Classifying Text Messages for Emergency Response

Cornelia Caragea, Hyun-Woo Kim, Prasenjit Mitra, and John Yen
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA-16801
{ccaragea,hxk263,pmitra,jyen}@ist.psu.edu

Abstract

In case of emergencies (e.g., earthquakes, flooding), rapid responses are needed in order to address victims' requests for help. Hence, the ability to classify tweets and text messages *automatically*, together with the ability to deliver the relevant information to the appropriate personnel are essential for enabling the personnel to timely and efficiently work to address the most urgent needs, and to understand the emergency situation better. The choice of *features* used to encode tweets and text message data is crucial for the performance of the learning algorithms. Here, we present a comparative study of four types of feature representations used to enable learning classifiers from such data. These feature representations are obtained using a "bag of words" approach, feature abstraction, feature selection, and Latent Dirichlet Allocation (LDA). The results of our experiments on a real-world text message data set show that feature abstraction can yield better performing models than those obtained by using a "bag of words", feature selection and LDA.

1 Introduction

The 7.0 Earthquake in Haiti has mobilized the entire world to support the relief effort, especially through novel uses of the cyberspace. Relief workers, reporters, and non-governmental organizations (NGOs) have used tweets and text messages extensively to spread and share information about the needs, events, and causalities in the Twitterworld. Regular citizens have also employed Twitter to rally others to support relief efforts. Both Haitians and relief workers have used mobile phones to send text messages regarding damages, resource needs, and security-related events. While there is useful information in these tweets and text messages, they are not well-organized to allow critical information (e.g., water, medical supply, food) to be delivered to those who need them in a timely and efficient fashion. Hence, the ability to classify tweets and text messages *automatically*, together with the ability to deliver the relevant information to the appropriate personnel are essential for enabling the personnel to timely and efficiently work to address the most urgent needs, and to understand the emergency situation better in the Emergency Response Sector.

Although tweets and text message classification can be performed with little or no effort by people, it still remains difficult for computers. Machine learning currently offers a promising approach to the design of algorithms for training computer programs to efficiently and accurately classify short text message data. Some of the main challenges in classifying such data are as follows: (i) tweets and text messages contain only a few words and, sometimes, require background information for accurate classification. For example, the message *"I live in Leogane, Route de Mellier Bongnotte #72, I need formula for my baby."* requires knowledge that *formula* refers to *baby food*. The choice of features used to encode such data is crucial for the performance of the learning algorithms; (ii) tweets and text messages may belong to multiple categories, i.e., the *multi-label classification*; (iii) there may be possible errors in the manually generated labels (i.e., categories) of text messages, which can impact the performance of the learning algorithms; (iv) the training set is often limited in size.

In this study, we focused on the choice of features that are used to represent short text messages. We used four types of feature representations to enable learning Naïve Bayes and Support Vector Machine classifiers to accurately classify text messages from Haiti earthquake, submitted to Ushahidi-Haiti (http://haiti.ushahidi.com) through phone, e-mail, Twitter, or web. These feature representations are obtained using: (i) a "bag of words", i.e., all words in the vocabulary [11]; (ii) feature abstraction methods, that find a partition of the set of words in the vocabulary by clustering words based on the similarity between the class distributions that they induce [15]; (iii) feature selection methods, that select a subset of features based on some chosen criteria [8]; and (iv) Latent Dirichlet Allocation (LDA) [1], that finds hidden topics in the data. The topic words, i.e., the words in each topic, can be seen as a set of discriminative features.

We compared the performance of the trained classifiers using the above feature representations on a real-world text message data set from Ushahidi-Haiti. The results of our experiments show that feature abstraction generates features that can yield better performing classifiers than those obtained by using a "bag of words", features chosen by feature selection, and features as topic words output by LDA. We also discuss the insights gained from these results and suggest directions of future research to enhance the accuracy and the coverage for classifying tweets and text messages for improved efficiency and coordination during the response, transition, and recovery of extreme events.

2 Methods

In this section, we describe the three feature representation methods, which are compared with the "bag of words" approach: (1) feature abstraction; (2) feature selection; and (3) Latent Dirichlet Allocation (LDA).

Feature Abstraction. Feature abstraction methods are potentially successful techniques for producing appropriate features for classification [15]. They reduce the classifier input size by grouping "similar" features to generate *abstract features* (also called abstractions). Silvescu et al. [15] proposed an approach to simplifying the data representation used by a learner by grouping features based on the Jensen-Shannon divergence [3] that result in minimal reduction in the mutual information between features and the class variable.

Specifically, they used hierarchical agglomerative clustering to group the most "similar" features at each step of the algorithm, based on the similarity between the conditional distributions of the class variable given the features. The most "similar" features are identified as those that have the smallest Jensen-Shannon divergence between the conditional distributions of the class that the features induce. As an effect, abstract features that are predictive of the class variable are obtained. An example of an abstract feature can be "food", which is more general than the specific features "rice" and "formula" (i.e., baby food). The abstract feature is identified by the group $\{rice, formula\}$.

Silvescu et al. [15] have shown that *abstraction* reduces the model input size and helps improve the statistical estimates of complex models (especially when data are sparse) by reducing the number of parameters to be estimated from data. In this study, we have applied the feature abstraction approach of Silvescu et al. to generate the abstract feature representation.

Feature Selection. Feature selection methods attempt to remove redundant or irrelevant features in order to improve classification performance of learning algorithms [5]. Feature selection selects a subset of the available features based on some chosen criteria, and can substantially reduce the number of model parameters. Kira and Rendell [8] proposed an algorithm for feature selection, called Relief, which is not heuristic-based, is robust to noise and to interaction among features.

Relief is a weight-based algorithm. At each step, Relief samples from the training data an instance x, and determines x's *near-hit* (the closest instance from the same class as x) and *near-miss* (the closest instance from the opposite class of x) in the training data, by using p-dimensional Euclidian distance. A feature weight vector is updated for each such triplet to determine the relevance of all features to the class variable. The algorithm terminates after k steps and returns those features whose relevance level is above some user-specified threshold. We have used the Relief algorithm to select a subset of features that are predictive to the class variable.

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) is an unsupervised method for detecting hidden topics in the data proposed by Blei et al. [1]. LDA is a generative probabilistic model of a collection of documents, which has been successfully used to perform dimensionality reduction for text classification [20], where documents are multiple paragraphs and pages in length.

LDA [1] models each document in a collection as a mixture of topics (drawn from a conjugate Dirichlet prior), and each topic as a distribution over words in the vocabulary. The topic distribution of a document can be seen as a lower dimensional representation of the document (where the dimensionality is equal to the number of topics). Furthermore, the union of the words with high probability in each topic can be seen as a set of discriminative features for the collection of documents. We have used these words to generate the topic words feature representation.

3 Experiments and Results

Ushahidi Text Message Data Set. The data set used in our experiments is the Ushahidi data set (http://haiti.ushahidi.com/), which consists of 3598 text messages from Haiti earthquake. We used a subset of 2116 text messages of the Ushahidi data set, for which the English translation is available. While text messages are available in both Haitian Kreyol and English languages, we used only the English version, as Munro and Manning [13] found no significant improvement from one language to another on a similar task. The messages are classified into 10 categories: (1) *medical emergency*; (2) *people trapped*; (3) *food shortage*; (4) *water shortage*; (5) *water sanitation*; (6) *shelter needed*; (7) *collapsed structure*; (8) *food distribution*; (9) *hospital/clinic services*; (10) *person news*. Note that a message may belong to multiple categories. For example, the message *"Good evening ONG, I'm very happy for the aid you're giving to the people, I thank you. But in my zone that's to say Lamenten 54 Rue St Juste we need shelter and food."* belongs to both *shelter needed* and *food shortage* categories.

Experimental Design. Our experiments are designed to explore what feature representations of short text messages, which are provided as input to machine learning classifiers, result in best classification performance. We used four types of feature representations to enable learning Naïve Bayes and Support Vector Machines (SVM) classifiers on the Ushahidi text message data set:

- a bag of words representation, i.e., all words in the vocabulary. After stemming and removing stop words, and words with document frequency less than 3, the vocabulary size is 1525 (BoW) [11];
- a bag of m words chosen using the RELIEF feature selection method (FS) [8];
- a bag of m abstractions over all words in the vocabulary, i.e., an m-size partition of the vocabulary obtained by grouping words into m abstract terms based on the similarity between the class distributions that they induce (FA) [15];
- a bag of m topic words output by Latent Dirichlet Allocation (LDA) as the top 20 words from k topics (the number of topic words m is bounded by $20 \times k$) (TW) [1].

In our experiments, we used WEKA implementation [6] of Naïve Bayes Multinomial and SVM with the default parameters, and MALLET implementation [10] of LDA. The LDA parameters are set to default, except for the number of iterations of Gibbs sampling, which is set to $3,000$, and the random seed, which is set to 1. The number of topics k is set to 9 (chosen to be close to the number of categories in the data set). This results in $m = 165$ topic words. Hence, we trained classifiers for $m = 165$ for all of the above feature representations. In the case of feature abstraction, the 165-size partition of the vocabulary produces classifiers that use smaller number of "features" compared to the "bag of words" representation, i.e., 1525 words, and at the same time, the model compression is not very stringent so as to lose important information in the data through *abstraction*).

Because a text message may belong to one or more categories, we trained 10 "one vs. others" binary classifiers, one for each category. For all experiments, we report the average F1 Measure obtained in a 5-fold cross-validation experiment.

Results. Table 1 shows the comparison of average F1 Measure (along with 95% confidence intervals) using binary SVMs and Naïve Bayes trained on the Ushahidi text message data set for each of the ten categories. The feature representations used to train the classifiers are as follows: (i) "bag of words" (BoW); (ii) abstractions used as "features" in the classification model, which are obtained by feature abstraction (FA); (iii) features selected by Relief feature selection (FS); and (iv) topic words, output by LDA (TW).

Class	Support Vector Machines			
	BoW	FA	FS	TW
medical emergency	**0.29 ± 0.06**	0.27 ± 0.08	0.12 ± 0.07	0.11 ± 0.05
people trapped	0.68 ± 0.11	**0.74 ± 0.09**	0.64 ± 0.14	0.62 ± 0.23
food shortage	0.71 ± 0.02	**0.73 ± 0.03**	0.71 ± 0.06	0.72 ± 0.07
water shortage	0.66 ± 0.03	**0.67 ± 0.02**	0.63 ± 0.04	0.65 ± 0.03
water sanitation	0.91 ± 0.01	0.94 ± 0.01	**0.96 ± 0.01**	0.95 ± 0.01
shelter needed	**0.52 ± 0.02**	**0.52 ± 0.05**	0.44 ± 0.07	0.48 ± 0.04
collapsed structure	**0.42 ± 0.08**	0.33 ± 0.15	0.31 ± 0.16	0.39 ± 0.20
food distribution	**0.27 ± 0.05**	**0.27 ± 0.03**	0.18 ± 0.07	0.17 ± 0.09
hospital/clinic services	0.56 ± 0.04	**0.59 ± 0.06**	0.47 ± 0.08	0.51 ± 0.05
person news	0.55 ± 0.06	**0.59 ± 0.04**	0.39 ± 0.10	0.45 ± 0.04

Class	Naïve Bayes			
	BoW	FA	FS	TW
medical emergency	0.29 ± 0.09	**0.31 ± 0.06**	0.25 ± 0.11	0.14 ± 0.03
people trapped	0.67 ± 0.06	**0.71 ± 0.08**	**0.71 ± 0.13**	0.65 ± 0.09
food shortage	**0.77 ± 0.03**	0.76 ± 0.02	0.73 ± 0.02	0.75 ± 0.04
water shortage	**0.69 ± 0.02**	0.67 ± 0.02	0.66 ± 0.03	0.66 ± 0.02
water sanitation	0.94 ± 0.01	0.94 ± 0.01	**0.95 ± 0.01**	**0.95 ± 0.01**
shelter needed	0.45 ± 0.09	**0.46 ± 0.09**	0.38 ± 0.10	0.44 ± 0.03
collapsed structure	0.35 ± 0.10	**0.45 ± 0.12**	0.43 ± 0.14	0.41 ± 0.11
food distribution	0.26 ± 0.05	**0.27 ± 0.04**	0.20 ± 0.07	0.20 ± 0.09
hospital/clinic services	0.59 ± 0.05	**0.61 ± 0.04**	0.54 ± 0.08	0.55 ± 0.04
person news	**0.61 ± 0.05**	**0.61 ± 0.05**	0.48 ± 0.10	0.52 ± 0.06

Table 1: Comparison of average F1 Measure (with 95% confidence intervals) obtained in 5-fold cross-validation experiments using binary Support Vector Machines and Naïve Bayes classifiers trained on the Ushahidi text message data set for each of the ten classes. The feature representations used to train the classifiers are as follows: (i) "bag of words" (BoW); (ii) abstractions used as "features" in the classification model, which are obtained by feature abstraction (FA); (iii) features selected by Relief feature selection (FS); and (iv) topic words, output by LDA (TW).

As can be seen from the table, FA significantly outperforms BoW for most of the categories from the Ushahidi text message data set, using both Naïve Bayes and SVM classifiers. For few categories, for example *shelter needed*, FA-based SVM matches the performance of BoW-based SVM with substantially smaller number of features, i.e., 165 and 1525 features are used for training FA-based SVM and BoW-based SVM, respectively.

Compared to FS and TW, FA significantly outperforms both of them for the same number of features used in the classification model, for both SVM and Naïve Bayes classifiers, on all categories except *water sanitation*. Although topic models have been successfully applied to documents that are multiple paragraphs and pages in length, we found that they do not work very well when applied to short text messages.

It is interesting to note that the performance of SVM is worse than that of NB for many categories using any of the feature representations used in this study. This could be due to *overfitting* (see [19] for a theoretical analysis of overfitting for the SVM algorithm). However, as already noted, FA-based SVM significantly outperforms BoW-based SVM for many of the categories. This suggests that FA can help minimize *overfitting* (through parameter smoothing).

4 Related Work

The problem of learning classifiers from short text messages has started to receive significant attention in the machine learning literature. Healy et al. [7], Hidalgo et al. [4], and Cormack et al. [2] have previously addressed the problem of identifying *spam* short messages, by employing various machine learning algorithms (such as Naïve Bayes, SVM, Logistic Regression, and Decision Trees) and various feature representations (such as "bag of words", "bag of words" augmented by statistical features, e.g., the proportion of upper case letters or punctuation in the text, orthogonal sparse word bigrams, character bigrams and trigrams). Gupta and Ratinov [14] have employed transfer learning techniques to classify short online dialogs, by enriching the set of features using external data sources. Munro and Manning [13] have focused on classifying medical text messages, written in Chichewa language, that were received by a clinic in Malawi, and have shown that incorporating morphological and phonological variation could improve classification performance. Furthermore, Munro [12] has presented a brief survey about the crowdsourced translation to English of text messages written in Haitian Kreyol during the January 12 earthquake in Haiti. Collaborating online, people around the world were able to translate more than $40,000$ messages in a short time, which led to saving hundreds of lives, and direct the food and medical aid to tens of thousands [12]. Starbird and Stamberger [16] introduced a Twitter hashtag syntax for reporting events related to crisis.

Unlike these works, we focused on determining a subset of features that are most informative for the target variable, either by selecting a subset of features from the entire vocabulary using feature selection or LDA, or by constructing abstract features using feature abstraction. In addition, our text message classification task is harder due to its multi-label nature (i.e., text messages may belong to multiple categories).

The topics related to emergency response (ER) form an ontology that can be applied to emergency response for a wide range of relief operations. Ontology development tool such as Protégé has been widely used for developing ontology for different domains. Li et al. (2008) [9] proposed an ontology for emergency response (ER). The top level concepts of the proposed ontology include: aftermath-handling, emergency-rescue, emergency-response, and response-preparation. Each concept is further refined by a set of subconcepts. Emergency-rescue, for instance, include medical-aid, evacuation, and victim-assistance. Turoff et al. (2006) have designed a dynamic emergency response management system: DERMIS [18], and have identified the characteristics of a good ER system.

5 Summary and Discussion

Summary. In this study, we compared four types of feature representations for learning Naïve Bayes and SVM classifiers to *accurately* classify text messages about Haiti disaster relief (originating in Haiti and elsewhere) so that they can be more easily accessed by NGOs, other relief workers, people in Haiti, and their friends and families. These feature representations are: "bag of words", abstract features (or abstractions), features selected using feature selection, and topic words output by LDA.

The results of our experiments on the Ushahidi text message data set show that using *abstract features* makes it possible to construct predictive models that use significantly smaller number of features than those obtained using a bag of words representation. The resulting models are competitive with, and often significantly outperform those that use the "bag of words" feature representation. Moreover, *abstract features* yield better performing models than features selected by Relief feature selection, and than topic words extracted using LDA.

Discussion. In learning from real-world text message data, other challenges may be encountered, hence, making the learning problem harder. We point out some of these challenges and provide potential solutions that will be addressed in future work: (i) Tweets and text messages may belong to multiple categories. For example, the text message *"I live in the site Marassa 7. I ask some help like water and food thank you"* belongs to both categories *Food Shortage* and *Water Shortage*. Hence, the classification problem can be formulated as a *multi-label* problem [17], where a collection of $|C|$ binary classifiers is trained (where $|C|$ is the number of categories). A test instance is classified using all $|C|$ classifiers. In future work, we will use the *multi-label* formulation. However, in this study, we performed binary classification for each category in order to determine which are the most "difficult" categories to be classified. (ii) As with many real-world data, there may be possible errors in labeling text messages. For example, the text message *"We in Canada turjo quote, we need food, water and tents. count on your participation"* belongs to *Food distribution*. However, this example is very similar to *"Good evening ONG, I'm very happy for the aid you're giving to the people, I thank you. But in my zone that's to say Lamenten 54 Rue St Juste we need shelter and food."*, which belongs to *Food Shortage*. In future work, we plan to create a new category that will contain examples from both *Food distribution* and *Food Shortage*.

Furthermore, possible errors in labeling may occur due to the presence of general terms in a text message. For example, the text message *"We need help at Mahotiere 79. Since the catastrophe, we have not seen anyone from the government"* is labeled as *Food distribution* and *Water sanitation* in the Ushahidi data set. However, there is no indication of the type of help needed. For example, people at Mahotiere 79 may need medical assistance or shelter. To distribute this message to food and water departments may be very inefficient if the people have other more urgent needs. Instead, we propose to use a general category, which consists of these types of messages. Hence, the general department can *efficiently* determine what the people needs are and act accordingly. Further research may also include: (i) Exploration of other types of abstraction based on semantically related words; (ii) Classification of tweets about Haiti, provided by Twitter.

Acknowledgments

This research is partially supported by NSF RAPID grant IIS 1026763. We would like to acknowledge Ushahidi for making the data available for this research, NetHOPE for general interests about this research. We also wish to thank members of EMERSE research team for their contribution to related discussions: Lee Giles, Jim Jansen, Andrea Tapia, Dinghao Wu, Anuj Jaiswal, Gregory Traylor, Anthony Maslowski, Nathan McNeese, and Louis-Marie Ngamassi Tchouakeu.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] G. V. Cormack, J. M. Gómez Hidalgo, and E. P. Sánz. Spam filtering for short messages. In *Proceedings of the 16th ACM CIKM '07*, pages 313–320, 2007.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[4] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García. Content based sms spam filtering. In *DocEng '06: Proceedings of the 2006 ACM symposium on Document engineering*, pages 107–114, 2006.

[5] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[7] M. Healy, S. J. Delany, and A. Zamolotskikh. An assessment of case base reasoning for short text message classification, 2005.

[8] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of AAAI*, pages 122–134, San Jose, CA, 1992. Morgan Kaufmann Publishers Inc.

[9] X. Li, G. Liu, A. Ling, J. Zhan, N. An, L. Li, and Y. Sha. Building a practical ontology for emergency response systems. In *Proceedings of IEEE International Conference on CSSE '08*, pages 222–225, 2008.

[10] A. McCallum. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*, 2002.

[11] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification, 1998.

[12] R. Munro. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *Relief 2.0 in Haiti*, 2010.

[13] R. Munro and C. D. Manning. Subword variation in text message classification. In *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*, pages 510–518, 2010.

[14] R.Gupta and L. Ratinov. Text categorization with knowledge transfer from heterogeneous data sources. In *Proceedings of AAAI 2008*, 2008.

[15] A. Silvescu, C. Caragea, and V. Honavar. Combining super-structuring and abstraction on sequence classification. In *ICDM*, pages 986–991, 2009.

[16] K. Starbird and J. Stamberger. Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the 7th Intl. ISCRAM Conf. '10*, 2010.

[17] G. Tsoumakas and I. Katakis. Multi label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[18] M. Turoff, M. Chumer, B. Van de Walle, and X. Yao. Design of a dynamic emergency response management information system (dermis). *Journal of Information Technology Theory and Application*, 2004.

[19] V. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, N.Y., 1998.

[20] X. Wei and W.B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of ACM SIGIR*, pages 178–185, 2009.