Modeling Microtext with Higher Order Learning

Christie Nelson
RUTCOR, Rutgers University
cnelson@dimacs.rutgers.edu

Hannah Keiler
Statistics, Columbia University
hpk2108@columbia.edu

William M. Pottenger
DIMACS & RUTCOR, Rutgers University
drwmp@cs.rutgers.edu

Abstract

Processing data manually is especially problematic during a natural disaster, where aid and response are quickly and urgently needed. In real time scenarios, a difficult yet important problem is to be able to get an accurate picture of needs from streaming data in a short time. When the streaming data includes microtext, this problem becomes even more challenging. In the application of emergency response, modeling microtext in real-time is especially important. Once messages have been classified and/or topics learned, the predicted categories and/or topics can be used by emergency responders to rapidly respond to needs.

In this effort, microtext from social media and text messages during the 2010 Haitian earthquake were modeled using novel machine learning algorithms: Higher-Order Naïve Bayes (HONB) and Higher-Order Latent Dirichlet Allocation (HO-LDA). Both illustrate that Higher-Order Learning can be valuable in classifying text data. Higher-Order Learning improves model generalization in online or real-time scenarios when smaller amounts of data are available for learning. Results from this research are promising in that when using samples of training data, the HONB classifier statistically significantly outperformed Naïve Bayes in all trials based on the accuracy metric. Promising results were also obtained in the comparison of HO-LDA versus traditional Latent Dirichlet Allocation.

Introduction

In real time scenarios, such as emergency response for disaster events, an accurate picture of the situation at hand is needed quickly. This can be obtained from streaming data, such as from text messages and social media sources. This type of data tends to be in the form of microtext, short pieces of text, so the problem of modeling the data becomes even more challenging.

In a real-time situation like responding to a natural disaster, the ability to build models on small amounts of data is critical since in general less data is available. Higher-Order Learning techniques are invaluable in this type of situation, as leveraging relational information in model construction results in models that generalize better on smaller amounts of data.

Here, various learning approaches involving Higher-Order Learning will be discussed. In this research, Naïve Bayes and its Higher-Order counterpart, Higher-Order Naïve Bayes (HONB), are used to classify text data from the 2010 Haitian earthquake. A second approach involves modeling the data using topic modeling approaches such as Latent Dirichlet Allocation and Higher-Order Latent Dirichlet Allocation (HO-LDA). Once messages have been classified and/or topics learned, the predicted categories and/or topics can be used by emergency responders to more rapidly respond to emerging needs.

This paper is organized as follows. In the following section, Background and Related Work are presented. Next, the Approach and then Results are presented. Conclusions and Future Work are discussed in closing.

Background and Related Work

Haitian Earthquake

On January 12, 2010, Haiti was hit with a devastating earthquake of magnitude 7.0. It was the worst to strike Haiti in 200 years and resulted in over 230,000 deaths and an additional 300,000 injured people (Heinzelman and Waters 2010). Despite the fact that many organizations arrived in the wake of the earthquake offering aid and supplies, there were many deaths after the earthquake due to unsanitary living conditions and unclean water. Emergency responders using traditional methods of disaster response had difficulty collectively prioritizing and physically locating people and areas of need.

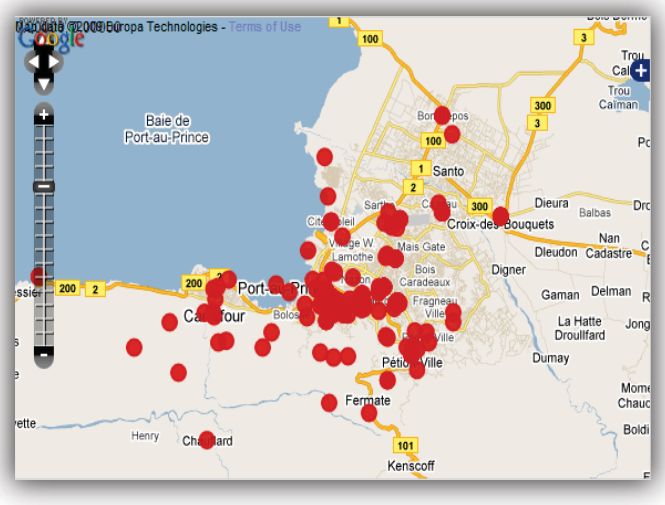
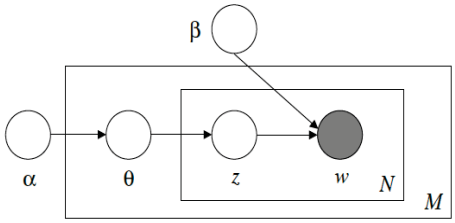


Figure 1. Ushahidi map of needs concentrated aroundPort-au-Prince



Choose θ ~ Dirichlet(α)
For each of the N words w_n:
Choose a topic z_n ~ Multinomial(θ)
Choose a word w_n from $p(w_n \mid z_n; \beta)$, a multinomial probability conditioned on the topic z_n and parameterized by the topic distributions β

Figure 2. The generation of a corpus of documents using the Latent Dirichlet Allocation Model

In an effort to combat the chaos that ensued from the earthquake, an organization called Ushahidi gathered information from social media sources such as blogs, Twitter, and Facebook, as well as text messages. This valuable information was then used to illustrate areas with the greatest need on a visual map, which was publicly available. An example of a map made by Ushahidi is shown in Figure 1. Ushahidi began gathering information and reports within two hours of the earthquake. Ushahidi volunteers found that they had the most difficulty in verifying and triaging the massive amount of information that came in during the early days of the disaster. Manual translation and triaging were performed, which was time consuming at a period when speed was vital.

The research goal of this work is to examine and improve the modeling of social media emergency text. The hope is that in the future there will be a way to automate categorization in an emergency. This will improve the response time within which emergency needs can be met.

Topic Modeling

A topic model is a statistical model for discovering the unobserved topics which explain why individual text documents within a collection of documents (called a corpus) are similar. This research focused on a type of topic modeling first proposed by Blei, Ng, and Jordan 2003 named LDA (Figure 2). An important assumption is that each document is a "bag of words," which means that the order of the words does not matter. Given a collection of documents, the posterior distribution of the latent variables (the underlying topics and the distribution with which each document exhibits them) given the words determines the decomposition of the underlying topics of the collection. In this application, the observed data were the words of each social media message, and the individual social media messages were treated as documents.

The underlying topics, their structure, and their distribution within a document can be learned by using posterior inference. In this project, Gibbs Sampling was used, which is a standard algorithm. In order to perform inference using a Gibbs Sampling algorithm, the conditional probability of occurrence of a topic for a given word in the corpus is used:

$$P(z_i = j \mid \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \quad (1)$$

The underlying distribution that is assumed to generate the corpus is parameterized by a Dirichlet parameter α. The number of topics is assumed to be some *T*. Suppose *W* is the size of the dictionary of words. α is a vector of length *T* and β is a *T* x *W* matrix where each row of the matrix is a Dirichlet parameter vector of length *W*. This formula is used in sampling the topic for the term *w* at position *i*. The term $n_{-i,j}^{(w_i)}$ corresponds to the number of occurrences of the term *w* that are assigned to the topic *j*, not including the current (i^{th}) occurrence and $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic *j*, not including the current one. d_i corresponds to the i^{th} document. Basically, a word is assigned to a topic with probability proportional to its frequency of occurrence in that topic.

The computational complexity of LDA is O(NKV) for N documents, V number of words in the vocabulary, and K number of topics (Blei 2008). Posterior multinomals do not need computed for each instance of each term in a document. Rather, they can be computed just once per unique term in a document.

Higher-Order Learning

Traditional machine learning techniques make assumptions that attributes, or words in our case, are IID (independent and identically distributed). These traditional methods can be thought of as "zero-order" since they do not leverage relationships between attributes across instances. The traditional IID assumption does not permit traditional machine learning methods to leverage "higher-order" relationships (Ganiz, George, and Pottenger 2011). Higher-Order Learning techniques can be useful in this area, as they do not assume that attributes are IID. Instead, Higher-Order methods take advantage of the latent information in higher-order paths between attributes. These paths leverage relationships between attribute values regardless of instance boundaries.

Higher-Order Learning has been shown to work well in several applications in the statistical relational learning field. Sometimes there is only a small amount of labeled training data available. When approaching this problem using traditional machine learning methods, traditional algorithms often do not perform very well. An underlying issue is the previously discussed assumption of IID attributes. Traditional machine learning methods do not leverage the relationships between attributes across instances. This is where Higher-Order Learning techniques can be very useful. Therefore, Higher-Order Learning techniques often outperform traditional machine learning techniques, providing much better results especially when data is sparse (Nelson et al. 2012).

In related prior work done by Ganiz, Lytkin, and Pottenger 2009, higher-order paths were successfully leveraged in entity classification using HONB, a generative learner, and Higher-Order Support Vector Machines (HOSVM), a discriminative learner. Table 1 portrays a sampling of the results, which demonstrate the value of the Higher-Order learning framework. Illustrated in the results of Table 1 are mean classification accuracies (reproduced from Ganiz, Lytkin, and Pottenger 2009) obtained on the *Science* subset of the 20 Newsgroups dataset. This dataset contained four classes; 5% (25 documents per class) of the data was used for training, the remaining 95% (475 documents per class) was used for testing. Performance of every pair of classifiers was significantly different at 95% confidence level. As Table 1 portrays, HONB statistically significantly outperformed both Naïve Bayes and SVM, despite the latter approach (SVM) being well-known to perform well on text classification tasks. Likewise, HOSVM outperformed Naïve Bayes and SVM.

	NB	HONB	SVM	HOSVM
Mean	0.632	0.833	0.751	0.792
Stdev.	0.071	0.043	0.029	0.039

Table 1: Mean Classification Accuracies obtained on the *Science* subset of the 20 Newsgroups dataset (Ganiz, Lytkin, and Pottenger 2009)

As detailed in (Ganiz, Lytkin, and Pottenger 2009), these results from Table 1 are representative in general of the performance of algorithms that leverage Higher-Order Learning techniques.

Prior Work – Nuclear Detection Using HONB

Prior research was performed on a nuclear detection dataset using HONB by Nelson and Pottenger 2011. The detection of potentially threatening nuclear materials is a challenging homeland security problem. This research involved the application of a novel statistical relational learning algorithm, HONB, to improve the detection and identification of nuclear isotopes. When classifying nuclear detection data, distinguishing potentially threatening from harmless radioisotopes is critical. This research applied Higher-Order Learning to nuclear detection data to improve the detection and identification of four isotopes: Ga67, I131, In111, and Tc99m.

In this nuclear detection research, traditional IID machine learning methods were applied, and the results were compared with the performance of leveraging Higher-Order dependencies between feature values using HONB. These findings gave insight on the performance of higher-order classifiers on datasets with small positive class size. In this research, Naïve Bayes was compared with its Higher-Order counterpart, HONB. HONB was found to perform statistically significantly better for isotope Ga67 when using a pre-processing methodology of discretizing then binarizing the input sensor data. Similar results were seen for various samples of training data for I131, In111, and Tc99m. HONB was also found to perform statistically significantly better for isotopes I131 and Tc99m when the pre-processing involved normalization, discretization then binarization. This study showed that Higher-Order Learning techniques can be very useful in the arena of nuclear detection.

Prior Work – Nuclear Detection Using HO-LDA

Prior work was performed on the same nuclear detection dataset by Nelson et al. 2012, only this time using a novel topic modeling approach, HO-LDA. In total, seventeen different nuclear radioisotopes were modeled, and the performance of Higher-Order versus traditional techniques was then evaluated.

This project employed LDA and HO-LDA on a nuclear detection numeric dataset to obtain a topic decomposition for each instance. For evaluation purposes these learned topics were then used as features in a traditional supervised classification algorithm. In essence, the LDA or HO-LDA topic assignments were used as features in supervised learning algorithms that predicted the class (isotope), treating LDA or HO-LDA as a feature space transform. Results demonstrated further evidence that Higher-Order Learning techniques can be usefully applied in topic modeling applied to nuclear detection.

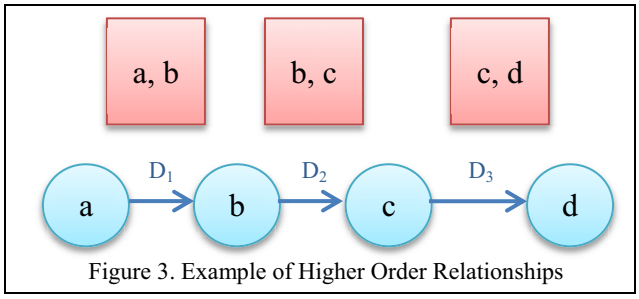


Figure 3. Example of Higher Order Relationships

Problem Definition

There are several objectives of this research. The overarching goal is to improve machine learning of microtext, a challenging research area. The second goal of this research is to illustrate that methods based on Higher Order Learning outperform traditional techniques. In particular, the goal is to show that topic modeling with HO-LDA outperforms LDA on our emergency response social media dataset, and similarly that HONB outperforms traditional Naïve Bayes on this data. However, since one is unsupervised and the other not, we will not be directly comparing HO-LDA with HONB.

Approach

Higher-Order Naïve Bayes

Naïve Bayes is a traditional classifier, based on Bayes Rule. Bayes Rule states that for some events A and B:

$$P(A|B) = P(B|A)P(A) / P(B) \quad (2)$$

Naïve Bayes assumes strong "naïve" independence, and also assumes that the absence (or presence) of a particular attribute is unrelated to the absence (or presence) of any other attribute. Assuming independence makes a good estimation challenging when considering real world applications. In practice, sometimes there is only a small amount of training data to use. In these cases, it can help to use other more sophisticated machine learning techniques that do not make the IID assumption.

As shown in Figure 3, HONB utilizes relationships between attribute values across instances. In Figure 3, there are three sample instances shown, instances D_1, D_2, and D_3. Instance D_1 has two attributes (a and b), instance D_2 has two attributes (b and c), and instance D_3 has two attributes (c and d). Traditional Naïve Bayes does not leverage the latent Higher-Order paths. However, HONB uses these Higher-Order paths to create a link between attributes. In this example, attributes a and d are linked by leveraging the higher-order paths between attributes in D_1, D_2, and D_3.

Higher-Order Latent Dirichlet Allocation

One potential shortcoming of LDA is that it only considers relationships between observations within individual data instances (documents) while disregarding the dependencies that link observations across documents. It assumes the instances are IID. In other words, LDA looks at "zero-order" relations. To address this, Higher-Order LDA was developed (Nelson et al. 2012). This novel approach to topic modeling modifies the Gibbs-sampling formula of LDA given in (1) by replacing feature frequencies in topics with their Higher-Order path counts (see Figure 3). In other words, in Equation 1, these counts were replaced with higher path counts for the feature *w* in topic *j*. Results demonstrated further evidence that Higher-Order Learning techniques can be usefully applied in topic modeling applied to nuclear detection.

$$P(z_i = j \,|\, \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \quad (3)$$

The Data

The data came from an organization called Ushahidi, which was created during the 2007-2008 post-election violence in Kenya. Ushahidi was designed as a way to report incidents and provide up-to-date information about the violence. It is "an open source crisis mapping platform" (Chaturvedi, Swami, and Singh 2011) that relies on crowdsourcing information from text messages and social media outlets to report and illustrate locations of need. Responders and the general public can use this to see where needs are concentrated and to find information not yet reported in the news.

Ushahidi was deployed within two hours of the 2010 Haitian earthquake. Volunteers quickly realized that much more manpower was needed to be able to handle the massive volume of information and messages coming in. Within four days of the earthquake, a collaborative effort between Tufts University, FrontlineSMS, the United States State Department, and Digicel set up a system that allowed Haitians to send text messages to an emergency number, and then volunteers manually classified the needs and then mapped them using Ushahidi. Approximately 85% of Haitian households had access to mobile phones after the earthquake, and many of the phone towers had been repaired by the time it became operational, aiding in the success of the Ushahidi Haitian deployment.

The Ushahidi Haitian data includes 3,598 texts from social media sources and text messages sent in the wake of the 2010 Haitian Earthquake. 3,561 of these messages are in English or were translated into English. The rest are in French or Haitian Creole and were not used in this research. Messages include medical needs, texts for help from survivors in the rubble, updates about infrastructure, and messages stating that supplies are available or are needed.

As messages could contain requests for multiple types of help needed, various labeling schemes were evaluated to deal with messages requiring multiple messages. For

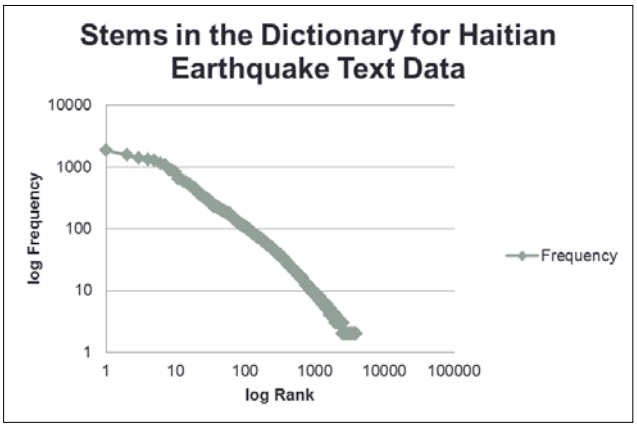


Figure 4. log-log plot of rank vs. frequency for the dictionary of stems made from the Haitian text data

example, one of the texts in the dataset "THERES A BRIDGE IN DELMAS 31 THAT IS DAMANGED, THERES NO FOOD AND NO GAS YET" reports both a damaged structure and a need for food and gas. The original Ushahidi data has 36 categories. However, after examining these messages, it became clear that in the haste of the emergency, many of the text messages were incorrectly labeled (i.e., a message of someone looking for information about a family member was labeled as "Services Available"). Initial trials were performed using the 36 categories created by Ushahidi, but classification accuracy was low. In addition, the original 36 categories were not rigorously defined, so a new labeling scheme was created, and the messages were entirely re-labeled to ensure correctness of the labeling scheme for experiments. 19 clearly defined categories were used for the scope of this research. Each message was labeled with all relevant categories.

As the classifiers employed in this project used only one label per message, three labeling schemes were examined to determine which performed best in early trials, which will be discussed later in the paper. First, the importance of the labels was priority ranked by most important to least important. For example, responding to violence would require more urgent attention than delivering clothing. Then a message was first labeled by the "most important" relevant class ("Rank 1"), with an inter-rater reliability of 92.7%, and a "second most important" relevant class ("Rank 2"), with an inter-rater reliability of 38.7%. A third labeling method was to qualitatively determine which label was the "Most Relevant." This was done by using the need that was emphasized or stated the most times within the message (having nothing to do with importance). Different trials were performed with all three different labels. Ultimately, the "Rank 2" labeling schema slightly outperformed the other two labeling approaches with respect to the metrics accuracy, precision, recall, and F-measure. The reason "Rank 2" performed the best requires further investigation, but performance was only slightly better. In fact, labeling was a challenging part of this research: reading and correctly labeling 3,561 messages was time-consuming, again illustrating the importance of using machine learning in the arena of text classification.

Two Approaches to Preprocessing

In order to prepare the text data to be used in the classification algorithms, each word in each message had to be represented as a draw from a multinomial distribution, which is the input that LDA and HO-LDA require. Two approaches were initially evaluated, and the same methods were used for both HONB and HO-LDA experiments.

To convert the text data for input into the various learners, a dictionary of words was created. This dictionary was created by first stemming the words, and then using all of the stems to create a dictionary. In particular, to determine the dictionary stems, common words (and, the, or, etc.), punctuation, and words occurring no more than once in the entire dataset were removed. The "common words" were those in the NLTK stopwords (www.nltk.org). The remaining words were then stemmed using the Porter Stemmer (Porter 1980) to create a dictionary of stems. Next, the remaining words appearing in each of the messages had to be given numeric values corresponding to the appropriate stem(s) and the dictionary index location of the stem(s).

The first approach for converting the text messages for input, Multiple Stem feature creation, used all stems that corresponded to a word. This means that one word may have multiple index values. For example, if the words "child" and "children" appeared separately in messages, then both the stems "child" and "childr" were in the dictionary of stems after the Porter Stemmer removed suffixes. Suppose "child" has index 20 and "childr" index 332. As a result, when using this method, the word "children" was represented by the two numbers of the index in the dictionary, 20 and 332.

The second method, called Longest Stem feature creation, used only the longest stem corresponding to each word. In the same example, the word "children" was represented only by one index value - the index number corresponding to the stem "childr" which was 332.

Our dictionary follows Zipf's Law, which states that in a natural language corpus, the frequencies of occurrences of words are inversely proportional to their rank (Zipf 1932), illustrated by the linearity of the log-log plot in Figure 4.

Pre-processing for Higher-Order Naïve Bayes

For the HONB trials, only the Longest Stem feature creation method combined with the Rank 2 labeling scheme were used for the trials. Trials were performed for training sample sizes 20% through 65% at 5% intervals.

Samples were selected in a randomized, stratified manner. The accuracy metric was examined.

Approach for Higher-Order Latent Dirichlet Allocation

After the initial preprocessing was performed on the text data, as discussed above, initial trials were performed examining both preprocessing methods (Multiple Stem and Longest Stem feature creation), as well as the three labeling approaches (Rank 1, Rank 2, and Most Relevant). After examining early results, Longest Stem was the feature creation selected with the Rank 2 labeling scheme. Results presented in this paper reflect these choices.

Trials were first performed for the entire training set, a sample size of 100% of the data. The preprocessed data was modeled using LDA and HO-LDA for 5, 10, and 20 topic models. Then, the topic probability assignments were used as attribute values, and the instances were classified using the decision tree induction algorithm J48 in the WEKA Workbench. 30 trials were performed for each experiment (e.g., 30 for the LDA 5 topic model, 30 for the HO-LDA 5 topic model, etc.) using standard 10 fold cross validation. Standard metrics of accuracy, precision, recall, and F-measure were recorded.

Sampling for the Haitian text data was also examined. For each training sample size (15%, 25%, and 50%), 30 randomized, stratified datasets were created. The remaining data that was not used in the training sample was used for testing. The data was again classified using the same experimental design. Standard metrics of accuracy, precision, recall, and F-measure were recorded.

Results

Higher-Order Naïve Bayes Results

In this section, results from the Longest Stems feature creation approach with the Rank 2 labeling scheme are presented. See Table 5 for complete results. In all cases, HONB statistically significantly outperformed Naïve Bayes. HONB accuracy results ranged from 39% to 47%, and Naïve Bayes accuracy results ranged from 27% to 45%, illustrating the difficulty of this classification application on microtext.

Higher-Order Latent Dirichlet Allocation Results

When using the Longest Stem feature creation approach with Rank 2 labels, HO-LDA consistently performed as well or statistically significantly better than LDA across all training set sample sizes of 15%, 25%, 50% and 100% (even though performance was not much better). These results indicate that leveraging Higher-Order Learning for topic modeling results in statistically significant improvements over the standard zero-order approach in many cases. Results are illustrated in Tables 2-4.

For a sample size of 100%, for all of the metrics: accuracy, F-measure, precision, and recall, HO-LDA statistically significantly (α=.05) outperformed LDA for 5 and 10 topics. For 20 topics, HO-LDA and LDA performed the same.

For a sample size of 15%, across all of the metrics, HO-LDA outperformed LDA statistically significantly for 10 and 20 topics. For 5 topics, there was no significant difference between HO-LDA and LDA.

Next, for training sample size of 25%, across all of the metrics, HO-LDA outperformed LDA statistically significantly for 20 topics. For 5 topics and 10 topics, there was no significant difference.

Finally, for training sample size of 50%, HO-LDA statistically significantly outperformed LDA for 5 topics with the precision metric. In other cases, there was no statistically significant difference. These results are consistent with prior research in Higher-Order Learning that reveal significant performance gains especially when the available training data is small, a situation that arises in online or real-time learning scenarios.

Conclusion and Future Work

In this effort we have demonstrated the value of applying Higher Order Learning techniques to the challenging problem of modeling microtext data for emergency response. Both Higher-Order Learning algorithms evaluated, HONB and HO-LDA, performed comparably or statistically significantly better than their zero-order counterparts. This is an important result given the need to model streaming data in real-time during emergency response.

There are many possibilities for future work in this interesting application domain. One immediate future work item is to incorporate data from the 2010 Chile earthquake, as well as the 2011 Japan earthquake. These earthquake datasets may be obtained from Ushahidi. The use of additional training data may improve classification accuracy, another important goal for future work.

Another aspect of future work is the use of the topics learned by topic modeling algorithms like HO-LDA as input to resource allocation frameworks. Such frameworks can benefit emergency response by allocating urgently needed resources based on real-time modeling of resource needs. This idea is currently being explored further.

A final component for future work is to look at an open problem in the fields of clustering and topic modeling: determining the optimal number of topics for LDA or HO-LDA. In topic modeling, hierarchical topic models have been proposed in (Blei et al. 2004) as a way to avoid having to select the number of topics as an input parameter prior to creating the model. The Chinese Restaurant Process (CRP) has been proposed as a basic way to think about a prior distribution for the number of topics. The idea of CRP is that there is a Chinese Restaurant with an

infinite number of tables. A person enters a restaurant and sits at an occupied table with probability proportional to the number of people already there or sits at a new table. The tables represent the topics, the people are the documents, and the number of tables is the number of topics. In (Teh et al. 2006), an analog of the CRP called the Chinese Restaurant Franchise (CRF) was developed for use with LDA as a prior for the number of topics. The schema for CRF is that there is an infinite number of restaurants in a franchise and a global menu of dishes. The first customer who sits at a table orders a dish, which is shared among the subsequent members arriving at the table. The dish corresponds to the topic, and the restaurant to the document, so topics can be shared across documents. In future endeavors, we hope to incorporate a version of CRF to HO-LDA.

Acknowledgements

This research was supported in part by the U.S. Department of Homeland Security (DHS) Center of Excellence for Advanced Data Analysis as well as by the National Science Foundation under Grant No. 1018445. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or DHS. The authors are grateful for the help of co-workers, family members, and friends. Co-author W. M. Pottenger also gratefully acknowledges the continuing help of his Lord and Savior, Yeshua the Messiah (Jesus the Christ) in his life and work.

Table 2. T-Test Results. HO-LDA vs. LDA using J48, 5 Topics

	HO-LDA Avg.	HO-LDA Standard Deviation	p Value	LDA Avg.	LDA Standard Deviation
Accuracy - 100%	0.2474	0.0055	0	0.2392	0.0064
F-Measure-100%	0.2386	0.005	0	0.2305	0.006
Precision - 100%	0.2329	0.0051	0	0.2254	0.0059
Recall - 100%	0.2473	0.0054	0	0.2392	0.0064
Accuracy - 50%	0.2384	0.0086	0.2258	0.2365	0.0107
F-Measure - 50%	0.2295	0.0086	0.0552	0.2248	0.0133
Precision - 50%	0.2238	0.0089	0.0427	0.2192	0.0113
Recall - 50%	0.2384	0.0086	0.2391	0.2366	0.0108
Accuracy - 25%	0.225	0.0112	0.1452	0.2222	0.009
F-Measure - 25%	0.2161	0.0105	0.0789	0.2126	0.0083
Precision - 25%	0.2105	0.0101	0.0539	0.2066	0.0083
Recall - 25%	0.2249	0.0112	0.1529	0.2222	0.0089
Accuracy - 15%	0.1993	0.0144	0.4373	0.1989	0.0078
F-Measure - 15%	0.1909	0.0102	0.1317	0.1882	0.0082
Precision - 15%	0.1856	0.0096	0.0515	0.1817	0.0086
Recall - 15%	0.1994	0.0114	0.4376	0.199	0.0079

Table 3. T-Test Results. HO-LDA vs. LDA using J48, 10 Topics

	HO-LDA Avg.	HO-LDA Standard Deviation	p Value	LDA Avg.	LDA Standard Deviation
Accuracy - 100%	0.2474	0.0055	0	0.2392	0.0064
F-Measure-100%	0.2523	0.0061	0.0005	0.2471	0.0055
Precision - 100%	0.2496	0.0061	0.0018	0.245	0.0056
Recall - 100%	0.2565	0.0062	0.0003	0.2509	0.0057
Accuracy - 50%	0.2379	0.0104	0.4669	0.2381	0.008
F-Measure - 50%	0.2336	0.01	0.4006	0.233	0.0083
Precision - 50%	0.2307	0.0098	0.3031	0.23	0.008
Recall - 50%	0.2379	0.0105	0.4211	0.2374	0.0088
Accuracy - 25%	0.2086	0.0107	0.4099	0.2079	0.0129
F-Measure - 25%	0.2042	0.0104	0.4077	0.2035	0.0126
Precision - 25%	0.2012	0.0102	0.4065	0.2005	0.0125
Recall - 25%	0.2086	0.0107	0.4099	0.2079	0.0129
Accuracy - 15%	0.187	0.0067	0.0012	0.1804	0.0092
F-Measure - 15%	0.1821	0.0067	0.0012	0.1754	0.0094
Precision - 15%	0.1789	0.0069	0.001	0.1719	0.0096
Recall - 15%	0.187	0.0068	0.0012	0.1804	0.0092

Table 4. T-Test Results. HO-LDA vs. LDA using J48, 20 Topics

	HO-LDA Avg.	HO-LDA Standard Deviation	p Value	LDA Avg.	LDA Standard Deviation
Accuracy - 100%	0.2474	0.0055	0	0.2392	0.0064
F-Measure-100%	0.2424	0.006	0.1125	0.2405	0.006
Precision - 100%	0.2423	0.006	0.0618	0.2399	0.0059
Recall - 100%	0.2442	0.0063	0.1483	0.2425	0.0062
Accuracy - 50%	0.2155	0.01	0.1015	0.2126	0.0072
F-Measure - 50%	0.2124	0.01	0.1177	0.2097	0.0072
Precision - 50%	0.2105	0.01	0.1188	0.2078	0.0073
Recall - 50%	0.2155	0.0101	0.0965	0.2125	0.0073
Accuracy - 25%	0.2311	0.0087	0	0.1865	0.0129
F-Measure - 25%	0.2285	0.0083	0	0.1833	0.0127
Precision - 25%	0.2273	0.0082	0	0.1811	0.0124
Recall - 25%	0.2311	0.0087	0	0.1865	0.0128
Accuracy - 15%	0.1772	0.0053	0	0.1666	0.0099
F-Measure - 15%	0.1737	0.0053	0	0.1628	0.0096
Precision - 15%	0.1714	0.0052	0	0.1601	0.0094
Recall - 15%	0.1772	0.0054	0	0.1666	0.0099

Table 5. T-Test Results – HONB vs. Naïve Bayes Using Accuracy Metric

Training Size	Avg. Accuracy HONB	Avg. Standard Deviation HONB	p Value	Avg. Accuracy Naïve Bayes	Avg. Standard Deviation Naïve Bayes
65%	0.476	0.013	0.003	0.457	0.012
60%	0.465	0.008	0	0.445	0.011
55%	0.457	0.010	0	0.432	0.011
50%	0.454	0.009	0	0.413	0.010
45%	0.441	0.009	0	0.399	0.008
40%	0.433	0.010	0	0.382	0.011
35%	0.422	0.009	0	0.360	0.009
30%	0.414	0.013	0	0.336	0.008
25%	0.403	0.012	0	0.314	0.008
20%	0.390	0.012	0	0.279	0.004

References

Aitchison, J. 1982. *The Statistical Analysis of Compositional Data.* Journal of the Royal Statistical Society.

Blei, David M. *Introduction to Probabilistic Topic Models.* To appear in Communications of the ACM.

Blei, D. "[Topic-models] The computational complexity of LDA." Message from Yangqiu Song to author. 2 Apr 2008. E-mail.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. 2010. *The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies.* Journal of the ACM.

Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J. 2004. *Hierarchical topic models and the nested Chinese restaurant process.* Advances in Neural Information Processing Systems.

Blei, D., and Lafferty, J. 2009. *Topic Models*. Text Mining: Classification, Clustering, and Applications.

Blei, D., Ng, A. Y., Jordan, M. I. 2003. *Latent Dirichlet Allocation.* Journal of Machine Learning Research.

Carpenter, Tamra, Cheng, Jerry, Roberts, Fred, Minge, Xie. 2009. *Sensor Management Problems of Nuclear Detection.* Unpublished manuscript.

Celikyilmaz, A., Hakkani-Tur, D. 2010. *A hybrid Hierarchical Model for Multi-Document Summarization*. Proceedings of the 48th Annual Meeting for the Association of Computational Linguistics.

Chaturvedi, S. K., Swami, D. K., Singh, G. 2011. *Dirichlet Distribution with Centroid Model (DDCM) based Summarization Technique for Web Document Classification.* Proceedings of the COMPUTE '11 Fourth Annual ACM Bangalore Conference.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, C., Wu, D., Tapai, A., Giles, L., Jansen, B., and Yen, J. 2011. *Classifying Text Messages for the Haiti Earthquake*. Proceedings of the 8th International ISCRAM Conference.

Ganiz, Murat Can, George, Cibin, Pottenger, William M. 2011. *Higher Order Naïve Bayes: A Novel Non-IID Approach to Text Classification.* IEEE Transactions on Knowledge and Data Engineering.

Ganiz, M. C., Lytkin, N. I., Pottenger, W. M. 2009. *Leveraging Higher Order Dependencies Between Features for Text Classification*. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Ganiz, M., and Pottenger, W.M. 2010. *A Novel Bayesian Classifier for Sparse Data*. IEEE Transactions of Knowledge and Data Engineering (TKDE).

Heinzelman, J., Waters, C. 2010. *Crowdsourcing Crisis Information in Disaster-Affected Haiti.* United States Institute of Peace Special Report.

Hoffman, M., Blei, D., Cook, P. 2008. *Content-based musical similarity computation using the Hierarchical dirichlet process.* Proceedings of the International Conference on Music Information Retrieval.

Hong, L., Davison, B. D. 2010. *Empirical Study of Topic Modeling in Twitter.* Proceedings of the 1st Workshop on Social Media Analytics (SOMA '10).

Jordan, M. I. *Dirichlet Processes, Chinese Restaurant Processes, and all that*. 2005. Proceedings of the 22nd International Conference on Machine Learning (ICML).

Jung, Y., Park, H., Du, D., Drake, B. L. 2003. *A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering.* Journal of Global Optimization.

Nelson, Christie and Pottenger, William M. 2011. *Nuclear Detection Using Higher Order Learning*. Proceedings of the IEEE International Conference on Technologies for Homeland Security (IEEE-HST).

Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. 2012. *Nuclear Detection Using Higher-Order Topic Modeling.* To appear in the proceedings of the IEEE International Conference on Technologies for Homeland Security (IEEE-HST).

Perry, Jason. 2009. *Clustering and Machine Learning for Gamma Ray Spectroscopy.* Unpublished manuscript.

Porter, M. 1980. *An algorithm for suffix stripping*. Program.

Pottenger, William M., Kantor, Paul, Li, Shenzhi, Kolipaka, Kashyap, Pandya, Chirag. *Final Report on Entity Resolution System.* U.S. Department of Justice, National Institute of Justice, Information Led Policing Research, Technology Department, Testing, and Evaluation.

Sugar, C. A., James, G. M. 2003. *Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach*. Journal of the American Statistical Association.

Teh, Y. W., Jordan, M. I, Beal, M. J., Blei, D. M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association.*

Thermo Scientific. 2011. User Manual Interceptor™ Spectroscopic Personal Radiation Detector.

Tibshirani, R., Walther, G., Hastie, T. 2001. *Estimating the Numbers of Clusters in a Data Set Via the Gap Statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Ushahidi. http://www.ushahidi.com/

Zipf, G. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.