Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 119-127 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Automated Text Annotation for Social Media Data during Natural Disasters

Si Si Mar Win*[,1], Than Nwe Aung[2]

[1]*University of Computer Studies, Mandalay, Web Data Mining Lab, 05071, Myanmar*

[2]*University of Computer Studies, Mandalay, Faculty of Computer Science, 05071, Myanmar*

ARTICLE INFO

Article history:
Received: 21 November, 2017
Accepted: 27 February, 2018
Online: 12 March, 2018

Keywords:
LIBLINEAR
Twitter
Annotation
POS
Ngram

ABSTRACT

Nowadays, text annotation plays an important role within real-time social media mining. Social media analysis provides actionable information to its users in times of natural disasters. This paper presents an approach to a real-time two layer text annotation system for social media stream to the domain of natural disasters. The proposed system annotates raw tweets from Twitter into two types such as Informative or Not Informative as first layer. And then it annotates again five information types based on Informative tweets only as second layer. Based on the first and second layer annotation results, this system provides the tweets with user desired informative type in real time. In this system, annotation is done at tweet level by using word and phrase level features with LIBLINEAR classifier. All features are in the form of Ngram nature based on part of speech (POS) tag, Sentiment Lexicon and especially created Disaster Lexicon. The validation of this system is performed based on different disaster related datasets and new Myanmar_Earthquake_2016 dataset derived from Twitter. The annotated datasets generated from this work can also be used by interested research communities to study the social media natural disaster related research.

1. Introduction

Today, online social networking sites like Twitter, YouTube Facebook and Weibo play the important news sources during mass emergencies. Among them, Twitter, the most popular social networking site, provides a wealth of information during a natural disaster. It is often the first medium to break important disaster events such as earthquakes often in a matter of seconds after they occur and more importantly. Recent observation proofs that some events and news emerge and spread first using this media channel rather than other the traditional media like online news sites, blogs or even television and radio breaking news.

People also used social media to share advice, opinions, news, moods, concerns, facts, rumors, and everything else imaginable. Corporations use social media to make announcements of products, services, events, and news media companies use social media to publish near real-time information about breaking news. However, due to questionable source, uncontrollable broadcasting, and small amount of informational tweets among large number of non-informational tweets, Twitter is hardly an actionable source of breaking news.

Tweets from Twitter are highly vary in terms of subject and content and the influx of tweets particularly in the event of a disaster may be overwhelming. It is impractical to generation of efficient features vector based on uniform vocabulary. Therefore, effective feature extraction is first challenge. It is infeasible to automatically classify these varied tweets by using particular annotated corpora for specific messages of every disaster events. Cross event classification is a major challenge.

Another challenge is occurred by the development of supervised learning based systems trained on a single corpus and able to achieve a good performance over a broad range of different events. The annotation of corpus of messages for every disaster by human annotators is obviously time-expensive and practically infeasible on real time manner. Therefore annotation of tweets corpora by human is additional challenge.

In summary, the proposed system is aimed to address these issues by using three main functions: 1. Create annotated disasters corpus of tweets with five labels for Informative tweets on real

*S. S. M. Win, U.C.S.M, Email: sisimarwin@gmai.com

https://dx.doi.org/10.25046/aj030214

time manner. 2. Competitive, easily implementable feature extraction method that act as a benchmark for automated accurate classification approaches for natural disaster related datasets by using natural disaster lexicon. 3. Creation of extended natural disasters lexicon based on publicly available annotated datasets and newly annotated corpus. This paper is an extension of the work originally presented in IEEE/ACIS 16th International Conference on In Computer and Information Science (ICIS) [1]. In the previous work, we identified the tweets into only three labels such as Informative, Not Informative and Other Information as single layer annotation. Therefore, we continue to identify the Informative tweets into more specific information types. Our annotation model in this paper is based on a more relevant and small set of features than our previous work.

The rest of the paper is organized as follow: Section 2 presents the overview of closely related work to this paper. Section 3 explains the methodology that we used in collecting, preprocessing, feature extraction, disaster lexicon creation, and classification scheme used for annotating the tweets. Section 4 describes the architecture of the proposed system. Section 5 expresses the datasets details, experiments and analysis performed. Section 6 summarizes the results from our analysis and highlights the implications of our results. In this section, we also describe the future work of the proposed system.

2. Related Work

This section presents the current state-of-the-art systems, algorithms and methodologies to access the social media data analysis. Social media allows users to exchange small digital content such as short texts, links, images, or videos. Although it is a relatively new communication medium compared with traditional media, microblogging has gained increased attention among users, organizations, and research scholars in different disciplines. There are several researches on social media mining for text based data for classification and prediction of informational posts in different domains.

Among them, the authors in [2] proposed Artificial Intelligence for Disaster Response (AIDR) system to annotate the posts from Twitter into a set of user defined categories such as damage, needs etc. by using hybrid unigram and bigram features. In AIDR system, the tweets were identified into Informative and Non-Informative types during Pakistan earthquake in 2013.

The authors in follow up study automatically and provided human annotated Twitter corpora for 19 different events that occurred between 2013 and 2015. They also experimented their corpora by using the similar features set [3].

The other authors also presented the Tweedr, twitter-mining tool, to retrieve actionable information from Twitter. They applied several different types of features for their CRF clustering. For each token in a tweet, they extracted capitalization, pluralization, whether it is numeric or includes a number, WordNet hypernyms, Ngrams, and part of speech tags to provide specific information about different classes of infrastructure damage, damage types, and casualties [4].

The authors in [5] used six types of features such as Tweet Meta-data Features, Tweet Content Features, User based Features, Network Features, Linguistic Features and External Resource Features for credibility analysis.

They also developed a real time web application, TweetCred, to provide the one of the seven credibility scores of user generated content on Twitter by using 45 features. They tested their application within three weeks period. Their result showed that high credibility tweets were 8% [6].

Moreover, hashtags have been effectively utilized as critical features for various tasks of text or social media analysis, including tweet classification system [7].

In [8], the authors studied the linguistic method to analyze the importance of linguistic and behavioral annotations. They applied the datasets of four crisis events such as Hurricane Gustav in 2008, the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods, and the 2010 Haiti Earthquake. They observed that the usage of specific vocabulary to convey tactical information on Twitter can achieve higher accuracy than the usage of bag of words (BOW) model for classification of context-aware tweets.

The classification of tweets into Credible or Not Credible was presented in [9]. However, most of the recent research focuses on the information extraction and detection of situational awareness during natural disasters, it is still needed to provide a cohesive pipeline that takes into consideration all of the facets of data extraction.

This system focuses on the content based features set such as Linguistic features, Disaster Lexicon based features, twitter specific features (hashtags and URLs), unigram POS tag features and other salient features from tweet content.

3. Methodology

At the core function of this system is the capability of annotating tweets into predefined information types in real time. We propose, implement and evaluate the approach for determining and assigning a label for each tweet, taking into account terms from the tweet itself and from disaster lexicon. For this study, we first collect the tweets from Twitter. And then we extracted content based features from the collected tweets.

3.1. Data Collection

This function works for tweets collection. It collects messages from Twitter for training and testing using the Twitter streaming API. At first, it collected different annotated datasets published by using annotated tweet_id from Imran et al. [3].

The new data collection process focuses on the exact matching of keywords to acquire tweets and build the query using user defined keywords or hashtags. Using the relevant keywords or hashtags for queries are the best way to extract the most relevant tweets during crisis or disasters. For example, #MyanmarEarthquake hashtag is applied to acquire the news of earthquake that struck in Myanmar.

3.2. Preprocessing

Firstly, this task removes the tweets which already contains the same text in the previous preprocessed tweets to reduce the redundancy and noise by using the cosine similarity.

Secondly, stop-words from tweets are removed to reduce dimensionality of the dataset and thus terms left in the tweets can be identified more easily by the feature extraction process. Stop-words are common and high frequency words such as "a", "the", "of", "and", "an" "in" etc. [10]. User mention and URLs are also eliminated.

Finally, we used lemmatization instead of stemming to convert all the inflected words present in the text into a base form called a lemma. For the purpose of lemmatization, the proposed system uses Stanford Core NLP.

3.3. Feature Extraction

The most important step in text analysis using supervised learning techniques is generating feature vectors from the text data or documents. This work is intended to build a real time system based on tweets from Twitter, feature extraction is therefore concerned with altering tweet contents into a simple numeric vector representation.

In our previous work, we used hybrid unigram and bigram, unigram Brown cluster, unigram part of speech (POS) tags, number of hashtags, number of URLs and two lexicon based features such as NRC hashtags lexicon and our disaster lexicon as our features set. We found these features outperform the neural word embeddings and only hybrid unigram and bigram features.

According to the constant vocabulary, hybrid unigram and bigram features outperforms the classification of same events (i.e. the training and test datasets are equal). However we need to annotate the unknown disaster events and the contents described for different events may have different vocabulary. Even the same type disaster events may contains the different language style.

The analysis of social media data is heavily rely on the ability to analyze text data. However, there are some unique considerations in the analysis of social media data that make it different than a normal text mining analysis. To overcome the informal social media data to be formal consideration, text in tweets are tokenized using ARK Tweet NLP [11]. This process receives the tweets from preprocessing step, it extracts the features by using ARK POS tagger and different lexicons. The features used in this work are only extracted from tweet contents.

To derive the most relevant feature, this work investigated the three types of feature extraction methods. The first one is BOW model with unigram and bigram based features used in AIDR. The second is the neural word embeddings (WE) model and the last one is the proposed content based features model.

This system proposes the features set based on the following observations:

1. Messages in tweets written by users for same disaster type may have composed of same terms. It is usually the case that the same disasters have the same terms such as shake, strike, magnitude for disaster earthquake.
2. Different natural disaster related tweets may have composed of same terms such as need, pray, pray for, damage, death, destroy, survivor, etc. and may have same syntactical style such as POS tag.
3. Similar words have similar distributions of words to their immediate left and right [11].
4. If a tweet contain more than two hashtags in its content, it may not be information tweet.
5. In crisis related tweets, hashtag may be assumed as topic word or keyword of these tweets.
6. Informative tweets may contain numerical word and URL.

Based on these observations 1 and 2, we decided to create and use Disaster Lexicon and word Ngrams. According to observation 3, we use Brown Word cluster. We also use number of hashtags and hashtag term, URL and numeral features due to the observation 4, 5 and 6. Ngrams POS features are used according to the observation 2. The proposed features used in this system are shown in Table 1.

Table 1. Features used in the proposed system

Feature	**Explanation**
Brown Cluster Ngrams	Unigram and Bigram of 1000 Brown clusters in Twitter Word Clusters made available by CMU ARK group
Count of disaster related terms	Number of disaster related terms as informative words in a created lexicon for disaster tweets.
Total PMI Score of disaster related terms	Total PMI scores of unigrams and bigrams words that occurred in the tweet and listed as strongly correlated with natural disaster in a disaster lexicon for tweets
Count of non-informational terms	Lexicon creation function of this system also identifies a set of terms which appear only in Not-Informative tweets across all natural disasters datasets.
Total PMI Score of non-informational terms	Total PMI Score for each set of unigrams, bigrams that mostly occur in the Not informative tweet.
Count of numerals	Expected to be higher in situational tweets which contain information such as the number of casualties, emergency contact numbers.
POS tag	Unigram part of speech tags that occur in the Tweet generated by CMU ARK POS-Tagger
Word Ngrams	Unigram and bigram of terms from Disaster Lexicon

To extract neural word embeddings (WE) features for baseline model, this system used Word2vec model in Deep Learning4J [12]. Word2Vec is the representations of words with the help of vectors in such manner that semantic relationship between words preserved as basic linear algebra operations. The following parameters were used while training for Word2Vec: 100 dimensional space, 10 minimum words and 10 words in context. After transforming 100 dimension feature vector of each word in the corpus, this system used t-Distributed Stochastic Neighbor embedding (t-SNE) technique to reduce from 100 dimensions of each feature vector to 10 dimensions feature vector.

3.4. Extended Disaster Lexicon Creation

This system creates the disaster lexicon which contains specific natural disaster related terms with a point wise mutual information (PMI) based score and frequency distribution of these terms based on the set of annotated disaster datasets. This lexicon creation process follows the method of Olteanu et al. [13]. In this process, we exploit their natural disaster related datasets, the other available natural disaster related datasets [2] and newly annotated dataset such as Myanmar_Earthquake_2016 dataset which are collected by proposed system for lexicon expansion or keywords (disaster related terms) adaptation.

The disaster creation process consists of two main parts. At first, to obtain the most relevant disaster terms, we create various disaster lexicons based on different datasets of same disaster type. For example, this work uses all available earthquake datasets such as 2015_Neapl, 2014_Chile, 2014_Calfornia and 2016_Myanmar earthquakes for creation of Earthquake Lexicon. In this phase, we used equal number of Informative and Not informative tweets for each disaster datasets based on same disaster types.

At second, we combined the different disaster lexicons into one master disaster related lexicon with unique unigram and bigram terms. In this step, we calculate the mean PMI score for the terms which are contained in the two or more lexicons.

The score of a term could be calculated from the PMI value of a term t in an informative context PMI (t, informative) and the same term in a non-informative context PMI (t, non-informative) using the equation:

$$\textit{InfoScore} = \textit{PMI (t, informative)} - \textit{PMI (t, non-informative)} \qquad (1)$$

Here PMI (t, informative) and PMI (t, non-informative) are calculated using:

$$PMI\ (t, orientation\ ) = log_2 \frac{freq(t, orientation).N}{freq(t).\ freq(orientation)} \qquad (2)$$

Where, freq (t) is the number of times term t appears in a tweet, while *N* is total number of terms in the tweet.

This automatically created lexicon is used in feature extraction process of the proposed system.

3.5. Feature Selection

Feature selection is an important problem for text classification. In feature selection, this work attempts to determine the features which are most relevant to the classification process. This is because some of the words are much more likely to be correlated to the class distribution than others. This system applied the information gain based feature selection method which is widely used for text classification.

Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. n this method, let P_i be the global probability of class *i*, and $P_i\ (w)$ be the probability of class *i*, given that the document contains the word w. Let *F(w)* be the fraction of the documents containing the word w. The information gain measure *I(w)* for a given word *w* is defined as follows:

$$I(w) = -\sum_{i=1}^{k} P_i\,.\log\,(P_i) +\ F(w).\sum_{i=1}^{k} P_i(\mathrm{w})\,.\log\,(P_i(w)) +$$

$$(1 - F(w)).\textstyle\sum_{i=1}^{k}(1 - P_i(\mathrm{w}))\,.\log\,(1 - (P_i(w)) \qquad (3)$$

The greater the value of the information gain *I(w)*, the greater the discriminatory power of the word *w*.

3.6. Annotation of Social Media Text

This system assess annotation of tweets by using supervised machine learning technique. This technique automatically classifies the information contained in tweets. To perform the annotation task, the proposed system trained a LIBLINEAR classifier operating on extracted features set. LIBLINEAR solves large-scale classification problems in many applications such as text classification. It is very efficient for training large scale. It takes only several seconds to train more than 600,000 examples while a Library for Support Vector Machines (LibSVM) takes several hours for same task [14].

Given a set of features and a learning corpus (i.e. the annotated dataset), the classifier trains a statistical model using the feature statistics extracted from the corpus and then annotates the tweets into Informative or Not Informative. This trained model is then employed in the classification of unknown tweets and, for each tweet, it assigns the probability of belonging to a class: Related and Informative as Informative, Not Related or Not applicable as Not Informative in first layer annotation. And then based on the Informative tweets, this system annotates again these informative into one of five types such as infrastructure damage, caution and advice, dead or injured people, needs and offer and Donations and volunteering as second layer annotation. The annotated datasets required by the system can be obtained from three sources such as AIDR, CrisisNLP which is the collection of tweets from 19 natural and man-made disasters and CrisisLexT26 which is the collection of tweets from 26 Crises [2, 3, 13]. This system uses datasets in English language only.

4. Architecture of the proposed System

The holy grail of text annotation is an automated system that accurately and reliably annotates very large numbers of cases using relatively small amounts of manually annotated training data. This work is intended to develop a two layer annotation system that automatically creates the different disaster datasets with annotated tweets. In this system, annotation is restricted to tweets in English language. Non-English tweets are not considered. Non-English tweets are not considered.

The system, illustrated in Figure, first collects the tweets from Twitter by using user desired query terms or target disaster related terms. After collecting the tweets, it removes the redundant tweets by using tweet_id and then it also eliminates the stop-words. In feature extraction, this system applies Linguistic features such as Brown cluster, Syntactic feature such as POS features, Lexical features using disaster lexicon and the other Twitter centric features such as Hashtags and URLs.

This system also analyzed which features are important in the data to annotation. It applied the annotated corpus to train a classifier that automatically annotates the tweets.

To improve model performance, the best set of 300 features were chosen by using Information gain theory based feature selection method.

In the ground truth annotation process, LIBLINEAR classifier uses these selected features subset for tweets categorization to create annotated corpus with Informative or Non-informative tweets and to provide informative tweets to the users.

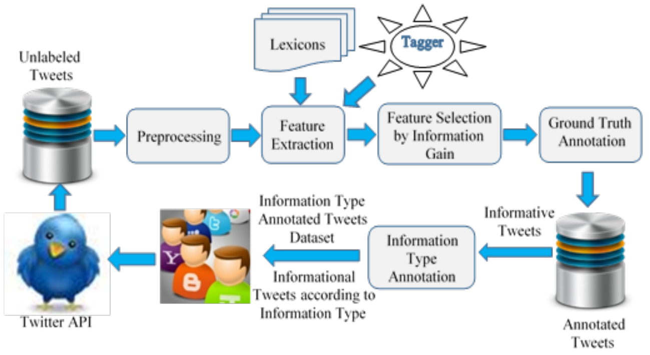
Lexicons
Tagger
Unlabeled Tweets
Preprocessing
Feature Extraction
Feature Selection by Information Gain
Ground Truth Annotation
Information Type Annotated Tweets Dataset
Informative Tweets
Information Type Annotation
Informational Tweets according to Information Type
Twitter API
Annotated Tweets

Figure 1: Architecture of the Proposed System

After annotating the collected tweets into one of the five information types, his system provides the informational tweets to users based on their desired type of information.

5. Experiments

This system performs a set of preliminary experiments to evaluate the effectiveness of feature extraction, feature selection model and classifier model on the performance of the proposed approach. For feature extraction, the proposed system applied three models such as neural word embedding, BOW with Unigram and Bigram model and the proposed model.

The final experiment is done under the best development settings in order to evaluate the classifier model with the best feature set. This section presents experiments and results for classification of four annotated datasets. The results along with the experimentation of different datasets are described based on accuracy, precision, recall, and F1 score of classifier model for feature extraction performance.

5.1. Datasets and Setting

In order to evaluate the effectiveness of the proposed social media text annotation strategies for identifying informative tweets during natural disasters events, the experiments of this system used people freely available 10 annotated natural disaster datasets . These datasets are already annotated with different information types.

To reduce the noise in training data, this system discarded all the following tweets.

1. The tweet where an information type clash is observed. An information type clash is a tweet that may happen two or more different type and may ambiguous in the dataset.

2. "Not labeled" tweets.

3. "Animal Management" are also eliminated.

The tweets with similar information types such as "Infrastructure damage" and "Infrastructure and utilities" are combined as "Infrastructure and utilities". "Injured or dead people", "missing or found people", "displaced people and evacuation" and "personal updates" tweets are combined as "Affected individuals" and "donation needs or offer volunteering services" and "Money" are also combined as "Donations and volunteering".

Before training the corpus for second layer annotation, the informative tweets with non-specific information type such as "Other Useful Information" are also discarded.

Detailed information of datasets is described in Table 2 and Table 3. In this table Type 1 refers to the information type "Affected individuals", Type 2 refers to "Infrastructure and utilities", Type 3 means "Donations and volunteering", Type 4 is "Caution and advice", and Type 5 refers to "Sympathy and emotional support".

Table 2. Natural disaster datasets details including disaster type, name, number of informative tweets, number of Not Informative tweets and total tweets.

Type	Disaster Name	Info	Not-Info	Total
Floods	2013_Queensland_floods (QF)	728	281	1009
Bushfire	2013_Australia_bushfire (AB)	691	261	952
Typhoon	2013_Typhoon_Yolanda (TY)	765	175	940
Wildfire	2012_Colorado_wildfires (CW)	685	247	932
Earthquake	2014_Chile_earthquake (ChiE)	1834	179	2013
Floods	2013_Colorado_floods (CF)	589	190	779
Earthquake	2014_Costa_Rica_earthquake (CE)	842	170	912
Floods	2014_Manila_floods (MF)	628	293	921
Floods	2012_Phillipines_Floods (PF)	761	145	906
Floods	2013_Alberta Floods (AF)	684	297	981
Floods	2014_India_floods (IF)	940	396	1336

Table 3. Natural disaster datasets statistics for five information types

Dataset	Type 1	Type 2	Type 3	Type 4	Type 5
QF	207	113	55	114	17
AB	199	65	35	70	33
TY	77	106	383	20	63
CW	44	128	62	69	25
NE	6	165	239	1215	458
IF	30	792	42	51	25
ChiE	55	3	70	14	58

The proposed system performed 10 fold cross validation to test the efficiency of the feature extraction and the model. In the experiments of classification, the proposed system used the set of tweets from five natural disasters such as 2013_Queensland_floods denoted by QF, 2013_Australia Bushfire as AB, the set of tweets for 2013_Typoon_Yolanda as

TY, 2012_Colorado_wildfires as denoted by CW, and the set of tweets from 2012_Costa_Rica_Earthquake as CE respectively.

5.2. Effectiveness of Feature Extraction

To choose the best classification model, we tested the extracted feature set on four different classifiers such as Random Forest, Sequential Minimal Optimization (SMO) which is the fast training algorithm for Support Vector Machine (SVM), Naïve Bayes and our LIBLINEAR classifier that are well known in text classification process. Due to the experiments on the previous work, the performance of Random Forest, Naïve Bayes and SMO was sensitive to the large number of features. Therefore, this system used LIBLINEAR classifier with Information Gain based feature selection method to get better performance and to reduce inconsistent features. This wok uses a well-known WEKA machine learning tools for implementation of Random Forests, Naïve Bayes, SMO, LIBLINEAR and Information Gain based feature selection methods [15]. The results of our previous work are described in [10]. Due to these results, we selected the LIBLINEAR classifier as our classification model.

In this work, the proposed feature extraction method and two baseline methods are evaluated by experiments on ten datasets. To compare the performance of the different feature-models (using LIBLINEAR classifier) under three scenarios such as (i) in-domain classification, (ii) cross event classification and (iii) cross-domain classification, where the classifier is trained with tweets of one event, and tested on another event are considered in this system.

5.2.1. In Domain Classification

In this type of classification, the classifier is trained and tested with the tweets of the same event. To evaluate the in-domain performance of each model, the proposed system followed a 10-fold cross validation process: each dataset was randomly split in 10 different non overlapping training and test sets. The Accuracy, Precision, Recall and F-Measure were calculated as the weighted average of these values over all the 10 test sets.

Table 4. Classification results in terms of Precision, Recall, F-Measure and Accuracy across Informative and Not Informative classes.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.791	**0.813**	0.612	0.812	**0.822**	0.782	0.785	**0.816**	0.687
AB	**0.772**	0.754	0.567	**0.79**	0.765	0.753	0.768	0.758	0.647
TY	0.797	**0.802**	0.685	**0.825**	0.816	0.828	0.788	**0.807**	0.750
CW	0.813	**0.815**	0.676	0.818	**0.820**	0.714	0.711	**0.817**	0.633

Table 4 represents a summary of evaluation for in domain classification by weighted average Precision (P), Recall (R) and F-Measure (F-M) of the classification results on four datasets.

According to the results, BOW with hybrid unigram and bigram model would perform relatively well in in-domain classification, since the training event and test event share a common vocabulary. However, the performances of the proposed features model is as good as BOW method.

5.2.2. Cross Event Classification

In this type of classification, where the classifier is trained with tweets of one event, and tested on another event. The result is significant since it shows that good classification can be achieved even without considering the type of disasters.

Table 5, Table 6, Table 7 and Table 8 also show the cross event classification performance on AB, TY, CW and QF dataset as training and the remaining datasets as testing data using the features sets extracted by the baseline method, proposed method and neural word embeddings method. The results in these tables indicated that the proposed method yields a high accuracy by using the LIBLINEAR algorithm in predicting certain classes in cross event classification.

Table 5. Classification results in terms of Precision, Recall, F-Measure using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Australia_Bushfire as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.77	**0.82**	0.76	0.73	**0.83**	0.50	0.74	**0.82**	0.52
TY	0.78	**0.81**	0.77	0.74	**0.83**	0.52	0.76	**0.81**	0.57
CW	0.79	**0.83**	0.75	0.80	**0.83**	0.66	0.80	**0.84**	0.66

Table 6. Classification results in terms of Precision, Recall, F-Measure across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Typhoon_Yolanda as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	**0.77**	**0.77**	**0.77**	0.64	**0.793**	0.79	0.67	0.755	**0.77**
AB	0.72	**0.77**	0.76	0.554	**0.785**	0.78	0.579	0.748	**0.75**
CW	0.74	**0.796**	0.76	0.589	**0.804**	0.78	0.605	**0.788**	0.76

Table 7. Classification results in terms of Precision, Recall, F-Measure across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2012_Colorado_Wildfire as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.759	**0.795**	0.74	0.73	**0.808**	0.37	0.745	**0.80**	0.36
AB	0.769	**0.797**	0.72	0.756	**0.806**	0.38	0.762	**0.799**	0.35
TY	0.762	**0.810**	0.76	0.673	**0.809**	0.36	0.703	**0.809**	0.37

Table 8. Classification results across two classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE) for 2013_Queensland_floods as training set.

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
AB	0.772	**0.817**	0.76	0.727	**0.831**	0.60	0.743	**0.818**	0.62
TY	0.771	**0.76**	0.76	0.641	**0.793**	0.65	0.672	**0.755**	0.68
CW	0.759	**0.795**	0.75	0.73	**0.808**	0.62	0.741	**0.799**	0.64

According to the experimental results, the performance of the BOW (hybrid unigram and bigram) and WE models is significantly inferior to the proposed model for cross-event classification. This is because the training and testing datasets (related to two different disaster events) have very different vocabularies. On the other hand, the classifier based on the proposed features significantly out-perform these two models in all cases. This implies that the selected features can separate between Informative and not-Informative tweets irrespective of the vocabulary and linguistic style related to specific events. Thus, classifiers can be trained over these features extracted from past disasters, and then deployed to classify tweets posted during future events.

5.2.3. Cross Domain Classification

In cross domain classification, to assign one of the two classes for first layer annotation and one of the five predefined categories (e.g. Affected individuals, Infrastructure and utilities, Donations and volunteering, Caution and advice, Sympathy and emotional support etc.) for second layer annotation to the tweet, the classifier requires sufficient training examples to learn about each pre-defined category. The proposed system used multiple past disasters of various types to train the classifier to robustly identify the different types of tweets for future natural disasters. The experiment for two classes classification (i.e. Informative and Not Informative), the proposed system used the set of tweets from Philippines (PF), Colorado (CF), and Queensland floods (QF) as the training set, denoted by PCQ, the set of tweets for Manila floods as the development set, denoted by MF, and the set of tweets from Alberta and Sardinia floods as two independent test sets, denoted by AF and SF, respectively. The results of this experiment are shown in Table 9.

Table 9. Experiments performed using the combined 2012_Philipinnes_flood, 2013_Colorado_floods and 2013_Queensland_floods as training set

Train /Test	Accuracy	Precision	Recall	F-Measure
PCQ/MF	80.34%	0.8609	0.803	0.808
PCQ/AF	76.47%	0.754	0.76476	0.7414
PCQ/SF	66.58%	0.6463	0.66579	0.6121

In the other experiments over five classes or multi-classes classification, this system combined the three or four datasets of different disaster types as training and the other one for testing data. For example, taking Colorado floods, Costa-Rica-earthquake, Philippine floods, Pablo-typhoon and Australia-Bushfire (CCPPA) as training dataset and the other datasets as individual test data. The classification results are shown in Table 10. Table 11 shows the classification results over the five classes of LIBLINEAR with 10-fold cross validation for six datasets by using three feature models.

Table 10. Classification results in terms of Precision, Recall, F-Measure across all five classes.

Train /Test	Accuracy	Precision	Recall	F-Measure
CCPPA/MF	72.3%	0.7308	0.723	0.719
CCPPA/TY	64.9%	0.72	0.649	0.663
CCPPA/CW	86.4%	86.8%	0.864	0.862
CCPPA/ChiE	79.5%	0.8214	0.795	0.7908

Table 11. Classification results across five classes using (i) (BOW), (ii) Proposed (P) (iii) Word Embeddings (WE)

Test Data	Feature Extraction Model								
	Weighted Avg. P			Weighted Avg. R			Weighted Avg. F-M		
	BOW	P	WE	BOW	P	WE	BOW	P	WE
QF	0.56	**0.581**	0.25	0.56	**0.586**	0.415	0.549	**0.576**	0.287
AB	0.537	**0.583**	0.22	0.569	**0.607**	0.465	0.532	**0.582**	0.295
TY	0.694	**0.701**	0.36	**0.724**	0.713	0.596	0.695	**0.705**	0.445
CW	0.653	**0.67**	0.18	0.64	**0.681**	0.427	0.631	**0.666**	0.255
NE	**0.749**	0.712	0.67	**0.761**	0.726	0.681	**0.738**	0.713	0.666
CE	0.757	0.78	0.58	0.783	**0.79**	0.76	0.754	**0.78**	0.658

According to the results in Table 10 and 11, the performance of LIBLINEAR classifier with proposed feature model outperforms the BOW and WE models in most cases and it can identify informational tweets at 67% accuracy on average. The performance of BOW and WE models sometime close to less than 20 %. This indicates that lexical features are critical to solve the ambiguous of information types.

5.3. Effectiveness of Annotation

In this section, the validation of this system on a real disaster study by classifying the data of Myanmar earthquake collected by Twitter API. The 6.8 magnitude earthquake that struck Myanmar on August 24th , 2016 is among the strongest in recent Myanmar history. The shaking was clearly perceived in all Central and Northern Myanmar and caused 4 deaths and several damage to the Pagodas of the area of Bagan. This dataset is crawled for a three days period from August 24th to 26th , 2016 by using the hashtags (#Myanmar, #Bagan, #earthquake, #Myanmarearthquake). And then it was randomly selected 1,800 tweets and was manually annotated based on the available news media in Myanmar such as Myanmar Times, The Global New Light of Myanmar and The Mirror.

Table 12. First Layer Annotation Results of Proposed features by LIBLINEAR

Dataset	Precision	Recall	F1	Accuracy
AB	0.897	0.895	0.895	89.45%
TY	0,912	0,92	0,913	92,02%
IF	0.908	0.908	0.908	90.82%
NE	0.748	0.751	0.749	75.05%

Table 13. Second Layer Annotation Results of Proposed features by LIBLINEAR

Dataset	Precision	Recall	F1	Accuracy
AB	0.682	0.693	0.685	69.31%
TY	0.768	0.786	0.774	78.60%
NE	0.815	0.843	0.828	84.26 %
IF	0.731	0.782	0.755	78.18 %

Base on cross domain classification over all five classes where we train the classifier on one dataset and test on another dataset, the experimental results using 2012-Costa-Rika- and 2014_Chile Earthquake as training data and Myanmar Earthquake as test data confirmed the expected classification of this work.

Myanmar_Earthquake_2016 was successfully annotated with predefined two labels at 75% accuracy on average and five labels also 74 % which is pretty high. The results of each annotation layer for four datasets are shown in Table 12 and Table 13.

5.4. Real Time Annotation

We showed a proof of real time model which takes a direct stream of new tweets as input test set and takes manually annotated tweets from previous disaster as training set, and then uses automated techniques to annotate the new tweets.

In the first layer, this system annotated the tweets into Informative if the tweets contain the information about the target disaster and Not informative for the remaining tweets. In the second layer, each informative tweet from previous layer is annotated into one of the five information types with respect to the information that contained in it.

To do this, we developed and deployed a real time system in the form of a Web application using Java 2 Platform Enterprise Edition (J2EE) and Twitter API binding library (Twitter4j). We analyzed the deployment and usage activity of our application from 24th August, 2016 to 28th August, 2016 which was the day of earthquake and days after an earthquake in Myanmar. For analysis and statistics, we collected the annotated datasets of our system for only three days period since August 24th to August 26th. In real time annotation process, we used the combined 2014_Chile_Earthquake and 2015_Nepal_Earthquake as training set. The number of collection and annotation times per day is 6. Each collection time is 4 minutes for 2000 tweets and annotation time is only 1 minute for each layer. Among the collected tweets, 62% of tweets are retweets and half of them are redundant tweets. Most of them are similar in text. Although the number of tweets collected in each time was 2000, the number of tweets annotated by our system is at most 750 because of the tweet cleaning step in preprocessing. The tweets in annotated datasets from each day are also redundant. The total unique automatically annotated tweets over three days, are nearly 2000.

After manually annotating the tweets, we compared the automatically annotated tweets from our system with manually annotated tweets to obtain the performance of our real time system. The manual annotation process is described in the previous section. According to the analysis results, the annotation accuracy of new tweets by our system is 80% in first layer and 74 % on second layer annotation.

5.5. Findings and Discussion

As mentioned above this system used 10 datasets for training and testing for evaluating classifier models and feature extraction models. Another new dataset for testing again for overall performance of the proposed system. According to the initial experimental results of BOW and word embeddings were very sensitive and depend on the vocabulary. The results of three feature extraction methods, the proposed method always outperforms the other two methods. Therefore, the proposed feature extraction model with LIBLINEAR classifier was chosen for second layer annotation process for categorizing the tweets into five specific frequently found information type.

6. Conclusion

Social media mining for disaster response and coordination has been receiving an increasing level of attention from the research community. It is still necessary to develop automated mechanisms to find critical and actionable information on Social Media in real-time. The proposed system combines effective feature extraction using NLP and machine learning approach to obtain the annotated datasets to improve disaster response efforts. Expanded disaster lexicon is also used to extract the relevant disaster related lexical features for annotation.

The proposed feature extraction method significantly outperforms the standard bag of words model and neural word embeddings model. By using LIBLINEAR classifier based on the proposed method, this system successfully annotated five information types to the Myanmar Earthquake data at 74% accuracy on average. In future, we will investigate the specific variation of terms over different disasters to perform annotation on all disaster types. We hope to formalize disaster lexicon in more detail to improve cross domain classification accuracy.

References

[1] S. S. M. Win, T.N. Aung, “Target Oriented Tweets Monitoring System during Natural Disasters”, In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference, IEEE, 2017.

[2] M. Imran, C. Castillo, J. Lucas, P. Meier and S. Vieweg, “AIDR: Artificial intelligence for disaster response”, In Proc. of WWW (companion). IW3C2, 2014, pp. 159–162.

[3] M. Imran, P. Mitra, C. Castillo: "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages." In Proceedings of the 10th Language Resources and Evaluation Conference (LREC), pp. 1638-1643. May 2016, Portorož, Slovenia.

[4] Z. Ashktorab, C. Brown, M. Nandi, A. Culotta, “Mining Twitter to Inform Disaster Response”, In Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA, May 2014.

[5] A. Gupta, P. Kumaraguru, “Credibility Ranking of Tweets during High Impact Events”, PSOSM’12, 2012.

[6] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, “TweetCred: Real-time credibility assessment of content on Twitter”, In Proc. Of SocInfo. Springer, 228–243, 2014.

[7] A. Stavrianou, C. Brun, T. Silander, C. Roux, “NLP-based Feature Extraction for Automated Tweet Classification”, Interactions between Data Mining and Natural Language Processing: 145.

[8] W. J. Corvey, S. Verma, S. Vieweg, M. Palmer and J. H. Martin, “Foundations of a Multilayer Annotation Framework for Twitter Communications During Crisis Events”, In Proc. International Conference on Language Resources and Evaluation (LREC'12), May 2012

[9] C. Castillo, M. Mendoza and B. Poblete, “Information Credibility on Twitter”, International World Wide Web Conference Committee (IW3C2), Hyderabad, India, 2011.

[10] C. C. Aggarwal and C.-X. Zhai, “Mining Text Data, Springer”, 2012.

[11] K. Gimpel et al., “Part-of-speech tagging for Twitter: Annotation, features, and experiments”, In Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume, Portland, June 2011.

[12] Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. http://deeplearning4j.org.

[13] A. Olteanu, C. Castillo, F. Diaz and S. Vieweg, “CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises”, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2014.

[14] R.E Fan, K.W Chang, C.J Hsieh, X.R Wang and C.J Lin, “LIBLINEAR: A Library for Large Linear Classification”, Journal of Machine Learning Research 9, 2008, pp. 1871-1874.

[15] E. Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data mining: Practical machine learning tools and techniques", Morgan Kaufmann, Fourth Edition, 2016.