise. Intuitively, this feature is indicative of conversational tweets.

- Instructional keywords: True if the tweet contains instructional words, such as "text," "call," and "donate," otherwise false.
- Phone numbers: True if the tweet contains a phone number, otherwise false. Phone numbers are identified using regular expressions (e.g., `\\d{3}-\\d{3}-\\d{4}`identifies numbers of the form xxx-xxx-xxxx).
- Internet slang:[1] True if the tweet contains abbreviations or slang, such as "OMG" (Oh My God), otherwise false. Abbreviations/slang are indicative of informalities, and occur mostly in conversational tweets.
- Retweet (RT): True is the tweet contains retweet patterns, such as "RT @" or "RT@," otherwise false.
- Profanity - True is the tweet contains informal/cuss words, otherwise false.
- Sentence structure: We use OpenNLP[2] Java Libraries to check whether the tweet contains a one-word sentence. We assign a value of 1 for the presence of one-word sentences, otherwise 0. Likewise, we also check for the presence of multiple sentences, and assign value 1 for multiple sentences, and otherwise 0.

User-based Features (UF)

The following attributes, provided by Twitter, are used as the set of user-based features.

- Followers count: Total number of followers of a user.
- Friends count: Total number of friends of a user.
- Favorites count: Total number of favorite tweets in the user account.
- Listed communities count: Total number of communities that the user is listed in.
- Statuses count: Total number of tweets posted by the user.
- Verified account: Twitter follows a procedure to verify the authenticity of the users who are famous personalities, brands, etc. Each user account is assigned with a special emblem, if the user meets the verification criteria. The possible value for this feature is 1 (if verified account) and 0 (if not verified).

Polarity-based Features (PF)

These features are formulated based on the polarity (positive or negative) of the words in a tweet. We identify the polarity words using lexicons of positive and negative words created by Hu and Liu (2004), and extract the following features:

- Positive word count: Number of positive words in a tweet.
- Negative word count: Number of negative words in a tweet.
- Positive Score: Positive score returned by the SentiStrength algorithm.
- Negative Score: Negative score returned by the SentiStrength algorithm.
- Emotional Divergence: We adopt the definition of "emotional divergence" (ED) from Pfitzner et al. (2012). ED is defined as "the (normalized) absolute difference between the positive and the negative sentiment score delivered by SentiStrength." It is calculated as $ED = \frac{(p-n)}{10}$, where p is a positive score and n is a negative score output by the SentiStrength algorithm. This feature is an aggregate of the above four polarity features.

Models

In this work, we evaluate the performance of several models for identifying informative tweets posted during disaster events. First, we evaluate the performance of models based on bag-of-words (BOW) and custom features (namely, TC, UF & PF), and use them as our baselines. Then, we evaluate the performance of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which take as input pre-trained embeddings.

[1]https://www.lifewire.com/urban-internet-slang-dictionary-3486341
[2]https://opennlp.apache.org/

Naïve Bayes Models

In the category of traditional supervised machine learning algorithms, we use the Naïve Bayes algorithm, as Naïve Bayes has been shown to be an effective classifier in crisis-related tweet classification (Li, D. Caragea, C. Caragea, and Herndon 2018). Furthermore, our own preliminary examination with Support Vector Machine and Random Forest classifiers confirmed that the Naïve Bayes classifier gives results competitive and sometimes better than the results given by other classifiers (results not shown due to space constraints). We used the following features and feature combinations with Naïve Bayes: individual feature types (namely, Only TC, Only UF, and Only PF), BOW (0/1 representation of tweets), and the combination of BOW with the other feature sets.

Deep Neural Network Models

CNN Model: Convolutional Neural Networks can be seen as a special category of Deep Neural Networks, based on the concepts of *local receptive field* and *weight replication*. CNNs consist of combinations of convolution and sub-sampling layers, which help extract meaningful representations of the input data. We use a model similar to the one described in (Kim 2014). Our model has a single convolution layer followed by max-over-time pooling layer and a fully connected layer. The input to convolution layer is a matrix, where each row is a vector representation of a word in the tweet, specifically an embedding trained using crisis-related tweets. Multiple filter of different sizes are used to convolve this matrix and detect local features. The output of each filter is fed to a pooling layer that scans for the maximum value. The output feature maps corresponding to the filters used are finally concatenated to get the feature representation of input. For regularization, we implement the dropout technique to the fully connected layer. Finally, a softmax layer calculates the probability distribution over the two classes, and also the loss function.

RNN Model: Recurrent Neural Networks represent another category of Deep Neural Networks. RNNs have a recurrent hidden state whose activation is controlled by the output from the current step as well as previous time steps. They keep track of contextual information from each time step in the past and are suitable for modeling sequences such as sentences, which are sequences of words. We create a Recurrent Neural Network Model with Gated Recurrent Unit (Cho et al. 2014). GRU has gating units to modulate the flow of information. The inputs to our RNN are word embeddings as for the CNN model.

EXPERIMENTS AND RESULTS

We used tweets from 26 disasters available from CrisisLexT26 (Olteanu et al. 2014), and separated them in two subsets based on the disaster type: one consisting of 7200 tweets from non-natural disasters and another one consisting of 6480 tweets from natural disasters. To have an equal amount of data for training in each subset and remove the sample size bias, we randomly sampled a subset of 6400 tweets from each subset. Each sample consists of 5000 informative tweets and 1400 non-informative tweets. Table 3 shows the results of the comparison between the Naïve Bayes models obtained using different feature sets and the deep neural network models, specifically CNN and RNN models, that take as input word embeddings. We evaluate the models using the 10-fold cross-validation strategy. We report the F1-scores of the classifiers on the informative and non-informative classes, and also the average scores over the two classes.

Naïve Bayes Models

We trained several Naïve Bayes models using bag-of-words, custom features and their combinations. For the bag-of-words, we removed stop words and words occurring in less than three tweets. As can be seen from Table 3, among our custom feature sets (TC, UF, PF), the classifiers trained using only TC perform better than using only UF and only PF, on both disaster types. For example, the average F1-score of only TC for non-natural is 0.776, which is much better than 0.583 (the average F1-score for only UF) and 0.742 (the average F1-score for only PF). When user features (UF) are added to any feature set, the performance does not improve for any disaster type, which indicates that the informativeness of a tweet is independent of the user features, given the other features used. Overall, we observe the best performance for TC+PF among our feature set combinations. When polarity features are added to TC, we observe that the classifiers' performance improves, implying that these two feature sets are mutually assisting each other to boost the classifier's performance. For example, for the non-natural disaster type, the average F1-score increases from 0.776 (for only TC) to 0.789 (for TC+PF), which represents a 1.68% increase. When TC features are combined with BOW features, the classifier's performance is similar to that obtain with Only BOW. By adding UF to BOW+TC, the classifier's performance is degraded. However, by adding PF features to BOW+TC, the performance improves slightly by comparison to only BOW features. Overall, by comparing Naïve Bayes models trained with different features, we observe that UF features are not useful for identifying the informative tweets, while TC and PF are slightly improving the results obtained with BOW only on this classification task.

Model Type	Feature/Model	Non-natural			Natural		
		Info.	Non-Info.	Avg.	Info.	Non-info.	Avg.
Naïve Bayes	Only TC	0.841	0.555	0.778	0.835	0.435	0.748
	Only UF	0.656	0.323	0.583	0.195	0.369	0.233
	Only PF	0.867	0.297	0.742	0.855	0.341	0.743
	TC+UF	0.808	0.539	0.749	0.679	0.470	0.633
	TC+PF	0.870	0.500	0.789	0.865	0.457	0.776
	UF+PF	0.801	0.390	0.711	0.279	0.375	0.300
	TC+UF+PF	0.837	0.577	0.781	0.718	0.494	0.669
	Only BOW	0.903	0.704	0.859	0.900	0.681	0.852
	BOW+TC	0.902	0.703	0.858	0.901	0.688	0.855
	BOW+TC+UF	0.894	0.696	0.850	0.880	0.670	0.834
	BOW+TC+PF	**0.905**	**0.710**	**0.862**	**0.902**	**0.688**	**0.855**
DNN	**CNN**	**0.937**	**0.747**	**0.895**	**0.928**	**0.687**	**0.875**
	RNN	0.927	0.714	0.880	0.918	0.676	0.865

Table 3. The performance (F1-score) of the Naïve Bayes (NB) models with various feature sets, by comparison with the performance of the deep neural network (DNN) models, specifically CNN and RNN. Experiments are performed using 10-fold cross-validation on the sets of *Non-natural* and *Natural* disasters, respectively. Among the Naïve Bayes models, the model that uses BOW+TC+PF gives the best results (values highlighted in blue), while the best results overall are obtained with the CNN model (values highlighted in red).

Deep Neural Network Models

We train two neural network models: CNN and RNN. For the CNN model, we trained several architectures and tuned different sets of parameters on a separate validation set. The best overall parameters, which were used in the final CNN model, are described in what follows. We used a word2vec (Mikolov et al. 2013) model trained using crisis-related messages from a collection of disaster tweets (i.e., tweets crawled from Hurricane Sandy and Boston bombing disasters, which consist of $\approx$ 30 million tweets). The dimension of each word vector is 150. We used three convolution filter sizes: $3 * 150$, $4 * 150$ and $5 * 150$. The number of feature maps for each filter size is 256 and the dropout ratio is 0.5 in all cases. For the RNN model, we used a single GRU cell with 32 hidden units, and maximum sequence length of 20 words. The results of deep neural networks are shown in the last two rows in Table 3. As can be seen from the table, the CNN model outperforms the other models. For instance, on non-natural disasters, the CNN shows approximately 3% better F1-score as compared to the Naïve Bayes model trained using the combination of BOW+TC+PF features. Furthermore, the CNN model performs better than the RNN model by approximately 1.5%.

Cross-Domain Experiments

Next, we investigate how the Naïve Bayes and the DNN models will generalize from one type of disaster to another (e.g., from natural to non-natural disasters). We use data from one type of disaster for training and data from the other type of disaster for testing. Table 4 shows the results of these experiments. As can be seen from Table 4, the best cross-domain results are also obtained with the DNN models, and the CNN model performs better than the RNN model. Among the Naïve Bayes models (Only BOW, Only TC, BOW+TC, BOW+TC+PF) trained and tested in the cross-domain setting, we observe that classifiers trained using feature sets TC and PF, in conjunction with BOW features, are performing better than the classifiers trained on Only BOW, for example in Train-*Non-natural*/Test-*Natural* experiment, the average F1-score is 0.707 for Only BOW and is 0.725 for BOW+TC+PF, which represents a 2.55% increase.

We also compare the cross-domain results with the 10-fold cross-validation results (Table 3), which were obtained when training and testing on the same type of disaster (this is a fair comparison as in the cross-validation each example is used exactly once in testing). As can be seen from Tables 3 and 4, generally, the cross-domain models trained using data from natural disasters and tested on data from non-natural disasters show similar performance to that of the classifiers evaluated using 10-fold cross-validation on the non-natural disaster data. The same trend is observed for the DNN models, when training on data from non-natural disasters and testing on data from natural disasters. However, in this cross-domain case, the Naïve Bayes models have significantly worse performance than that of the models trained and tested using 10-fold cross-validation.

These results suggest that the features extracted from natural disasters can be used for developing Naïve Bayes models that could predict both disaster types (non-natural or natural), whereas features extracted from non-natural

Model Type	Features/Model	Train-*Natural*/Test-*Non-natural*			Train-*Non-natural*/Test-*Natural*		
		Info.	Non-Info.	Avg.	Info.	Non-Info.	Avg.
Naive Bayes	Only BOW	0.891	0.656	0.839	0.753	0.540	0.707
	Only TC	0.871	0.446	0.778	0.817	0.497	0.74
	BOW+TC	0.894	0.666	0.844	0.768	0.546	0.719
	BOW+TC+PF	0.897	0.669	0.847	0.773	0.555	0.725
DNN	**CNN**	**0.929**	**0.676**	**0.874**	**0.912**	**0.682**	**0.862**
	RNN	0.918	0.685	0.867	0.911	0.66	0.856

Table 4. Summary of results for cross-domain experiments with Train-on-all/Test-on-all strategy

disasters do not generalize well for natural disasters. We further investigated this aspect by analyzing several patterns in natural versus non-natural disasters. Our analysis revealed some variation in terms of hashtags usage: the usage seems to be greater in natural disaster tweets as compared to the usage in non-natural disaster tweets. Specifically, in our natural disaster subset, there are 204 unique hashtags with a total of 6174 occurrences, while in the non-natural disaster subset, there are 154 unique hashtags with a total of 3758 occurrences. Another observation we made was that the non-natural disaster tweets have a lot of slang words, whereas a more limited slang usage can be seen in natural disaster tweets. Therefore, the fact that the models trained on non-natural disasters do not generalize to natural disasters can be explained through variations in usage of hashtags and slang, but more analysis is needed to confirm this hypothesis, which will be interesting to explore in future.

CONCLUSIONS

Previous research suggests that data gleaned from social media contributions have both significant value to emergency responders and are difficult to use. Responders seek an enhanced operational picture during any disaster, which grants them better situational awareness. In this paper, we designed several feature sets for identifying informative tweets using CrisisLexT26 dataset. We experimented with BOW and various combinations of our designed feature sets - TC, UF, and PF to obtain the best performing model. Among our feature set combinations, TC+PF is performing better than other feature set combinations and adding BOW to TC+PF show improved classifier performance over TC+PF. We find that using user features is not useful for this classification task, that is, in general, the informativeness of a tweet is independent of user characteristics. From all the experiments, it is evident that CNN outperformed the other models. We also explored how the models developed for natural disasters are useful for non-natural disasters and vice-versa. For the traditional machine learning (non-neutral) models, we found that the natural disasters generalize well on both natual and non-natural disasters, but non-natural disasters do not. In contrast, neural models generalize well across both types of disasters.

In future, it would be interesting to explore a combination of DNNs with hand-crafted features. One more interesting direction would be to identify the tweets which convey actionable information and particularly those that ask for help, which will make it easier for the first responders to find those victims who are seeking help.

ACKNOWLEDGMENTS

We thank the National Science Foundation for support from the grants IIS-1802284, IIS-1741345, IIS-1526542 and CMMI-1541155. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the National Science Foundation. We also wish to thank our anonymous reviewers for their constructive comments.

REFERENCES

Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). “Tweedr: Mining twitter to inform disaster response”. In: *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management*. ISCRAM ’14. University Park, Pennsylvania.

Beigi, G., Hu, X., Maciejewski, R., and Liu, H. (2016). “An overview of sentiment analysis in social media and its applications in disaster relief”. In: *Sentiment Analysis and Ontology Engineering*. Springer, pp. 313–340.

Ben Lazreg, M., Goodwin, M., and Granmo, O.-C. (2016). “Information Abstraction from Crises Related Tweets Using Recurrent Neural Network”. In: *Proceedings of the Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016*. Ed. by L. Iliadis and I. Maglogiannis. Cham: Springer International Publishing, pp. 441–452.

Bullock, J., Haddow, G., and Coppola, D. P. (2012). *Homeland security: the essentials*. Butterworth-Heinemann.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., Wu, D., Tapia, A. H., Giles, C. L., Jansen, B. J., et al. (2011). "Classifying Text Messages for the Haiti Earthquake". In: *Proc. of the 8th International Conference on Information Systems for Crisis Response and Management*. ISCRAM '11. Lisbon, Portugal.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). "Identifying Informative Messages in Disasters using Convolutional Neural Networks". In: *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016.*

Caragea, C., Squicciarini, A. C., Stehle, S., Neppalli, K., and Tapia, A. H. (2014). "Mapping moods: Geo-mapped sentiment analysis during hurricane sandy". In: *11th Proceedings of the International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, May 18-21, 2014.*

Castillo, C., Mendoza, M., and Poblete, B. (2011). "Information Credibility on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW'11. Hyderabad, India: ACM, pp. 675–684.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259*.

Hu, M. and Liu, B. (2004). "Mining and Summarizing Customer Reviews". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: ACM, pp. 168–177.

Huang, Q. and Xiao, Y. (2015). "Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1549–1568.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Computing Surveys (CSUR)* 47.4, p. 67.

Imran, M., Chawla, S., and Castillo, C. (2016). "A Robust Framework for Classifying Evolving Document Streams in an Expert-Machine-Crowd Setting". In: *Proceedings of the 18th International Conference on Data Mining (ICDM)*. Barcelona, Spain.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Practical extraction of disaster-relevant information from social media". In: *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pp. 1021–1024.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

Li, H., Caragea, D., and Caragea, C. (2017). "Towards Practical Usage of a Domain Adaptation Algorithm in the Early Hours of a Disaster". In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2017)*. France.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018). "Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach". In: *Journal of Contingencies and Crisis Management (JCCM), Special Issue on HCI in Critical Systems. In press* 26.1, pp. 16–27.

Meier, P. (2013). "Crisis Maps: Harnessing the Power of Big Data to Deliver Humanitarian Assistance". In: *Forbes Magazine*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Neppalli, K., Caragea, C., Caragea, D., Medeiros, M. C., Tapia, A., and Halse, S. (2017). "Predicting Tweet Retweetability during Hurricane Disasters." In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM 2017)*.

Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2017). "Robust Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks". In:

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises." In:

Pfitzner, R., Garas, A., and Schweitzer, F. (2012). "Emotional Divergence Influences Information Spreading in Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI.

Qadir, J., Ali, A., Rasool, R. U., Zwitter, A., Sathiaseelan, A., and Crowcroft, J. (2016). "Crisis Analytics: Big Data Driven Crisis Response". In: *CoRR* abs/1602.07813.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., and Ghosh, S. (2015). "Extracting Situational Information from Microblogs During Disaster Events: A Classification-Summarization Approach". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. Melbourne, Australia, pp. 583–592.

Sen, A., Rudra, K., and Ghosh, S. (2015). "Extracting situational awareness from microblogs during disaster events". In: *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*. IEEE, pp. 1–6.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., and Anderson, K. M. (2011). "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*.

Watson, H., Finn, R. L., and Wadhwa, K. (2017). "Organizational and Societal Impacts of Big Data in Crisis Management". In: *Journal of Contingencies and Crisis Management* 25.1, pp. 15–22.

Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. (2012). "Using social media to enhance emergency situation awareness". In: *IEEE Intelligent Systems* 27.6, pp. 52–59.

Yuan, S., Wu, X., and Xiang, Y. (2016). "Incorporating Pre-Training in Long Short-Term Memory Networks for Tweets Classification". In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, pp. 1329–1334.