Automatic Classification of Disaster-Related Tweets

Beverly Estephany Parilla-Ferrer, Proceso L. Fernandez Jr., PhD, and Jaime T. Ballena IV, PhD

Abstract**—** The social networking site Twitter has become one of the quickest sources of news and other information. Twitter information feeds known as tweets, are voluntarily sent by registered users and reach even non-registered users, sometimes ahead of traditional sources of mass news. In this study, we develop some machine learning models that can automatically detect informative disaster-related tweets.

A dataset of tweets, collected during the Habagat flooding of Metro Manila in 2012, was used in building the classifier models. A random subset of this dataset was manually labeled as either informative or uninformative to produce the ground truth. Two machine learning algorithms, Naive Bayes and Support Vector Machine (SVM), were used to build models for the automatic classification of the tweets, and these models were evaluated across the metrics of accuracy, precision, recall, area under curve and F-measure. Experimental results show that the model generated from SVM has significantly better results compared to that of the Naive Bayes.

This study also revealed that uninformative tweets outnumbered informative tweets, suggesting that the subscribers used Twitter to broadcast more of tweets that express their subjective messages and emotions regarding the Habagat event. However, the informative tweets were more likely to be retweeted than uninformative tweets, indicating that subscribers retweet messages they deem informative and useful for public awareness. These insights, together with the built classifier models, can help in the development of a system that can sift through the voluminous Twitter data and in real-time detect informative disaster-related tweets so that appropriate action may be done promptly.

Keywords**—** Disaster, Machine learning, Text mining, Tweets

I. Introduction

DURING disasters and emergencies, microblogs, have been used by people whether from the private or public sector, local or international community, as a medium to broadcast their messages. This social medium is being considered as a means for emergency communications because of its growing ubiquity, communications rapidity, and cross-platform accessibility [1]. The interactions on social media being highly distributed, decentralised and occurring in real time, provide the necessary breadth and immediacy of information required in times of emergencies [2].

Twitter is one microblogging service that allows its subscribers to broadcast short messages, called tweets, of up to 140 characters. These tweets are used to share relevant information and report news [3]. In emergency situations, tweets provide either first-person observations or bring relevant knowledge from external sources [1]. Twitter is becoming a valuable tool in disaster and emergency situations, as there is increasing evidence that it is not just a social network, it is also a news service [4]. Relevant tweets shared by users is a vital source of information and is useful in understanding and visualizing the situation of affected parties. This medium is seen as a place for "harvesting" information during a crisis event to determine what is happening on the ground [5]. The growing use of social media during crises offers new information sources from which the right authorities can enhance emergency situation awareness which is significantly recognized as a critical part of making successful and effective decisions for emergency response [6].

Tweets highly vary in terms of subject and content and the influx of tweets particularly in the event of a disaster may be overwhelming. It consists of socio-behaviors that include intensified information search and information contagion [7]. Microblogging offers ways to retrieve, produce and spread information; the nature of that sharing has a lifecycle of information production and consumption that is rapid and repetitive [1]. Since these varied tweets are rapidly broadcasted, it is imperative to automatically classify the tweets in order to extract needed information. The availability and accessibility of disaster-relevant information can contribute to an effective and efficient disaster response mechanism, which eventually can alleviate damages or loss of life and property during a disaster or crisis.

Disaster-related tweets are one of the many subjects of text mining researches nowadays. Specifically on the area of classifying and extracting information from disaster-related tweets, Caragea et al conducted a study to classify the set of tweets collected during the Haiti earthquake for the emergency response sector [8]. The authors compared different feature representations for training SVM classifiers

Beverly Estephany Parilla-Ferrer is with the Department of Information Technology and Computer Science, School of Information and Computing Science Saint Louis University, Baguio City, Philippines(+637409205884351; e-mail:beverlyestephanyferrer@gmail.com).

Proceso L. Fernandez Jr., is with the Department of Information Systems and Computer Science, School of Science and Engineering, Ateneo de Manila University, Quezon City, Philippines (e-mail:pfernandez@ateneo.edu).

Jaime T. Ballena IV is with the Math Department, School of Information and Computing Science Saint Louis University, Baguio City Philippines (e-mail: jimwhaleiv@gmail.com).

http://dx.doi.org/10.15242/IIE.E1214072

to classify tweets. Imran et al extracted information using disaster-related message ontology to classify tweets using the Joplin dataset [9]. Multi-level and multi-label classification using Naive Bayes classifier in Weka was used in the study. Another paper of Imran et al focused on the classification and extraction of disaster-relevant information from the Joplin and Sandy dataset using conditional random fields for training [10].

These studies have clearly presented that disaster–relevant information can be classified and can provide information that can augment people's awareness on incidents. However, these studies did not cover statistical analysis on the big data of tweets and the performance evaluation of machine learning algorithms for the classification of tweets. Although there are several studies that have evaluated machine learning algorithms for the classification of tweets, these studies dealt with sentiment or opinion analysis. In this study, we aim to create a machine learning model to classify disaster-related tweets as informative or uninformative and compare the performance of two of the most common machine classifying algorithms Naive Bayes and Support Vector Machine. Performance evaluation is based on the validation of results across the metrics of accuracy, precision, recall, area under curve and F-measure, with the application of statistical tools. Furthermore, the research investigates the information that can be extracted from the statistics of broadcasted tweets during the Habagat incident which caused widespread flooding in Metro Manila in 2012.

II. Related Works

There are several researches on text mining for classification and prediction on various domains such as in the medical, business, crime investigation, e-mail detection, etc. The following works are focused on the classification of tweets and the comparison of classifying algorithms.

Sriram et al extracted features from author's profile and text using bag of words(BOW) and 8F approach [11]. With the use of Naive Bayes in Weka, the 8F approach dominated BOW in classifying a tweet as News, Events, Opinions, Deals or Private messages. A study conducted by Mahendran et al classified microblogs as Positive, Negative or Neutral using bag of words as features and Naive Bayes and MaxEnt as classifiers [12]. The results of the experiment showed that the domain- specific features extracted from the author's profile and text effectively help in the accurate classification of the microblogs. Lee et al classified Twitter trending topics into 18 general categories such as sports, politics, etc., using text-based and network-based classification using various classification models [13]. Based on experimental results, Network-based classifier (71%) outperformed text-based classifier (65%). Imran and Castillo used the Artificial Intelligence for Disaster Response(AIDR) platform to classify in real-time social media messages for public health, specifically in this experiment detection of flu [14]. This platform learns automatically from human annotated examples and classifies tweets into user-defined categories. The experimental results returned 75% classification accuracy based on AUC metric. Prasetyo et al used SVM algorithm to classify software related microblogs as relevant or irrelevant to engineering software systems with text from URLs and microblogs as features [15]. Training and testing was performed using 10-fold cross validation and the model revealed a significantly good performance based on accuracy, precision, recall and F-measure.

On the area of comparing machine learning algorithms for classification using short text messages as dataset, the following researches were conducted. Duwairi and Qarqaz compared Naive Bayes, SVM and K-nearest Neighbor as implemented in Rapidminer to classify sentiments of tweets as positive, negative or neutral using a dataset on general topics such as education, sports and political news [16]. With 10-fold cross validation, SVM returned the highest precision, while K-Nearest Neighbor (KNN) with the highest recall. A study on the classification of Reuters headline news as dataset, Khamar compared SVM, K-Nearest Neighbor and other algorithms [17]. After a process of training and testing, KNN returned a higher accuracy compared to Naive Bayes and SVM. Lu conducted a study to identify online messages using C4.5, Naive Bayes and SVM [18]. Based on experiment, SVM outperformed C4.5 and Naive Bayes in terms of accuracy and F-measure. Zielenski and Bugel investigated on classifying tweets posted in 4 different languages (English and 3 Mediterranean languages in Turkey, Greece and Romania) as relevant or not relevant to an earthquake event by testing a language-specific detection classifier with keywords that are synonyms or translations of the word earthquake as features in the classification [19]. Training and testing used different datasets using regular expression and Naive Bayes. The results showed a best performance on the official languages but worst in the English language.

III. Method

The automatic classification of tweets begins with the manual classification of a dataset which serves as the ground truth for evaluating the performance of two machine classifying algorithms, Naive Bayes(NB) and Support Vector Machine(SVM). The following sub-sections describe the dataset and the approach used in the study.

3.1 *Data Source*

Habagat hit the Philippine's capital Manila and its neighboring provinces last August 1-8, 2012. The monsoon brought about eight days of torrential rain and thunderstorms which caused flooding in several areas and consequently caused massive damages and loss of properties and lives. At the onset of the Habagat until its aftermath, subscribers of Twitter used this social medium to send relevant or personal messages to their intended recipients. A sample of Habagat tweets were collected by the researchers of Ateneo de Manila University using the Twitter API. The sample has a total of

612,622 tweets, of which 373,771 are unique tweets and 238,851 are retweets. Unique tweets are the original messages that are sent by the author of a tweet which can be viewed by his or her followers and followees. Retweets on the other hand are messages received by a subscriber and are forwarded to another user or set of users.

3.2 *Manual Classification*

From the collected Habagat tweets, a sample of 4,000 tweets was randomly selected. Annotators initially classified the randomly selected tweets as to whether they are encoded in English, Tagalog, combination of English-Tagalog or other languages or dialects. The annotators further classified the English tweets as informative or uninformative based on the given definitions. Informative tweets are tweets that provide useful information to the public and are relevant to the event, while uninformative tweets are tweets that are not relevant to the disaster and these do not convey enough information or are personal in nature and may only be beneficial to the family or friends of the sender.

Using the data tool of a spreadsheet application, the unique tweets and retweets were extracted from the same dataset. Three annotators were asked to label the tweets taking into consideration the text of the tweets and links that may be included in the tweet were disregarded. The tweets were independently labeled by the annotators and were eventually collated for agreement. In cases where labels were non-uniform, the annotators argued on the items and came up with a unanimous label. In this paper, the unique English tweets labeled as informative or not informative served as the dataset for the machine learning.

After labeling, the Intraclass Correlation Coefficient(ICC) or multi-rater Kappa coefficient was computed to estimate the multi-rater strength of agreement among the annotators. Its most common application is, like the Kappa tests, the test of agreement between *k* raters on *n*-subjects [20]. Lu states that the best measure of reliability for continuous data is the intra-class correlation coefficient [21]. This is essential in the process because a substantial level of agreement among the annotators signifies the reliability of the labeling by different annotators.

3.3 *Information Extraction*

Using conditional probability and Bayes' theorem, information can be extracted from the statistics of manually classified tweets. Conditional probability is defined as $P(A|B) = P(A \cap B)/P(B)$, provided $P(B) > 0$. Bayes' theorem , also known as Bayes' rule or Bayes' law, is a result in probability theory that relates conditional probability. If A and B denote two events, P(A|B) denotes the conditional probability of A occurring, given that B occurs [22]. Bayes theorem is mathematically defined as:

$$P(A|B) = P(B|A)\, P(A)/ P(B)$$

where:

P(A) is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B

P(A|B) is the conditional probability of A, given B

P(B|A) is the conditional probability of B given A

P(B) is the prior or marginal probability of B, and acts as a normalizing constant

In the context of this study, *P(A)* is the probability of a tweet being informative, while *P(B)* is the probability of a tweet being unique. Therefore, information of the probabilities of tweets being informative or not informative, given that these are unique or are retweets were then extracted.

3.4 *Machine Learning and Classification*

3.4.1 *Preprocessing methods and Generation of Features*

Pre-processing procedure as implemented in Rapidminer such as tokenization, stemming, and stop words removal were applied on the corpus to create the word vector. Tokenization is the process of breaking up of text into words or phrases called tokens while stemming is the conversion of words into its base form. The purpose of this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time [23]. In this experiment, Porter's algorithm was used, being the most popularly used stemming algorithm. Stop words removal is the process of removing words in the corpus which are language specific functional words that carry no information [23]. Examples of stopwords are the words a, an, the, and, so, etc.

The following is an actual example of a Habagat tweet which was preprocessed.

Actual Habagat Tweet:
RAINFALL WARNING SIGNAL Code Red Warning For Metro Manila Issued at PM
Stop words removal:
RAINFALL WARNING SIGNAL Code Red Warning Metro Manila Issued PM
Porter's stemming:
rainfal warn signal code red warn metro manila issu pm

The bag of words technique was used to produce the main features in the word vector. However in this experiment, words which appeared once in the entire corpus and words with length less than two characters were pruned. The word vector was generated using the binary term occurrence as weights for the features and the generated word vector was used as the training dataset for machine learning.

3.4.2 *Machine Learning Algorithms for Classification*

3.4.2.1 *Supervised Learning*

Supervised learning was used in training the machine to classify a tweet as informative or not informative. Supervised learning is a training in which the class attribute values for the dataset are known (labeled data) before running the algorithm [24]. Supervised learning builds a model that maps **x** to *y*;

where **x** is a vector and *y* is the class attribute. A model is generated when the supervised learning algorithm is run on a training set, which maps the feature values (**x**) to the class attribute values (*y*). After training, the model is tested on a dataset which will predict class attributes. In the context of this study, **x** = vector of features and $y \in$ {informative, uninformative}.

In order to minimize bias related to the sampling of data, the stratified 10-fold cross validation was used to estimate the performance of the model. In a 10-fold cross validation, the dataset is randomly split into 10 mutually exclusive subsets (DS_1, DS_2...DS_{10}) of approximately equal sizes and with proportional representation of the tweet classes. Using the data set, the classification model is trained and tested 10 times, with the 9-folds used as the training data set and the remaining 1-fold as the testing data set. The algorithms Naive Bayes and Support Vector Machine (SVM) were compared in terms of the different metrics of evaluation.

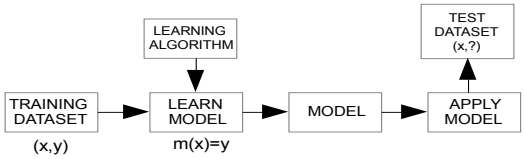

Fig. 1 Supervised Learning

3.4.2.2 *Machine Learning Algorithms*

Naive Bayes' and Support Vector Machine are two of the most commonly used machine learning algorithms for classification. Naive Bayes classifier is robust and has a good performance in several real-world classification tasks. A Naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions [25]. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [26].

Support Vector Machine is a learning method used for binary classification. The basic idea is to find a hyperplane which optimally separates the *d*-dimensional data into its two classes [26]. However, since example data is often not linearly separable, SVM incorporates the notion of a kernel induced feature space which projects the data into a higher dimensional space where the data is more easily separable [27].

3.4.3 *Evaluation of the Machine Learning Algorithms*

In this study, accuracy, recall, precision, area under curve (AUC) and F-measure were used as metrics in the empirical evaluation of the classification algorithms Naive Bayes and Support Vector Machine. Table I presents the description of each metric of evaluation, as described in Rapidminer.

TABLE I
METRICS OF EVALUATION

Metric	Description
Accuracy	Relative number of correctly classified examples or in other words percentage of correct predictions.
AUC	AUC is the Area Under the Curve of the Receiver Operating Characteristics (ROC) graph which is a technique for visualizing, organizing and selecting classifiers based on their performance.
Precision	Relative number of correctly as positive classified examples among all examples classified as positive
Recall	This parameter specifies the relative number of correctly as positive classified examples among all positive examples
F-measure	This parameter is a combination of the *precision* and the *recall* i.e. $f=2pr/(p+r)$ where f,r and p are *f-measure*, *recall* and *precision* respectively

In reference to the confusion matrix in Table II, the evaluation metrics can be defined as:

$$Accuracy = (TP + TN)/ (TP + TN + FP + FN)$$
$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$
$$F\text{-}measure = 2*Precision/(Precision + Recall)$$

TABLE II
CONFUSION MATRIX

		Prediction	
		Positive	Negative
Actual	Positive	TRUE POSITIVE(TP)	TRUE NEGATIVE(TN)
	Negative	FALSE POSITIVE(FP)	FALSE NEGATIVE(FN)

With 10-fold cross validation, the performance of the classification model is tested in each fold and the significant differences between the two machine learning algorithms were computed by means of statistical testing. The performance results data were tested for normality prior to the computation of the significant differences between the Naive Bayes and SVM. Field and Oztuna, Elhan and Tuccar state that it is impossible to draw accurate and reliable conclusions without taking into consideration normality [28][29]. Normality testing is used to determine if the data is normally distributed and consequently, whether to use parametric or non-parametric testing of significant difference [28][30][31][32]. If data is normally distributed, parametric tests are utilized else otherwise.

Fig. 2 presents the structure of the methodology of the study.

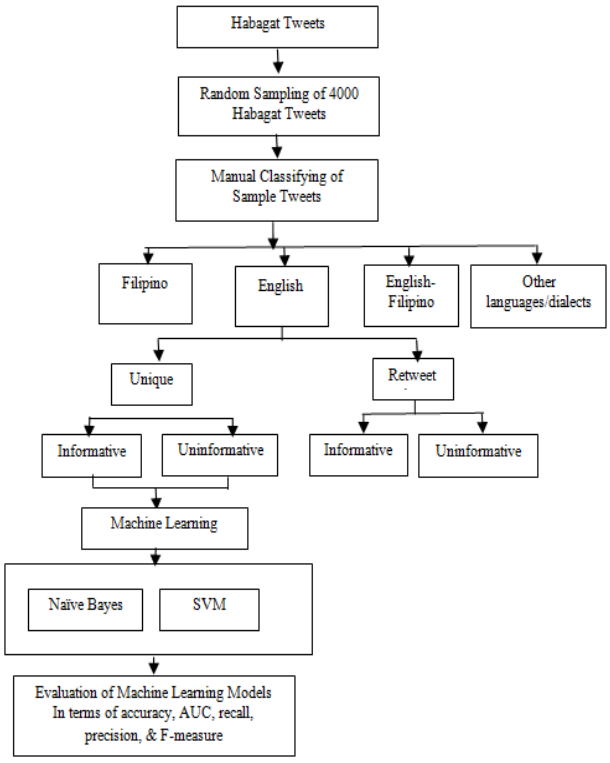

Fig. 2 Methodology Structure

IV. RESULTS AND DISCUSSION

4.1 *Manual Classification of Habagat Tweets*

From the 4000 tweets randomly selected, there were 1,563 English tweets, 1,393 Tagalog tweets, 913 tweets using a combination of English and Filipino and 121 tweets using other languages or dialects. Table III presents a summary of the manually classified English tweets.

TABLE III
MANUALLY CLASSIFIED ENGLISH TWEETS

	Informative Tweets	Uninformative Tweets	TOTAL
Unique	315	799	1114
Retweets	228	221	449
TOTAL	543	1020	1563

Based on the labeling of the annotators, the computed ICC or multi-rater Kappa coefficient is 0.671, which apparently is substantial [33][34][35] or there is a good level of agreement among the annotators in classifying whether a tweet is informative or not.

In case of conflict in label, a discussion among the three annotators was necessary to resolve such differences. After thorough discussion, annotators agreed on a specific label for the tweet.

4.2 *Extracted Information*

Applying conditional probability and Bayes' theorem, information were extracted from the statistics of the manually classified tweets. Based on these statistics, uninformative tweets outnumbered informative tweets by a ratio of 65% to 35%. An actual example of an uninformative tweet is "*Stay safe evryone!!! #PrayForThePhilippines #TrustGOD".*

The unique tweets are more likely uninformative (71.72%) and the unique tweets are more likely to be informative with the probability of 28.28%. It can also be noted that the probabilities of retweets being uninformative and informative are 49.22% and 50.78% respectively, which are relatively equal.

Though uninformative tweets tend to outnumber the informative tweets sent, the informative tweets are more likely to be retweeted (41.99%) than uninformative tweets (21.67%). Informative tweets that are retweeted imply the degree of significance and urgency of the situation, which then may provide information that may enhance the situational awareness of the public and disaster response units.

Previous studies on Twitter use during crises have shown that Twitter users often respond to information they deem important or relevant by retweeting other's tweets (RT @username) or by mentioning other usernames (@username) in their own tweets [36]. Retweets seem important because these tweets identify and thereby informally recommend the original author as an important source [37][38].

The results also suggest that the subscribers used Twitter to broadcast more of tweets that express their subjective messages and emotions regarding the Habagat event. These results seem to confirm the findings of Hughes and Palen on hurricanes [39], Starbird and Palen on flooding and wildfires [37] and Starbird and Palen on Haiti earthquake [40]. These studies revealed that users tweet to share information about the crisis, to express their opinions and feelings, and to help those in need of aid.

4.3 *Evaluation of Machine Learning Algorithms*

Table IV presents the results of the 10-fold cross validation for all folds for all the metrics of evaluation. Using the Kolomogorov- Smirnov and Shapiro Wilk for normality testing, the data is normally distributed and this is true to all the five evaluation metrics. The normality of these variables has also been validated by their Normal Probability Plots.

Since the data are normally distributed, parametric testing was performed. The parametric t-test was specifically used to determine the significant differences between Naive Bayes and SVM. Table V presents the results of the experimentation.

The paired t-test results shown in Table V demonstrate that there is a significant difference between Naive Bayes and SVM ($p<0.001$). This is true to all the five parameters namely, accuracy, AUC, precision, recall, and F-measure. In particular, SVM is significantly higher than Naive Bayes in accuracy, AUC, recall, and F-Measure, though Naive Bayes is significantly higher than SVM in precision. Table VI shows the mean values for the paired sample statistics.

The results can further be explained by referring to the

TABLE IV
RESULTS OF 10-FOLD VALIDATION

	ACCURACY		AUC		PRECISION		RECALL		F-MEASURE	
FOLD	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM
1	0.5946	0.8198	0.7280	0.9170	0.9696	0.8105	0.5063	0.8747	0.6400	0.8851
2	0.5536	0.7857	0.7200	0.8870	0.8571	0.7745	0.4500	0.9875	0.5902	0.8681
3	0.5856	0.7658	0.6300	0.8570	0.8113	0.7573	0.5443	0.9873	0.6515	0.8571
4	0.5536	0.8125	0.6420	0.9350	0.8000	0.7921	0.5000	1.0000	0.6154	0.8840
5	0.4955	0.8018	0.6730	0.8860	0.8519	0.8000	0.3924	0.9620	0.5254	0.8736
6	0.6216	0.7928	0.6650	0.8850	0.8846	0.7822	0.5450	0.9875	0.6866	0.8729
7	0.6339	0.8214	0.7360	0.9080	0.9024	0.8211	0.5679	0.9630	0.6917	0.8864
8	0.5766	0.8018	0.7720	0.8920	0.7333	0.8021	0.4625	0.9625	0.6160	0.8750
9	0.4643	0.7946	0.5310	0.8380	0.7667	0.7959	0.4074	0.9630	0.5238	0.8715
10	0.6036	0.8108	0.6660	0.8350	0.8462	0.8041	0.5500	0.9750	0.6667	0.8596
MEAN	0.5683	0.8007	0.6763	0.8840	0.8323	0.7940	0.4926	0.9663	0.6207	0.8733

TABLE V
PAIRED T-TEST RESULTS

					95% Confidence Interval of Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	T	df	Sig(2-tailed)
Pair 1	Accuracy (SVM-NB)	0.232410	0.05267610	0.0166577	0.194727	0.194727	13.952	9	<0.001
Pair 2	AUC (SVM-NB)	0.207000	0.05788700	0.0183055	0.166290	0.166290	11.346	9	<0.001
Pair 3	Precision (SVM-NB)	-0.038330	0.05348900	0.0169147	-0.077659	-0.077659	-2.266	9	<0.001
Pair 4	Recall (SVM-NB)	0.473670	0.06905060	0.0218357	0.424274	0.424274	21.692	9	<0.001
Pair 5	F-measure (SVM-NB)	0.2522600	0.06027520	0.0190607	0.209481	0.209481	13.252	9	<0.001

TABLE VI
PAIRED SAMPLES STATISTICS

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	NB_Accuracy	0.60172	10	0.5014710	0.0158579
	SVM_Accuracy	0.80207	10	0.0265764	0.0084042
Pair 2	NB_AUC	0.72350	10	0.0954443	0.0310821
	SVM_AUC	0.86880	10	0.0620767	0.0196304
Pair 3	NB_Precision	0.86546	10	0.0648431	0.0205052
	SVM_Precision	0.79652	10	0.0186738	0.0059052
Pair 4	NB_Recall	0.51856	10	0.0510700	0.0161498
	SVM_Recall	0.96747	10	0.0253584	0.0080190
Pair 5	NB_F-Measure	0.64702	10	0.0479798	0.0151726
	SVM_F-Measure	0.87349	10	0.0162222	0.0051299

confusion matrices of SVM and Naive Bayes as shown in Table VII and Table VIII respectively. Using the same training data set for both algorithms, the SVM model achieved an average accuracy of 80%, while Naive Bayes had 57% average accuracy. This indicates that the SVM model returned 892 correct classifications out of 1,114 unique tweets while Naive Bayes model correctly classified only 633 tweets.

In terms of recall, the SVM model correctly classified 780 uninformative tweets and only 19 labeled uninformative tweets as informative resulting to a recall value of 97.62% for the uninformative class. For Naive Bayes, the model correctly classified 396 uninformative tweets over 799 uninformative tweets yielding a 49.56% recall value.

AUC is a measure of quality of a probabilistic classifier. A random classifier has an area under curve 0.5, while a perfect classifier has 1.

Binary classifiers used in practice should therefore have an area somewhere in between, preferably close to 1 [41]. In this experiment, SVM demonstrated an average AUC of 0.884 which indicates that the classifier ranked positive examples higher than the negative examples.

F-measure is a combination of the metrics precision and recall (F-measure=*2*Precision/(Precision + Recall)*. Though Naive Bayes outperformed SVM in precision, SVM yielded a higher average F-measure (87.33%) than Naive Bayes (62.07%).

For the precision metric, Naive Bayes tend to have a higher precision than SVM. Naive Bayes returned a classification of

396 correct uninformative tweets and classified 78 labeled informative tweets as uninformative resulting to a precision of 84%. For the SVM model's precision of 79.4%, it displayed 780 correctly classified uninformative tweets but classified 203 labeled informative tweets as uninformative.

TABLE VII
CONFUSION MATRIX OF SVM

	Pred. INFORMATIVE	Pred. UNINFORMATIVE
True INFORMATIVE	112	203
True UNINFORMATIVE	19	780

TABLE VIII
CONFUSION MATRIX OF NAÏVE BAYES

	Pred. INFORMATIVE	Pred. UNINFORMATIVE
True INFORMATIVE	237	72
True UNINFORMATIVE	78	396

The results of this experiment confirm the findings of several studies that have concluded that SVM classifier significantly outperforms Naïve Bayes on the classification of short, unstructured and noisy text. Lu conducted an experiment to identify online messages using C4.5, Naive Bayes and SVM [42]. Based on experiment, SVM outperformed C4.5 and Naive Bayes in terms of accuracy and F-measure. Mostefai and Elberrichi compared SVM and Naive Bayes in sentiment analysis on Twitter by applying semantics and Wordnet [43]. SVM ranked first with an F-measure of 90.75%. Another study of Go et al used SVM, Naive Bayes and MaxEnt for sentiment analysis taking into consideration unigrams, bigrams and emoticons [44]. Experimental results of their study demonstrated SVM has outperformed the other classifiers.

V. CONCLUSION

Twitter is a medium used by subscribers to broadcast disaster-related tweets. During the Habagat incident, subscribers used this medium to broadcast both informative and uninformative tweets. Based on the Habagat statistics, uninformative tweets outnumbered informative tweets with a 65% to 35% ratio. Subscribers expressed their opinions and emotions through their posted tweets. Although more uninformative tweets were posted, the informative tweets were rapidly and repeatedly sent because these were retweeted. This evidently implies that the informative tweets contain vital and urgent information that can provide significantly needed information for situational awareness of the public and disaster response units.

Moreover, disaster-related tweets can be automatically classified using the bag of words approach and classifying algorithms SVM and Naïve Bayes. With a 10-fold cross validation, SVM outperformed Naïve Bayes in terms of accuracy, recall, AUC and F-measure, while Naïve Bayes performed better in precision.

Future directions of the research will be the exploration of other features and weights to generate a word vector and investigate their effects on the evaluation metrics. Feature selection, parameter optimization and semantics will be another focus of the research and the evaluation of other machine learning algorithms on the basis of other metrics of evaluation other than accuracy, recall, precision, AUC and F-measure. It is also essential to determine the priority of evaluation metrics which will guide data mining researchers to choose an algorithm for specific operations. Multi-label classification of English and multi-lingual tweets is also imperative to the extraction of relevant information which can eventually aid in increasing situational awareness. A real-time system that can detect and filter information from the disaster-relevant tweets may then be developed for an effective and efficient disaster response management.

ACKNOWLEDGMENT

This work is supported by the Commission on Higher Education of the Philippines under its Faculty Development Program and also under its Philippine Higher Education Research Network (PHERNet) program.

REFERENCES

[1] Vieweg, S. (2010). Microblogged Contributions to the Emergency arena: Discovery, Interpretation and Implications. Paper presented at the Computer Supported Collaborative Work.

[2] Palen, L., & Vieweg, S. (2008). The Emergence of Online Widescale Interaction in Unexpected Events: Assistance, Alliance & Retreat. Paper presented at the CSCW 2008.

[3] Castillo, C., Mendoza, M., & Poblete, B. (2011, 28 March - 1 April). Information Credibility on Twitter. Paper presented at the WWW 2011, Hyderabad, India.

[4] Yates, D., & Paquette, S. (2011). Emergency Knowledge Management and Social Media Technologies: A case study of the 2010 Haitian earthquake. International Journal of Information Management, 31(1), 6-13. http://dx.doi.org/10.1016/j.ijinfomgt.2010.10.001

[5] Palen, L., K. Anderson, G. Mark, J. Martin, D. Sicker, & D. , Grunwald. A Vision for Technology- Mediated Public Participation and Assistance in Mass Emergencies and Disasters, University of Colorado manuscript. 2010.

[6] Yin, J., Lampert, A., Cameron, M., Robinson, B., Power R. (2012). Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems. http://dx.doi.org/10.1109/MIS.2012.6

[7] Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. Paper presented at the CSCW 2010, Savannah, Georgia, USA.

[8] Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, L., Jansen, B., Yen, J.(2011).Classifying Text Messages for the Haiti Earthquake. Proceedings of the 8th International ISCRAM Conference Lisbon, Portugal.

[9] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.(2013). Extracting Information Nuggets from Disaster Related Messages in Social Media. Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013

[10] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.(2013). Practical Extraction of Disaster-Relevant Information from Social Media. International World Wide Web Conference Rio de Janeiro,

Brazil., May 13–17, 2013 ACM 978-1-4503-2038-2/13/05.
[11] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.(2010). Short Text Classification in Twitter to Improve Information Filtering. SIGIR 2010, Geneva, Switzerland.
[12] Mahendran, A., Duraiswamy, A., Reddy, A., Gonsalves, C.(2013). International Journal of Scientific Engineering and Technology Volume No.2, Issue No.6, pp : 589-594.
[13] Lee, K., Palsetia, D., Narayanan,R., Patwary, M., Agrawal, A. , Choudhary, A.(2011). Twitter Trending Topic Classification. 11th IEEE International Conference on Data Mining Workshops.
[14] Imran, M., Castillo, C. (2014). Volunteer-powered Automatic Classification of Social Media Messages for Public Health in AIDR . International World Wide Web Conference Seoul, Korea.
[15] Prasetyo, P., Lo, D., Achananuparp,P. , Tian, Y., Lim, E.(2012). Automatic Classi?cation of Software Related Microblogs. 28th IEEE International Conference on Software Maintenance (ICSM).
[16] Duwairi, R. , Qarqaz, I.(2014). Arabic Sentiment Analysis using Supervised Classification. The 1st International Workshop on Social Networks Analysis, Management and Security, Barcelona, Spain.
[17] Khamar, K.(2013). Short Text Classification Using kNN Based on Distance Function. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4.
[18] Lu, L., Shara, N. Reliability analysis: Calculate and Compare Intra-class Correlation Coefficients (ICC) in SAS. Retrieved Nov. 7, 2014 from http://www.lexjansen.com/nesug/nesug07/sa/sa13.pdf.
[19] Zielinski, A., Ulrich, B. (2012). Multilingual Analysis of Twitter News in Support of Mass Emergency Events. Proceedings of the 9th International ISCRAM Conference, Vancouver, Canada.
[20] Intraclass Correlation Coefficients(n.d.) retrieved Nov. 4, 2014 from http://www.unistat.com/guide/intraclass-correlation-coefficients/
[21] Lu, L., Shara, N.(n.d.) Reliability analysis: Calculate and Compare Intra-class Correlation Coefficients (ICC) in SAS. Retrieved Nov. 7, 2014 from http://www.lexjansen.com/nesug/nesug07/sa/sa13.pdf.
[22] Bayes' Theorem(n.d.) Retrieved Oct. 28, 2014 from http://www.eng.utah.edu/~cs5961/Resources/bayes.pdf
[23] Ramasubramanian, C. and Ramya,R. (2013). Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12.
[24] Zafarani, R., Abbasi, M., Liu, H.(2014). Social Media Mining: An Introduction. Cambridge University Press, 2014.
[25] Sankar, K., Kannan, S., Jennifer, P.(2014). Prediction of Code Fault Using Naive Bayes and SVM Classifiers. Middle-East Journal of Scientific Research IDOSI Publications, 2014.
[26] KumarGiri, R. and Saikia, M.(2013). Multipathrouting and for Admission Control and Load Balancing in Wireless Technology Meshnetworks. International Review on Computers.
[27] Kumaravel, A. and Rangarajan, K.(2013). Algorithm for Automation Specification for Exploring Dynamic Labyrinths, Indian Journal of Science and Technology.
[28] Field, A. (2009). Discovering statistics using SPSS. 3rd ed. London: SAGE publications Ltd; 2009 p.822.
[29] Oztuna D, Elhan AH, Tuccar E. Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. Turkish Journal of Medical Sciences.2006;36(3):171–6.
[30] Altman DG, Bland JM. Statistics Notes: The Normal Distribution.Bmj.1995;310(6975):298. http://dx.doi.org/10.1136/bmj.310.6975.298
[31] Pallant J.SPSS Survival Manual, A Step by Step Guide to Data Analysis using SPSS for Windows(2007).3 ed. Sydney: McGraw Hill; 2007. pp. 179–200.
[32]http://dx.doi.org/10.1136/emj.17.3.205-a%09Driscoll%20P,%20Lecky%20F.,%20Crosby%20M.(2000).%20An%20Introduction%20to%20Everyday%20Statistics-1.%20Emergency%20Medicine%20Journal.%20Retrieved%20November%208,%202014%20from%20http:/emj.bmj.com/content/17/3/205.2.full. http://dx.doi.org/10.1136/emj.17.3.205-a
[33] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data.Biometrics, 33, 1159-1174. http://dx.doi.org/10.2307/2529310
[34] Cicchetti, D. V. (1984). On a Model for Assessing the Aecurity of Infantile Attachment: Issues of Observer Reliability and Validity. Behavioral and Brain Sciences, 7, 149-150. http://dx.doi.org/10.1017/S0140525X00026558
[35] Fleiss, J. (1971). Measuring Nominal Scale Agreement among Many Raters. Psychological Bulletin, 76, 378-382. http://dx.doi.org/10.1037/h0031619
[36] Terpstra, T., de Vries, A., Stronkman, R., Paradies, R.(2012). Towards a Realtime Twitter Analysis during Crises for Operational Crisis Management. Proceedings of the 9th International ISCRAM Conference Vancouver, Canada.
[37] Starbird, K. and L. Palen (2010). Pass It On?: Retweeting in Mass Emergencies. Proceedings of the 7th International Information Systems for Crisis Response and Management Conference (ISCRAM), Seattle, WA.
[38] Vieweg, S., A. L. Hughes, K. Starbird and L. Palen (2010). Microblogging during Two Natural hazards Events: What Twitter may Contribute to Situational Awareness. Proceedings of the 28th Annual Conference on Computer Human Interaction Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA.
[39] Hughes, A. L. and L. Palen (2009). Twitter Adoption and Use in Mass Convergence and Emergency Events. Proceedings of the 6th International Information Systems for Crisis Response and Management Conference(ISCRAM) Conference, Gothenburg, Sweden.
[40] Starbird, K. and L. Palen (2011). "Voluntweeters:" Self-Organizing by Digital Volunteers in Times of Crisis. Proceedings of the ACM 2011 Conference on Computer Human Interaction (CHI 2011). Vancouver, BC, Canada, May 7–12, 2011.
[41] Vuk, M. and Curk, T. (2006). ROC Curve, Lift Chart and Calibration Plot. Metodoloski zvezki, Vol. 3, No. 1, 2006, 89-108. Retrieved November 26, 2014 from http://www.stat-d.si/mz/mz3.1/vuk.pdf.
[42] Lu, Y. (2013). Automatic Identification of Health Related Messages in Online Help Community Using Text Classification. Retrieved November 5, 2014 from http://www.springerplus.com/content/2/1/309.
[43] Mostefai,I. and Elberrichi, Z.(2014). Introducing The Semantics In Sentiment Analysis On Twitter Using WordNet. Retrieved November 5, 2014 from http://manifest.univouargla.dz/documents/Archive/Archive%20Faculte%20des%20Sciences%20et%20Technologies%20et%20des%20Sciences%20de%20le%20Matiere/2emes-journees-internationales-de-chimie-organometallique-et-catalyse-jicoc-2014/icaiit2014_submission_32.pdf
[44] Go,A., Richa, B. and Lei,H.(2009). Twitter Sentiment Classification using Distant Supervision. In Processing, 2009, pp.1–6. Retrieved November 3, 2014 from http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf