Classification of Twitter Disaster Data Using a Hybrid Feature-Instance Adaptation Approach

Reza Mazloom
Kansas State University, US
rmazloom@ksu.edu

HongMin Li
Kansas State University
hongminli@ksu.edu

Doina Caragea*
Kansas State University
dcaragea@ksu.edu

Muhammad Imran
Qatar Computing Research Institute (HBKU)
mimran@hbku.edu.qa

Cornelia Caragea
Kansas State University
ccaragea@ksu.edu

ABSTRACT

Huge amounts of data that are generated on social media during emergency situations are regarded as troves of critical information. The use of supervised machine learning techniques in the early stages of a disaster is challenged by the lack of labeled data for that particular disaster. Furthermore, supervised models trained on labeled data from a prior disaster may not produce accurate results, given the inherent variation between the current and the prior disasters. To address the challenges posed by the lack of labeled data for a target disaster, we propose to use a hybrid feature-instance adaptation approach based on matrix factorization and the k-nearest neighbors algorithm, respectively. The proposed hybrid adaptation approach is used to select a subset of the source disaster data that is representative for the target disaster. The selected subset is subsequently used to learn accurate Naïve Bayes classifiers for the target disaster.

Keywords

Tweet classification, Domain adaptation, Matrix factorization, k-Nearest Neighbors, Disaster response

INTRODUCTION

Social media is becoming a more prevalent part of our everyday life, due to the advancements in technology and virtualization. The availability of the Internet, cameras and real-time message boards at our fingertips has brought about live and parallel reporting, and witness testimonies during many events. These reports can be useful to responders and can help create awareness among the populace, especially in emergency situations (Meier 2015; Watson et al. 2017). Despite the potential benefits, major response groups and organizations under-utilize these sources of information, as therein lie many administrative and technical challenges (Meier 2013). Among the challenges, there are reliability issues associated with public and unstructured data, as well as information overload issues, as millions of messages are posted during a crisis situation (Bullock et al. 2012).

There are many recent studies that propose the use of machine learning techniques to provide automated methods for analyzing social media data to reduce the information overload (Imran, Castillo, et al. 2015; Beigi et al. 2016). Machine learning techniques can help transform raw data into usable information by labeling, prioritizing and structuring data, and making them beneficial to responders and to the populace in times of need (Qadir et al. 2016).

*corresponding author

However, supervised learning algorithms rely on labeled training data to build predictive models. Accurate labeling of data for an emerging disaster is both time consuming and expensive, and, hence, it is not appropriate to assume that labeled data for a current disaster will be promptly available to be used for analysis. The lack of labeled data for emerging disasters prohibits the use of supervised learning techniques.

To address this challenge, several works proposed to use labeled data from prior "source" disasters to learn *supervised* classifiers for a "target" disaster (Verma et al. 2011; Imran, Elbassuoni, et al. 2013; Imran, Mitra, et al. 2016). However, due to the divergence of each disaster domain in terms of location, nature, season, etc. (Palen and Anderson 2016), the source disaster might not accurately represent the characteristics of the target disaster (Qadir et al. 2016; Imran, Castillo, et al. 2015). Domain adaptation techniques (Pan and Yang 2010; Jiang 2008) are designed to circumvent the lack of labeled target data by making use of unlabeled target data as guideposts for the readily available labeled source data. Studies in the disaster space have shown that using domain adaptation techniques, which use together target unlabeled data and source labeled data, significantly improve classification results as compared to supervised techniques that use solely labeled source data (Li, Guevara, et al. 2015; Li, D. Caragea, C. Caragea, and Herndon 2017). Unlabeled data from the target disaster become more abundant as the event unfolds, and it can enable the use of domain adaptation techniques during emerging or occurring disasters.

There are several ways in which the unlabeled target data can be used with domain adaptation techniques, including parameter-based adaptation, instance-based adaptation and feature-based adaptation (Pan and Yang 2010). In the parameter-based adaptation, the labeled source data is used together with the unlabeled target data to identify shared parameters that result in good predictions for the target data. In the instance-based adaptation, the unlabeled target data is used to identify and/or reweigh the most relevant source labeled instances with respect to the target classification task, while in feature-based adaptation, the target unlabeled data and source labeled data are used together to find a feature representation that minimizes the difference between the two domains. Prior work on disaster tweet classification using domain adaptation has relied on parameter-based adaptation. Specifically, Li, D. Caragea, C. Caragea, and Herndon (2017) proposed to learn weighted source and target Naïve Bayes classifiers with the iterative method of Expectation-Maximization (EM) (Dempster et al. 1977), and showed that the resulting classifiers can accurately predict the target.

In this study, we propose to use a combination of two domain adaptation approaches, specifically a hybrid between feature-based adaptation and instance-based adaptation, to reduce the variation between the two domains. First, the Alternating Nonnegative Least Squares Matrix Factorization (LSNMF) (Lin 2007) is used on the combined source and target data, represented using binary vectors, to create a dense and reduced conceptual representation of source and target instances. Subsequently, the k-Nearest Neighbors algorithm (kNN) is used to select a subset of the source instances which are most similar to the target instances, according to the cosine similarity calculated based on the reduced common representation. The objective is to gain an understanding of the benefits provided by the hybrid feature-instance adaptation approach, as compared to the independent feature or instance adaptation approaches. Furthermore, given that both the LSNMF approach and the kNN approach have parameters that need to be tuned, specifically, the number of reduced features f for LSNMF and the number of neighbors k for kNN, we aim to study the variation of performance with these parameters and identify overall good values that can be used in practice.

As an application, we focus on the task of classifying disaster tweets as being relevant to the disaster of interest (*i.e.*, on-topic) or not relevant (*i.e.*, off-topic). This is one of the most basic but crucial classifications needed during a disaster, as subsequent analysis should be done only on data relevant to the disaster in question. Furthermore, this classification is not trivial: supervised classifiers may not achieve accurate results due to domain variations.

To summarize, our main contributions are as follows:

- We design a hybrid feature-instance adaptation approach to adapt the source disaster data to the target disaster data. Specifically, we use a matrix factorization approach to construct a shared representation of source and target instances, and subsequently use the kNN algorithm to select source instances that are most similar to target instances. Finally, we train supervised Naïve Bayes classifiers on the modified source data.
- We perform an extensive set of experiments on pairs of source-target disasters from the CrisisLexT6 datasets to evaluate the feature-instance adaptation approach by comparison with approaches that make use of either feature-based adaptation or instance-based adaptation, but not both.
- We study the variation of performance with the parameters of the feature-based adaptation (specifically, the number of features, f), and instance-based adaptation (specifically, the number of neighbors, k), respectively, to identify parameters that result in good overall performance.

METHODS

There are many traditional machine learning techniques that can be used for disaster tweet classification, such as Naïve Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), etc. Compared to other algorithms, Naïve Bayes has the advantage of not requiring hyper-parameter tuning. Furthermore, a recent study on disaster tweet classification (Li, D. Caragea, C. Caragea, and Herndon 2017) has shown that the results obtained with Naïve Bayes are comparable, and sometimes better, than the results obtained with other more sophisticated algorithms used with default parameters. Therefore, in this work, we will use Naïve Bayes together with a hybrid feature-instance adaptation approach to learn classifiers for disaster data, as described below.

Given a source and target pair of disasters, our goal is to adapt the source data by reducing the variance with respect to the target data, and then train Naïve Bayes classifiers on the adapted source data. The source adaptation is guided by the target unlabeled data. More specifically, we propose a hybrid feature-instance adaptation approach to select a subset of the source instances, which are most similar to the target instances. First, the target instances are used to construct a target vocabulary V, which is subsequently used to represent both source and target data as bag-of-words binary vectors. As part of the feature adaptation step, the resulting data matrix D is decomposed using the popular Least Squares Non-negative Matrix Factorization (LSNMF) proposed by Lin (2007). The implementation of this method is available in Python under the “nimfa” package. Intuitively, the decomposition will produce a reduced dense representation of the data, which is more suitable for identifying similar instances as compared to the sparse binary representation (Guo and Diab 2012).

As part of the instance adaptation step, the reduced representation is used to identify source instances that are most similar to the target. More precisely, for each target (unlabeled) instance, we calculate the cosine similarity to the source instances and select the k nearest neighbors from the source. If two different target instances have the same source instance among the k nearest neighbors, the selected subset of the source may contain duplicate instances. We experiment with two settings, one in which we retain duplicates (i.e., we reweigh source instances), and another one in which we remove duplicates (under the assumption that duplicates can bias the classifier).

Finally, we use the Naïve Bayes algorithm to learn classifiers from the selected subset of the source. Here, we also experiment with two settings: one in which the Gaussian Naïve Bayes algorithm is used on the reduced representation of the selected source instances, and another one in which the Bernoulli Naïve Bayes algorithm is used on the original binary representation of the selected source instances. The reason we also experiment with the binary representation of the adapted source is that in preliminary experimentation the binary representation gave better results than the numeric TF-IDF representation (results not shown due to space constrains). Finally, the resulting classifiers are tested on separate target test data. The approach is summarized in Algorithm 1.

Algorithm 1: Hybrid feature-instance adaptation with Naïve Bayes classifiers

1. Given: Target unlabeled data TU, source labeled data SL, and target test data TT.
2. Use target unlabeled data TU to construct the vocabulary V.
3. Represent source SL and target TU data as binary vectors. The resulting data matrix is denoted by D.
4. *Feature adaptation*: Use the Least Squares Non-negative Matrix Factorization to obtain a reduced representation of the source and target data. The dimension of the reduced representation is denoted by f.
5. *Instance adaptation*: For each target instance in TU, find its k nearest neighbors and add them to the selected subset of source instances $Sel\text{-}SL$, by retaining duplicates or by removing duplicates, respectively.
6. *Naïve Bayes*: Use the selected subset of source instances $Sel\text{-}SL$, with the reduced representation or the original binary representation, respectively, to learn a classifier for the target data.
7. Evaluate the resulting Naïve Bayes classifier on the target test data TT.

DATASET

The CrisisLexT6 dataset (Olteanu et al. 2014) is a collection of six disasters that occurred between October 2012 and July 2013 in United States, Canada and Australia. This dataset was collected through Twitter API based on disaster keywords and the geographic locations of the affected areas. Each disaster’s data contains approximately 10,000 tweets which were manually labeled as on-topic or off-topic using CrowdFlower, a popular crowdsourcing

Table 1. Summary of source and target disaster pairs used in the experiments, together with information about instances and features in the combined source and target datasets

Crisis			Instances			Features
Abbreviation	Source	Target	On-topic	Off-topic	Total	
BB-AF	Boston Bombings	Alberta Floods	7938	9023	16961	1322
BB-OT		Oklahoma Tornado	7650	9358	17008	1143
BB-WT		West Texas Explosion	8564	9042	17606	1239
OT-AF	Oklahoma Tornado	Alberta Floods	6706	9763	16469	1322
QF-AF	Queensland Floods	Alberta Floods	6733	9264	15997	1322
QF-BB		Boston Bombings	7677	8859	16536	1317
QF-OT		Oklahoma Tornado	6445	9599	16044	1143
SH-AF	Sandy Hurricane	Alberta Floods	8758	8466	17224	1322
SH-BB		Boston Bombings	9702	8061	17763	1317
SH-OT		Oklahoma Tornado	8470	8801	17271	1143
SH-QF		Queensland Floods	8497	8302	16799	1242
SH-WT		West Texas Explosion	9384	8485	17869	1239

platform. The data was cleaned according to the pre-processing steps described in (Li, Guevara, et al. 2015), which included removing re-tweets (RT), duplicate tweets, non-printable ASCII characters, and replacing URL, email addresses and usernames with placeholders pertaining to each. Furthermore, the dataset is split into combinations of consecutive source-target pairs of all six disasters and converted into bag-of-words binary (word existence) representations. Each feature (word) must appear at least 10 times in any given pair of disasters to be included in the vocabulary as a feature. Hence, the feature set is different from one source-target pair to the another, although, on average, pairs have approximately 1200-1300 features.

EXPERIMENTAL SETUP

In this section, we state the research questions that are driving our experiments, describe the evaluation setup and also the parameter setting for the constituent approaches of the experiments, and finally our baselines.

Research Questions

Our experiments are designed to answer the following research questions:

- Are the adaptation approaches more effective than the baseline, where Bernoulli Naïve Bayes is used to learn classifiers from the binary representation of the source data?
- Is the hybrid feature-instance adaptation approach more effective than the individual feature adaptation and instance adaptation approaches? Between Gaussian Naïve Bayes on the reduced representation of the selected source data and Bernoulli Naïve Bayes on the binary representation of the selected source data, which classifier gives better results?
- Between the feature adaptation approach and the instance adaptation approach, which one is more effective? What parameter values result in better performance for the two approaches, respectively?
- When using the instance adaptation approach, is it better to keep duplicate neighbors or to remove them?

Evaluation Strategy

We consider the six disasters in our dataset in chronological order and create 12 pairs of source and target disasters, by ensuring that the source disaster has occurred before the target disaster (under the assumption that a later disaster may mention an earlier disaster but not the other way around). This strategy creates pairs of natural or man-made disasters, but also pairs that contain a combination of natural and man-made disasters. In our result tables, we use the abbreviations shown in Table 1 to specify the source and target disasters in a pair, respectively.

We used the 5-fold cross-validation technique for each pair of disasters to select target test and target unlabeled data. Similar to (Li, D. Caragea, C. Caragea, and Herndon 2017), the folds are rotated five times to obtain five combinations of consecutive folds, within each selecting the first three folds as target unlabeled TU data, the next fold as target test TT data, and last one as target labeled TL data (reserved for future work). Each domain has between 8000-9000 instances, as can be seen from Table 1. Only target unlabeled data is used with the instance adaptation approach, which means that the classifiers are different for each test fold, as they are trained from different subsets of the source instances, as guided by the corresponding unlabeled target data. We report accuracy results averaged over 5 folds.

Matrix Factorization Setup

Data from each pair of disasters, represented as a binary matrix which consists of bag-of-words vectors, is reduced using the LSMNF technique. Specifically, the number of features f for each pair is reduced from approximately 1200-1300 to 30, 50, 100, 200, and 500 features, respectively.

K-Nearest Neighbors Setup

The kNN algorithm is used to select the k nearest neighbors from the entire source for each of the instances in the target unlabeled dataset. We experiment with the following values for k: 1, 3, 5, 7, 9, 11, to understand what value of k results in best overall performance. As there is a possibility of having the same source neighbor for multiple target instances, duplicates may exist in the source subset. Hence, we experiment with two options: retaining duplicates (*d*) or not retaining duplicates (*n*) to understand which one is more appropriate.

Bernoulli Naïve Bayes and Gaussian Naïve Bayes

After selecting a subset of the source instances using the hybrid feature-instance adaptation, the next step is to learn a Naïve Bayes classifier from the adapted source. We experiment with two options. First, we use the reduced representation of the selected source subset (*r*) to train Gaussian Naïve Bayes classifiers. Furthermore, we also use the original binary representation of the instances in the selected source subset (*b*) to train Bernoulli Naïve Bayes classifiers, given preliminary experimentation that showed better results with Bernoulli Naïve Bayes on the binary representation, as compared to Gaussian Naïve Bayes on the TF-IDF representation.

Baselines

We compare our proposed approach against the following baselines:

- *Supervised Bernoulli Naïve Bayes classifiers* learned from the binary representation of the source and evaluated on the test target data.
- *Instance adaptation with Bernoulli Naïve Bayes classifiers*, where we first use the binary representation of the source to identify a subset of instances most similar to the target instances, and subsequently learn Bernoulli Naïve Bayes classifiers from the selected source subset.
- *Feature-adaptation with Gaussian Naïve Bayes classifiers*, where we first use the binary representation of the source and target to find a reduced dense representation, and subsequently learn Gaussian Naïve Bayes classifiers from the selected source subset.

EXPERIMENTAL RESULTS AND DISCUSSION

Instance Adaptation with Bernoulli Naïve Bayes Classifiers

Instance adaptation is performed on the original binary representation of the combined source and (unlabeled) target datasets using kNN. Specifically, for each target instance we select the k nearest neighbors from the corresponding source. Subsequently, Bernoulli Naïve Bayes is used on the selected source subset, with duplicates (*d*) or no duplicates (*n*). Table 2 shows the results of this set of experiments. As can be seen, the best results overall are obtained for the model labeled *3k-n* which is a model where the 3 nearest neighbors are selected for each target instance, and duplicates are not kept in the selected source subset. Furthermore, the performance slightly decreases for values of k greater than 3 (regardless of the fact that duplicates are retained or removed), suggesting that noisy source instances are added to the selected subset when more than 3 neighbors are included. Given this observation (and other preliminary experiments now shown), the subsequent experiments that make use of kNN will be run with $k = 3$. When comparing the instance adaption results with the results of Bernoulli Naïve Bayes on the original binary data (labeled *Original* in Table 2), it can be seen that the instance adaptation consistently improves the classification accuracy by as much as 7.4% and 6% in the case of SH-BB and BB-AF pairs, respectively.

Table 2. Instance-based adaptation using kNN on the binary representation, followed by Bernoulli Naïve Bayes, on the selected source subset. Accuracy results on 12 source-target pairs are shown for three values of k, specifically, 3, 5, and 7, and two instance-selection settings, specifically, with duplicates (denoted by d) and with no duplicates (denoted by n). For example, *3k-d* means that 3 nearest neighbors are selected for each target instance, and duplicates are retained, while *3k-n* means that 3 nearest neighbors are selected, but duplicates are removed. The *Original* results are obtained when Bernoulli Naïve Bayes is used on the original binary data. Significant best results for each pair are highlighted in bold (based on a t-test with $p \leq 0.05$).

Source	BB	BB	BB	OT	QF	QF	QF	SH	SH	SH	SH	SH
Target	AF	OT	WT	AF	AF	BB	OT	AF	BB	OT	QF	WT
Original	0.738	**0.843**	**0.948**	0.872	0.789	0.750	0.841	0.711	0.687	0.808	0.768	0.772
3k-d	**0.744**	0.842	0.946	**0.871**	**0.813**	0.728	**0.848**	**0.759**	**0.755**	**0.832**	**0.823**	0.820
5k-d	0.736	**0.847**	**0.949**	0.868	**0.811**	0.717	**0.848**	**0.756**	**0.758**	**0.830**	**0.819**	0.809
7k-d	0.732	0.844	**0.948**	0.869	**0.810**	0.712	**0.850**	**0.757**	**0.753**	**0.828**	0.816	0.830
3k-n	**0.752**	**0.842**	**0.946**	**0.874**	**0.806**	**0.780**	**0.853**	0.726	**0.747**	**0.829**	0.789	**0.846**
5k-n	**0.749**	**0.846**	**0.946**	**0.874**	0.804	0.770	0.850	0.724	0.739	0.823	0.786	0.833
7k-n	**0.746**	**0.848**	**0.947**	**0.874**	0.800	0.772	**0.849**	0.720	0.733	0.817	0.782	0.820

Feature Adaptation with Gaussian Naïve Bayes Classifiers

Similar to the instance-based adaptation, the feature-based adaptation is also performed on the original binary data matrix, consisting of source and (unlabeled) target data. The goal of this adaptation is to create a denser feature set that better captures the similarity between target and source instances, and ultimately produces better classification results. We use a wide range of dimensions, specifically 30, 50, 100, 200 and 500. Table 3 shows the results of the Gaussian Naïve Bayes classifiers trained on the reduced representations, by comparison with the results of the Bernoulli Naïve Bayes classifiers trained on the original binary representation. As can be seen, the highest accuracy overall is obtained with the reduced representation, although there are pairs for which the original representation gives better results. This suggests that the reduced representation by itself is not always enough to ensure best results on the target. The results in Table 3 also show that the classifiers trained with 200 reduced features (i.e., *200f*) give the best results overall, while sometimes the models trained with 50 or 100 reduced features give the best results for specific pairs. In subsequent experiments we will only train classifiers with 50 and 200 features to reduce the number of experiments (by eliminating several values from the original feature adaptation experiment).

Table 3. Feature-based adaptation using LSNMF on the original binary representation, followed by Gaussian Naïve Bayes on the reduced representation. Accuracy results on 12 source-target pairs are shown for six values of f, specifically, 30, 50, 100, 200, 500, and 1000. For example, $50f$ means that the LSNMF decomposition has 50 reduced features. The *Original* results are obtained with Bernoulli Naïve Bayes on the original binary data. Significant best results for each pair are highlighted in bold (based on a t-test with $p \leq 0.05$).

Source	BB	BB	BB	OT	QF	QF	QF	SH	SH	SH	SH	SH
Target	AF	OT	WT	AF	AF	BB	OT	AF	BB	OT	QF	WT
Original	0.738	**0.843**	**0.948**	**0.872**	0.789	**0.750**	**0.841**	0.711	0.687	**0.808**	0.768	0.772
30f	**0.792**	0.801	0.922	0.756	**0.860**	0.444	0.770	0.720	**0.807**	0.751	0.781	0.704
50f	0.764	**0.850**	0.921	0.774	0.796	0.457	0.760	0.790	0.565	0.766	0.809	0.736
100f	0.563	0.829	**0.948**	0.798	0.849	0.643	0.808	**0.828**	0.698	0.758	0.819	**0.843**
200f	0.618	**0.851**	0.932	0.814	0.841	0.729	0.815	**0.834**	0.669	0.743	**0.833**	**0.846**
500f	0.721	0.807	0.936	0.815	0.824	0.463	0.694	0.799	0.671	0.742	**0.840**	0.825

Table 4. Hybrid feature-instance adaptation. Accuracy results on 12 source-target pairs are shown for $k = 3$, $f = 50, 200, 500$, respectively, combined with settings with duplicates (denoted by *d*) or with no duplicates (denoted by *n*). Naïve Bayes is run on the selected source subset with reduced (denoted by *r*) and binary (denoted by *b*) representations, respectively. For example, *50f-3k-d-r* means that LSNMF gives 50 reduced features, kNN selects 3 nearest neighbors, duplicated are retained, and the Gaussian Naïve Bayes is trained on the reduced representation, while *50f-3k-n-b* means that there are no duplicates and Bernoulli Naïve Bayes is trained on the binary representation of the selected source subset. The *Original* results are obtained when Bernoulli Naïve Bayes is used on the original data. Significant best results for each pair are highlighted in bold (based on a t-test with $p \leq 0.05$).

Source Target	BB AF	BB OT	BB WT	OT AF	QF AF	QF BB	QF OT	SH AF	SH BB	SH OT	SH QF	SH WT
Original	0.738	**0.843**	**0.948**	**0.872**	0.789	0.750	0.841	0.711	0.687	0.808	0.768	0.772
50f-3k-d-r	0.757	0.822	0.928	0.753	0.780	0.749	0.758	0.775	0.645	0.696	0.796	0.883
50f-3k-n-r	0.742	0.828	0.934	0.771	0.775	0.618	0.758	0.757	0.653	0.757	0.765	**0.884**
200f-3k-d-r	0.705	0.816	0.923	0.799	**0.834**	**0.771**	0.803	**0.838**	**0.797**	0.707	0.833	0.828
200f-3k-n-r	0.669	0.789	0.931	0.815	0.820	0.762	0.788	0.803	0.753	0.691	0.790	0.806
50f-3k-d-b	**0.782**	**0.846**	0.940	0.868	0.810	0.716	**0.848**	0.805	0.746	0.815	**0.851**	**0.895**
50f-3k-n-b	0.764	**0.840**	**0.945**	**0.873**	0.807	**0.773**	**0.850**	0.773	0.762	**0.846**	0.834	**0.891**
200f-3k-d-b	0.758	0.836	0.926	0.865	0.773	0.694	0.822	0.789	0.766	0.815	**0.858**	0.868
200f-3k-n-b	0.766	0.830	0.939	**0.874**	0.795	**0.762**	0.831	0.784	0.776	**0.843**	0.839	**0.897**

Hybrid Feature-Instance Adaptation with Bernoulli/Gaussian Naïve Bayes

Finally, we experiment with our proposed hybrid feature-instance adaptation approach combined with Gaussian and Bernoulli Naïve Bayes classifiers, respectively. We fix the value of k to 3, as this value gave the best results in our instance adaptation experiments, and fix f to 50 or 200 reduced features, respectively. For kNN, we experiment with duplicates (*d*) and with no-duplicates (*n*) options. Finally, once we select a subset of the source, we train Gaussian Naïve Bayes classifiers on the reduced representation of that subset (*r*), and Bernoulli Naïve Bayes classifiers on the binary representation of that subset (*b*). The results of the experiments are shown in Table 4. As can be seen, the results of the hybrid approach are overall better than the results of the original models. In two cases, specifically, SH-AF and SH-WT, the increase in performance is close to 13%. Between duplicates and no-duplicates options, the no-duplicates option is usually better than the duplicates option, suggesting that the combination of feature and instance adaptation is good at identifying source instances that are representative for the target and prevents the need for changing the weights of the source instances (which was already apparent in the instance adaptation approach that used the sparse binary representation to find neighbors). Regarding the number of reduced features f, the results obtained with 50 features are overall better than the results obtained with 200 features. However, when looking at duplicate retainment and feature reduction together, we observe that they affect each other. For example, we can compare the difference between *50f-3k-d-b* and *50-3k-n-b*, on one hand, and *200f-3k-d-b* and *200-3k-n-b*, on the other hand. It can be observed that in the case of *50f* features the performance is overall higher for the no-duplicates option, as compared to the duplicates option, while this is not the case when considering *200f* features. Intuitively, a higher-level representation (i.e., smaller number of features) helps identify good nearest neighbors, which in turn helps obtain good performance.

Finally, when comparing the performance of the Gaussian Naïve Bayes classifiers with the performance of the Bernoulli Naïve Bayes classifiers, the results are not conclusive: Gaussian Naïve Bayes classifiers give better results for half of the pairs, while Bernoulli Naïve Bayes classifiers give better results for the other half.

Summary of the Results and Discussion

A summary of our results is shown in Table 5, where we compare the original classifiers with the feature adaptation, instance adaptation and hybrid feature-instance adaptation classifiers. We will use the results in this table to answer our original research questions.

Are the adaptation approaches more effective than the baseline, where Bernoulli Naïve Bayes is used to learn classifiers from the binary representation of the source data? As can be seen from Table 5, the adaptation-based classifiers are generally significantly better than the original classifiers.

Table 5. Summary of the results for 12 source-target pairs. The upper section of the table contains the individual feature adaptation (*50f* and *200f*) and instance adaptation (*3k-d* and *3k-n*) approaches, while the bottom section contains the hybrid approach for *50f* and *200f*, respectively, and *3k*. Significant best results for each pair are highlighted in bold (based on a t-test with $p \leq 0.05$).

Source	BB	BB	BB	OT	QF	QF	QF	SH	SH	SH	SH	SH
Target	AF	OT	WT	AF	AF	BB	OT	AF	BB	OT	QF	WT
Original	0.738	**0.843**	**0.948**	**0.872**	0.789	0.750	0.841	0.711	0.687	0.808	0.768	0.772
3k-d	0.744	0.842	**0.946**	0.871	0.813	0.728	**0.848**	0.759	0.755	0.832	0.823	0.820
3k-n	0.752	0.842	**0.946**	**0.874**	0.806	**0.780**	**0.853**	0.726	0.747	0.829	0.789	0.846
50f	0.764	**0.850**	0.921	0.774	0.796	0.457	0.760	0.790	0.565	0.766	0.809	0.736
200f	0.618	**0.851**	0.932	0.814	**0.841**	0.729	0.815	**0.834**	0.669	0.743	0.833	0.846
50f-3k-d-r	0.757	0.822	0.928	0.753	0.780	0.749	0.758	0.775	0.645	0.696	0.796	0.883
50f-3k-n-r	0.742	0.828	0.934	0.771	0.775	0.618	0.758	0.757	0.653	0.757	0.765	**0.884**
50f-3k-d-b	**0.782**	**0.846**	0.940	0.868	0.810	0.716	**0.848**	0.805	0.746	0.815	**0.851**	**0.895**
50f-3k-n-b	0.764	0.840	**0.945**	**0.873**	0.807	**0.773**	**0.850**	0.773	**0.762**	**0.846**	0.834	**0.891**
200f-3k-d-r	0.705	0.816	0.923	0.799	**0.834**	**0.771**	0.803	**0.838**	**0.797**	0.707	0.833	**0.828**
200f-3k-n-r	0.669	0.789	0.931	0.815	0.820	0.762	0.788	0.803	0.753	0.691	0.790	0.806
200f-3k-d-b	0.758	0.836	0.926	0.865	0.773	0.694	0.822	0.789	0.766	0.815	**0.858**	0.868
200f-3k-n-b	0.766	0.830	0.939	**0.874**	0.795	0.762	0.831	0.784	0.776	**0.843**	0.839	**0.897**

Is the hybrid feature-instance adaptation approach more effective than the individual feature adaptation and instance adaptation approaches? To easily answer this question based on Table 5, we have separated the table into two sections: one for the individual feature adaptation and instance adaptation approaches, and the other one for the hybrid approach. The best results for each pair (based on a t-test with $p \leq 0.05$) are highlighted in bold. As can be seen, the hybrid approach achieves best results for all 12 pairs, while the feature adaptation approach achieves best results for only 3 pairs and the instance adaptation approach achieves best results for only 4 pairs. While the individual adaptation approaches with *200f* and *3k-n* achieve best results for 7 pairs combined, the other results obtained with these approaches are not competitive. The hybrid approach with *50f-3k-d-b* and *50f-3k-n-b* settings achieves either best values for almost all pairs or values closest to the best values for other pairs. In other words, the individual adaptation approaches can produce very good results in some cases and poor results in other cases, while the hybrid feature-instance adaptation approach with *50f*, *3k* and no duplicates can produce competitive results consistently, suggesting that this approach is more reliable.

Between Gaussian Naïve Bayes on the reduced representation of the selected source data and Bernoulli Naïve Bayes on the binary representation of the selected source data, which classifier gives better results? As mentioned above, this question does not have a definite answer, as the Gaussian Naïve Bayes classifiers give better results for half pairs and the Bernoulli Naïve Bayes classifiers give better results for the other half pairs.

Between the feature adaptation approach and the instance adaptation approach, which one is more effective? What parameter values result in better performance for the two approaches, respectively? The instance adaptation approach gives better results than the feature adaptation approach for 7 out of 12 pairs and they have a tie for 2 pairs. Thus, we can say that the two approaches have complementary strengths, as the instance adaptation has performed well on pairs where feature adaptation has not performed well, and vice-versa. Furthermore, we observe that the feature adaptation performs better on pairs with more dissimilar source and target datasets, as opposed to the instance adaptation which performs better on pairs with more similar source and target datasets. Consequently, combining the instance based and the feature based approaches should ensure good results, as seen in our experiments. In terms of parameters, for the instance adaptation approach, the best results were obtained for $k = 3$. As for the number of reduced features, when comparing the hybrid models with *50f* versus *200f*, the results are visibly better for *50f*. The opposite is true for the feature adaptation models, where better results are observed for *200f* as compared to *50f*.

When using the instance adaptation approach, is it better to keep duplicate neighbors or to remove them? When using the instance adaptation approach on the original binary representation of the data, it is better to remove duplicates. Similarly, when using the instance adaptation approach in combination with the feature adaptation approach on the original binary representation, the results are better when removing duplicates. However, the option where duplicates are retained is more beneficial when using the reduced representation with Gaussian Naïve Bayes.

RELATED WORK

Machine learning algorithms have been used to help responders sift through the huge amount of crisis data, and prioritize information that may be useful for response and relief (Verma et al. 2011; C. Caragea, McNeese, et al. 2011; Vieweg 2012; Terpstra et al. 2012; Purohit et al. 2013; Imran, Elbassuoni, et al. 2013; C. Caragea, Squicciarini, et al. 2014; Ashktorab et al. 2014; Sen et al. 2015; Huang and Xiao 2015; Imran, Chawla, et al. 2016). For example, Imran, Elbassuoni, et al. (2013) used conditional random fields to find tweets within specific situational awareness categories. Sen et al. (2015) used Support Vector Machine (SVM) classifiers to differentiate between situational and non-situational tweets. Huang and Xiao (2015) introduced a detailed list of situational awareness categories, divided based on three stages of a disaster (preparedness, emergency response, and recovery), and used k-Nearest Neighbors, Logistic Regression and Naïve Bayes classifiers to automatically classify tweets with respect to the categories defined.

While research on supervised machine learning in the area of emergency response has shown that it is possible to automatically classify disaster-related data, it has also emphasized one of the most important challenges that precludes the use of supervised machine learning in real time in an emerging crisis situation: *the lack of labeled data to train reliable supervised models as the crisis unfolds*. To address this challenge, several works proposed to use labeled data from prior "source" crises to learn *supervised* classifiers for a "target" crisis (Verma et al. 2011; Imran, Mitra, et al. 2016; C. Caragea, Silvescu, et al. 2016; Nguyen et al. 2017). One drawback of this approach is that supervised classifiers learned in one crisis event, do not generalize well to other events (Qadir et al. 2016; Imran, Castillo, et al. 2015), as each event has unique characteristics (Palen and Anderson 2016). Domain adaptation approaches (Pan and Yang 2010; Jiang 2008) that make use of unlabeled data from the target disaster in addition to label data from a source disaster are desirable. Some recent works (Li, Guevara, et al. 2015; Li, D. Caragea, and C. Caragea 2017; Li, D. Caragea, C. Caragea, and Herndon 2017) have shown the using domain adaptation approaches can significantly improve the results of the supervised classifiers learned from source only. According to Pan and Yang (2010), domain adaptation is achieved by performing parameter adaptation, feature adaptation or instance adaptation. A comprehensive description of works in each category can be found in (Pan and Yang 2010).

In the space of disasters, the domain adaptation approaches proposed by Li et al. (2015; 2017) can be seen as parameter-based adaptation approaches. To the best of our knowledge, there are no instance-based or feature-based adaptation approaches that have been used for classifying disaster related data. As a consequence, in this study we focus specifically on a hybrid approach that combines feature-based adaptation based on matrix factorization with instance-based adaptation based on the kNN algorithm, and compare the hybrid approach with the individual feature-based and instance-based approaches.

CONCLUSIONS AND FUTURE WORK

Social media data taken from sources such as Twitter contain invaluable data which can be used in times of crisis and emergency situations to improve response and awareness. Despite many supervised learning approaches being proposed, not many agencies and groups use these approaches to identify useful information, due to lack of labeled data for training the supervised models. In this study, we proposed a simple but powerful feature-instance adaptation approach to reduce the variation between source and target disasters. Combined with Naïve Bayes classifiers, the proposed adaptation approach produces accuracy results that are significantly better than the results of the supervised models learned from source alone, in some cases by more than 12%, when used for the task of identifying tweets related to a particular disaster.

The CrisisLexT6 dataset was used to construct twelve pairs of disasters that we experimented with. Our results showed that adaptation-based models perform significantly better than the supervised models. We also showed that feature adaptation and instance adaptation approaches have complementary strengths that can be combined to produce better results. We argued that the hybrid feature-instance adaptation approaches are more reliable due to their consistent competitive results, especially when not considering duplicates for the instance adaptation step. Overall, the results of this study can be used to recommend the best options and parameters for the adaptation approaches, based on our observations on 12 different pairs of disasters.

In future work, more experiments can be done using different classifiers, including deep learning classifiers, on the selected source data. Furthermore, different matrix factorization and clustering approaches (potentially, with different distance metrics) can be explored.

ACKNOWLEDGMENTS

The computing for this project was performed on the Beocat Research Cluster at Kansas State University. We thank the National Science Foundation for support from grant CNS-1429316, which partially funded the cluster. We also thank the National Science Foundation for support from the grants IIS-1802284, IIS-1741345, IIS-1526542 and CMMI-1541155. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the National Science Foundation. We also wish to thank our anonymous reviewers for their constructive comments.

REFERENCES

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

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

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

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

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

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Guo, W. and Diab, M. (2012). "A simple unsupervised latent semantics based approach for sentence similarity". In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 586–590.

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

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

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

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

Imran, M., Mitra, P., and Srivastava, J. (2016). "Cross-Language Domain Adaptation for Classifying Crisis-Related Short Messages". In: *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016.*

Jiang, J. (2008). "A literature survey on domain adaptation of statistical classifiers". In: *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey* 3.

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

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2017). "Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach". In: *Journal of Contingencies and Crisis Management (JCCM), Special Issue on HCI in Critical Systems.* 26.1, pp. 16–27.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A., and Tapia, A. (2015). “Twitter Mining for Disaster Response: A Domain Adaptation Approach”. In: *Proceedings of the 12th International Conference on Information Systems for Crisis Response and Management, Kristiansand, Norway*.

Lin, C.-J. (2007). “Projected gradient methods for nonnegative matrix factorization”. In: *Neural computation* 19.10, pp. 2756–2779.

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

Meier, P. (2015). *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. Boca Raton, FL, USA: CRC Press, Inc.

Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2017). “Robust Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks”. In: *11th International AAAI Conference on Web and Social Media (ICWSM)*. Montreal, CA.

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

Palen, L. and Anderson, K. M. (2016). “Crisis informatics-New data for extraordinary times”. In: *Science* 353.6296, pp. 224–225.

Pan, S. J. and Yang, Q. (2010). “A Survey on Transfer Learning”. In: *IEEE Trans. Knowl. Data Eng.* 22.10, pp. 1345–1359.

Purohit, H., Castillo, C., Diaz, F., Sheth, A., and Meier, P. (2013). “Emergency-relief coordination on social media: Automatically matching resource requests and offers”. In: *First Monday* 19.1.

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

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

Terpstra, T., De Vries, A., Stronkman, R., and Paradies, G. (2012). *Towards a realtime Twitter analysis during crises for operational crisis management*. Simon Fraser University.

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

Vieweg, S. E. (2012). “Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications”. PhD thesis. University of Colorado.

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