A Genetic Algorithm based Domain Adaptation Framework for Classification of Disaster Topic Text Tweets

Lokabhiram Dwarakanath
Department of Computer System and Technology, Universiti Malaya, Malaysia
lokabhiram@siswa.um.edu.my

Amirrudin Kamsin
Department of Computer System and Technology, Universiti Malaya, Malaysia
amir@um.edu.my

Liyana Shuib
Department of Information Systems, Universiti Malaya, Malaysia
liyanashuib@um.edu.my

Abstract: *The ability to post short text and media messages on Social media platforms like Twitter, Facebook, etc., plays a huge role in the exchange of information following a mass emergency event like hurricane, earthquake, tsunami etc. Disaster victims, families, and other relief operation teams utilize social media to help and support one another. Despite the benefits offered by these communication media, the disaster topic related posts (posts that indicate conversations about the disaster event in the aftermath of the disaster) gets lost in the deluge of posts since there would be a surge in the amount of data that gets exchanged following a mass emergency event. This hampers the emergency relief effort, which in turn affects the delivery of useful information to the disaster victims. Research in emergency coordination via social media has received growing interest in recent years, mainly focusing on developing machine learning-based models that can separate disaster-related topic posts from non-disaster related topic posts. Of these, supervised machine learning approaches performed well when the machine learning model trained using source disaster dataset and target disaster dataset are similar. However, in the real world, it may not be feasible as different disasters have different characteristics. So, models developed using supervised machine learning approaches do not perform well in unseen disaster datasets. Therefore, domain adaptation approaches, which address the above limitation by learning classifiers from unlabeled target data in addition to source labelled data, represent a promising direction for social media crisis data classification tasks. The existing domain adaptation techniques for the classification of disaster tweets are experimented with using single disaster event dataset pairs; then, self-training is performed on the source target dataset pairs by considering the highly confident instances in subsequent iterations of training. This could be improved with better feature engineering. Thus, this research proposes a Genetic Algorithm based Domain Adaptation Framework (GADA) for the classification of disaster tweets. The proposed GADA combines the power of 1) Hybrid Feature Selection component using the Genetic Algorithm and Chi-Square Feature Evaluator for feature selection and 2) the Classifier component using Random Forest to classify disaster-related posts from noise on Twitter. The proposed framework addresses the challenge of the lack of* labeled *data in the target disaster event by proposing a Genetic Algorithm based approach. Experimental results on Twitter datasets corresponding to four disaster domain pair shows that the proposed framework improves the overall performance of the previous supervised approaches and significantly reduces the training time over the previous domain adaptation techniques that do not use the Genetic Algorithm (GA) for feature selection.*

Keywords: *Crisis informatics, disaster management, machine learning, domain adaptation approaches, social media, genetic algorithm.*

Received December 5, 2021; accepted February 20, 2022
https://doi.org/10.34028/iajit/20/1/7

1. Introduction

Machine learning was introduced in disaster management two decades ago, and since then, it has evolved into becoming one of the most effective methods for filtering irrelevant social media data, thereby increasing the data analysis speed in disaster situations. Twitter posts such as “Flood waters head for the cotton fields: It’s a wait and see game for cotton growers on the Queensland-NS”, “Australia issues flood warning as mini-tornado hits” is disaster topic related example posts from Queensland Flood disaster event from CrisisLex [18] That can be categorized as useful information for crisis management. Since 2014, researchers have proposed various automated machine learning methods for the delivery of useful information in a disaster situation. Parilla-Ferrer *et al.* [21] developed a supervised classification model to automatically classify the informative and non-informative tweets from the deluge of disaster twitter datasets. The study revealed that non-informative tweets outnumbered the informative tweets; however, the informative tweets get re-tweeted, often indicating the informativeness in the time of a disaster. Thus, the authors developed a supervised classifier model to classify the disaster-related tweets from non-disaster-related tweets. In another study, Rudra *et al.* [22], utilized a supervised classifier to classify situation tweets from non-situational tweets in the aftermath of a disaster event. naïve bayes, random forest, support vector machine, linear regression were some of the classifiers which various studies utilized for supervised classification. Majority of the studies utilized CrisisLexT6 datasets from six disaster events (2013 Boston Bombing [18], 2012 Hurricane Sandy [18], 2013 Queensland Floods [18], 2013 Oklahoma Tornado [18], 2013 West Texas Explosion [18], 2013 Alberta Floods [18]) for binary classification and CrisisLexT26 containing 26 disaster events multi-label datasets. However, supervised approaches faced the issue of models not generalizing well on unseen disaster event datasets [15, 16]. In other words, when a supervised model is developed using a disaster event Flood when tested on a disaster event like Wildfire, the model performance degrades. This is because supervised models were developed by utilizing large samples of labeled datasets from prior disaster events, which may have different characteristics from the specific disaster event on which the model is tested. Thus, various studies proposed semi-supervised domain adaptation approaches to overcome the issues in the supervised approaches.

Domain adaptation approaches for classification in machine learning were first proposed by Blum and Mitchell [4] and Yarowsky [28], where the authors in the experiment combined unlabelled and labelled data for effective classification. Later in machine learning, the domain adaptation approaches were extensively utilized in various researches. A survey by Pan *et al.* [19] presents various domain adaptation approaches from various fields of research.

Domain adaptation techniques have been successfully used in text classification, sentiment analysis, etc., and hold a great promise for classification problems in disaster and crisis management. In the context of disaster tweet classification research, three state-of-art works were published [13, 14, 16]. One of the earliest works in domain adaptation approaches was proposed by Li *et al.* [13]. The authors proposed a model that combined the unlabelled data from target disaster events and labeled data from the source disaster in order to improve the performance of classification. Li *et al.* [13] utilized a Naïve Bayes classifier for classification, and the experiments were conducted on two disaster event datasets, from 2012 Hurricane Sandy (as source data) and 2013 Boston Marathon Bombing (as target data) to conduct the study. However, due to the lack of more data from various disaster events, the testing was not done extensively. Li *et al.* [14] improved the previous domain adaptation approach by incorporating weighted Naïve Bayes and tested using six disaster event datasets from Crisislex [18]. Unlike Li *et al.* [13], where all data from target disaster event was labeled using Naïve Bayes and added to source data, Li *et al.* [14] proposed an approach where only the instances labeled with high confidence was considered and added to the source disaster event data and training was performed in subsequent iterations. Later, Mazloom *et al.* [16] proposed a Hybrid approach which is a feature-instance-parameter adaptation approach, wherein the authors extended previous domain adaptation approaches by using a Random Forest classifier instead of Naïve Bayes. In the feature-instance-parameter adaptation approach, the kNN instances from the target disaster event are added to the labeled dataset from the source in subsequent iterations. Random Forest, being an Ensemble classifier, outperformed the previous approaches in terms of accuracy. The work was tested on 16 datasets - 6 disaster events from Crisislex [18] and ten events from 2CTweets Crisis [28]. This approach utilizes Matrix Factorization for Feature reduction and reduces the number of features from 1000 to matrices of 30, 50, 100, 200 and 500 features. The performance accuracy was plotted against several iterations of the Random Forest classifier. While the previous domain adaptation approaches achieved reasonable success, they still face one challenge that needs to be addressed; it would be desirable to do better feature engineering so that one can achieve the best performance in a minimal number of iterations. On the other hand, Schulz *et al*. [23] developed Semantic Abstraction model to improve the generalization of tweet classification. Meanwhile, Stowe *et al.* [24] proposed an annotation schema for identifying relevant tweets. Existing domain adaptation approaches achieve the highest accuracy in 10 to 25 iterations in work by Li *et al.* [14] and 50 to 70 iterations in work by Mazloom *et al.* [16]. Achieving best performance accuracy with minimal iterations would help in delivering useful time-critical information to the disaster victims and emergency teams much quicker so that the victims would receive the support needed.

During the dimensionality reduction and feature selection process, using a single feature selection technique does not guarantee universally optimal feature subset selection [7]. The importance of feature selection for the improvement in the performance of the algorithms in the existing literatures [6]. It is often found in the research literature that a hybrid feature selection technique offers a robust and optimal solution towards improving the overall accuracy of the model, while a single technique often achieves an immature solution. This is because, in the hybrid technique, feature selection runs on different feature selection techniques, and each one of them produces feature subsets. Then they are combined to achieve the final list of feature subsets [7]. While matrix factorization is the feature selection approach that is used in Li *et al*. [14], there are other dimensionality reduction approaches in the literature-Particle Swarm Optimization (PSO) algorithm [9], Genetic Algorithm (GA) [5, 8, 11, 12], Incremental Wrapper Subset Selection [7], etc., could be utilized as well.

Of these, the GA is used as a Wrapper class of algorithms for dimensionality reduction and feature selection in various researches [2, 3, 11, 12]. This is due to the fact that GA can handle large amounts of data [17]. Just like in genetics, in

GA, there is a population of possible parent solutions, and these go through a combination of mutation and cross-over to yield children across various generations. After several generations of re-combinations and mutations, a stopping criterion is utilized to stop the algorithm after which the 'fittest' individual solution is chosen. The advantages of GA are that they are faster in operation, gets better over time and are suitable for problems that are NP-hard [20]. GA provides a near-optimal solution to a problem in consideration. GA, being an evolutionary algorithm, mimics the process of natural selection while obtaining sets of solutions (population). Each solution is called a chromosome that consists of feature sets called genes. GA generates solutions, evaluates the fitness of a feature by using a fitness function as a guide [5, 8]. GA search utilizes crossover and mutation as operators for feature subset selection [8]. Crossover is a mechanism for swapping genes (features) between two parent chromosomes to produce new child chromosomes for the next generation [25, 26]. On the other hand, a mutation operator is utilized for flipping one or more genes to obtain the next generation of chromosomes [8]. Genetic operators search through the entire search space and find a globally optimum solution while choosing the feature subset. Thus, GA is utilized in large scale problems.

One of the limitations of the existing domain adaptation approaches is that they do not have a robust feature engineering that handles the high dimensional disaster datasets [14, 15, 16]. Due to this, the experiments incur several iterations of more than 50 [16] before convergence. In the real world, experiments with a fewer number of iterations are desirable as it helps in quickly getting the datasets from target disaster event classified with the highest accuracy. This study improves the existing domain adaptation approaches by utilizing GAs for feature selection which can result in more comprehensible features that can enable faster execution time which makes it suitable for large datasets.

The rest of the article is divided into the following: Section 2 describes the proposed Genetic Algorithm Based Domain Adaptation (GADA) framework, section 3 explains the proposed methods, and section 4 describes the experiment results and evaluation methods. Section 5 discusses the performance improvements achieved through feature engineering using GADA from the findings of the study, and section 6 concludes the article.

2. Proposed GADA Framework

The GADA Framework is shown in Figure 1, and detailed functionalities of GADA are shown in Figure 2. The *aim* of the framework is to investigate the steps involved in the separation of tweets into disaster "On-topic" (On-topic indicates posts relating to a disaster event) and "Off-topic" (Off-topic indicates the posts that are not related to disaster events) to deliver it to the disaster victims and emergency teams. In order to achieve this aim, the proposed framework is built using the following components:

1) Hybrid Feature Selection component consisting of Genetic Algorithm and Chi-Square Feature Evaluator search component combination.
   a) We use Chi-Square Feature selection [10, 28, 29] to downsize the number of features to less than 1000. Following the Chi-Square, the features are passed on to the next phase, which contains GA. Mathematically, the proposed hybrid feature selection across *N iterations* can be represented as below:

$$\sum_{i=1}^{N} (\boldsymbol{Chi\,Sq} + \boldsymbol{GA})$$

   b) GA utilizes the fitness function to evaluate the worth of the feature sets. The proposed GA is a Filter class that consists of Correlation-based Feature Selection (CFS) as the fitness function.
2) Classifier component consists of Random Forest as classifiers.
3) Data Ensemble component ensembles data from source disaster event and target disaster event.

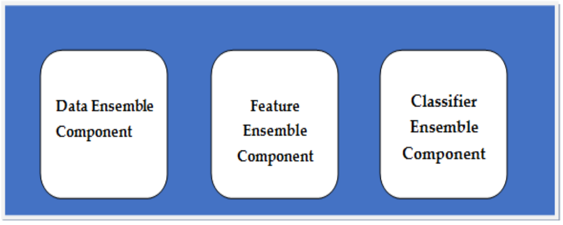

Figure 1. Simplified GADA framework.

2.1. Datasets Utilized

In this research, Twitter datasets from 4 different disaster events from diverse domains (2013 Alberta flood [18], 2013 Boston bombing [18], 2013 Oklahama Tornado [18], 2013 Queensland Floods [18], 2013 West Texas Explosion [18]) are considered. These datasets are available publicly on CrisisLex [18]. Table 1 shows the number of data instances (tweets) considered for research from each disaster event taken from CrisisLex. These datasets are manually labeled datasets labeled as "on-topic" and "off-topic".

Table 1. Distribution of labelled tweet datasets for four disaster events from CrisisLex [18].

Disaster event	No. of Instances in data	On-topic	Off-topic
Alberta Floods	10027	5189	4838
West Texas Explosion	10004	5246	4758
Boston Marathon shooting	10012	5648	4364
Queensland Floods	10008	5389	4619
Oklahoma Tornado	10007	5007	5000

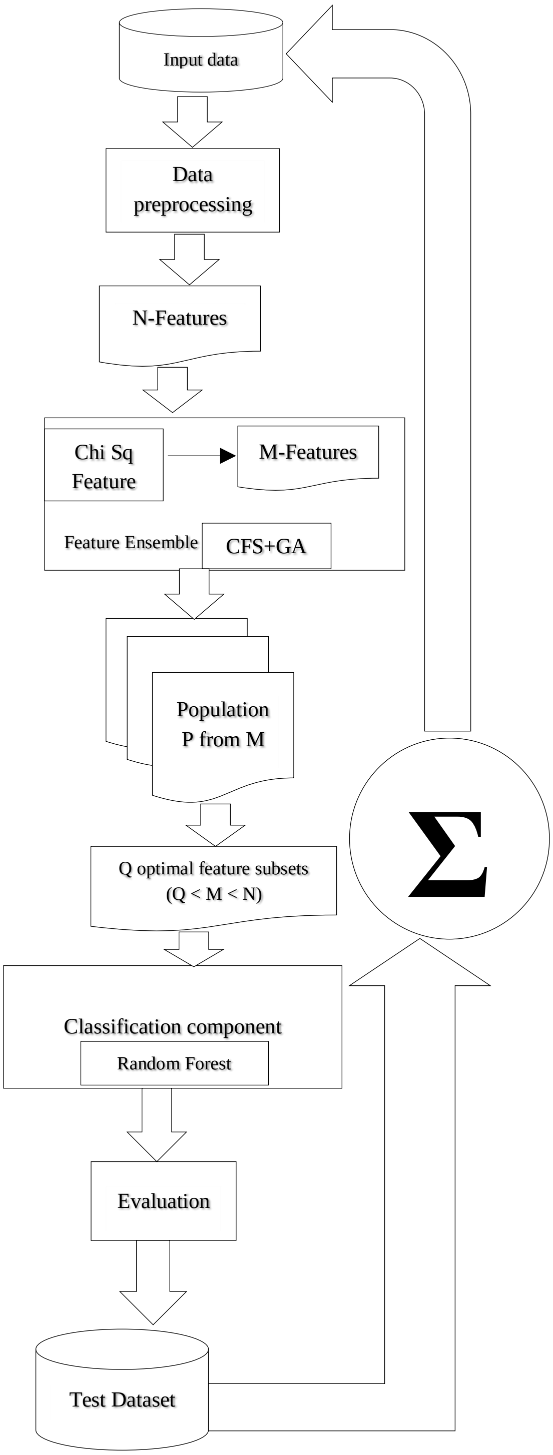

Figure 2. Genetic algorithm based domain adaptation framework functionalities.

3. Methodology

Based upon the GADA functionalities in Figure 2, the methods/steps for classifying disaster-related tweets can be described as below:

1. In total, the proposed framework aims to automatically label 10000 tweet instances from unlabelled target disaster events into "On-topic" and "Off-topic". During the first iteration of the experiment, the dataset from the Labeled Source Disaster (SDL) event is sent as input– initially, 1000 labeled instances from the source disaster event are considered as input, and 1000 unlabelled instances from the target event is considered as input during evaluation.
2. Preprocessing of the data is done on the input SDL data. During preprocessing, firstly, the string data is converted into numeric data as the machine learning classifiers could handle only numeric data (word2vec). Stop words were removed using Rainbow stop words list [30], Word tokenizer is used for tokenization. The minimum term frequency parameter is tuned to choose the frequently occurring terms in the list of instances as features. Following preprocessing stage, N number of features is generated, which is fed as input to the feature ensemble framework.
3. Within the feature ensemble framework, after the N features pass through Chi-Square feature evaluator, M number of features are selected where M<N. Following Chi-Square feature evaluator, the feature transformation is performed using GA. Premature convergence of the Genetic Algorithm while choosing features has been reported as a problem while using GA [1]. We utilize a Correlation-based feature selector to choose the fittest feature, which ensures that the experiment doesn't end prematurely due to convergence during feature transformation using GA [25]. Q number of features are finally given out, which is utilized for the classification of the tweet instances.
4. With the final number of Q features, the classification of tweet instances is performed by the Classification component, where Random Forest is utilized for classification. The number of iterations used within RF is 100.
5. Following the above step, self-training is performed in subsequent iterations. In self-training, the classifier is trained with a portion of labeled data which is used to predict a part of unlabelled data. The predicted labeled data is added together with original labeled data and training is repeated.
6. Following the first iteration during training, unlabelled test data from the target disaster event (TU) is input during evaluation, and the labeled test data comes out as output (TL).
7. During the second iteration of the experiment, TL is added with SDL (TL+SDL) by ensuring that the following condition is met – all the labeled 'on-topic' posts from TL are considered, and an equal number of 'off-topic' posts from TL is considered while adding to SDL.
8. During the third iteration of the experiment, 10000 instances from TU are considered as input. Duplicates are removed in every iteration.
9. Then, the experiment from Step 1 is repeated across n iterations.
10. The termination criteria are when the proposed GADA framework has reached accuracy more than the previous domain adaptation approaches in a particular iteration.

3.1. Experiment Setup

This study uses Waikato Environment for Knowledge Analysis (WEKA) for the classification and separation of actionable posts. WEKA is a collection of machine learning algorithms for data mining tasks [27] suitable for the Text classification task.

As a part of the experiment, datasets from four disaster event pairs are considered as source disaster events and target disaster event pairs, as mentioned in Table 2.

Table 2. Disaster event pairs based on five disaster related datasets mentioned in CrisisLex repository [18].

Pairs	Source Disaster	Target Disaster
BB-WTE	2013_Boston Bombing	2013 West Texas Explosion
QF-AF	2013 Queensland Floods	2013 Alberta Floods
QF-BB	2013 Queensland Floods	2013 Boston Bombing
QF-OKT	2013 Queensland Floods	2013 Oklahama Tornado

3.2. Evaluation Metrics

Performance metrics provide a practical method to check the efficiency of a model. In this experiment, True Positive (TP) and True Negative (TN) are metrics that are referred to the number of correctly classified “on-topic” samples and “off-topic” samples respectively. False Positive (FP) represents the number of “off-topic” instances classified as “on-topic”, while False Negative (FN) represents the number of “on-topic” instances classified as “off-topic”. The parameters TP, TN, FP, and FN can be used to derive some standard metrics, Precision, Recall and F1-Measure, respectively, as shown in Equations (1) to (4).

$$\text{TPR / Recall (R)} = TP / (TP+FN) \tag{1}$$

$$\text{TNR} = TN / (TN+FP) \tag{2}$$

$$\text{Precision (P)} = TP / (TP+FP) \tag{3}$$

$$\text{F1-measure} = 2PR / (P+R) \tag{4}$$

4. Experiment Results

The objective of this experiment is to evaluate the improved domain adaptation technique through the proposed GADA for the task of identifying tweets relevant to the target disaster event (on-topic vs off-topic tweets). We pair the SDL with TU for experiments where TU is the unlabeled instance. Table 2 describes the dataset pairs considered for the experiment. Table 3 to Table 6 describes the Accuracy and Weighted auROC results for the pairs of experiments considered. Figure 3 shows the accuracy plot for BB-WTE pair of datasets derived based upon Table 3 results. Figure 4 illustrates the accuracy (%) plot for QF-AF pair of datasets derived based upon Table 4 results. Figure 5 shows the accuracy (%) plot for QF-BB pair of datasets derived based upon Table 5 results and Figure 6 depicts the accuracy plot of QF-OKT pair of disaster datasets based on Table 6. For each pair of datasets, a domain adaptation experiment is conducted using the proposed GADA framework. The experiments were tested on 3-fold and 5-fold cross validation. The proposed GADA framework achieves the best performance over previous approaches with 3-fold cross-validation itself. Thus, during evaluation 3-fold CV results are compared with prior works.

Table 3. Weighted auROC results and accuracy (%) (over 3-fold & 5-fold Cross Validation) for BB-WTE pair of disaster event using proposed GADA framework across eight iterations (using 1000 instances from Boston Bombing labelled dataset as source).

GADA Iterations (BB-WTE)	Weighted Avg. ROC		Accuracy (%)	
	3-fold CV	5-fold CV	3-fold CV	5-fold CV
Iteration 1	0.797	0.794	62.46	62.48
Iteration 2	0.900	0.91	79.85	79.81
Iteration 3	0.967	0.928	93.71	92.7
Iteration 4	0.971	0.937	94.32	93.7
Iteration 5	0.977	0.953	95.36	95.30
Iteration 6	0.976	0.958	95.74	95.76
Iteration 7	0.982	0.957	95.83	95.71
Iteration 8	0.983	0.960	96.11	96.001

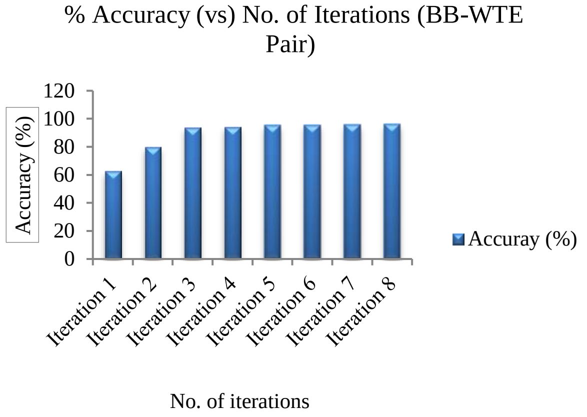

Figure 3. Accuracy (%) plot for BB-WTE pair of datasets derived based upon Table 3 results.

Table 4. Weighted auROC results and accuracy (%) (Over 3-fold & 5-fold CV) for QF-AF pair of disaster event using proposed GADA framework across twelve iterations (using 1000 instances from Queensland Flood labelled dataset as source).

GADA Iterations (QF-AF)	Weighted Avg. ROC		Accuracy (%)	
	3-fold CV	5-fold CV	3-fold CV	5-fold CV
Iteration 1	0.873	0.871	74.38	74.39
Iteration 2	0.823	0.823	72.23	72.29
Iteration 3	0.869	0.869	74.50	74.52
Iteration 4	0.875	0.876	73.81	73.83
Iteration 5	0.831	0.830	73.95	73.96
Iteration 6	0.827	0.828	74.798	74.83
Iteration 7	0.831	0.832	74.58	74.62
Iteration 8	0.856	0.855	74.37	74.35
Iteration 9	0.842	0.843	75.11	75.16
Iteration 10	0.889	0.889	76.13	76.38
Iteration 11	0.896	0.896	81.68	81.69
Iteration 12	0.868	0.868	81.97	81.98

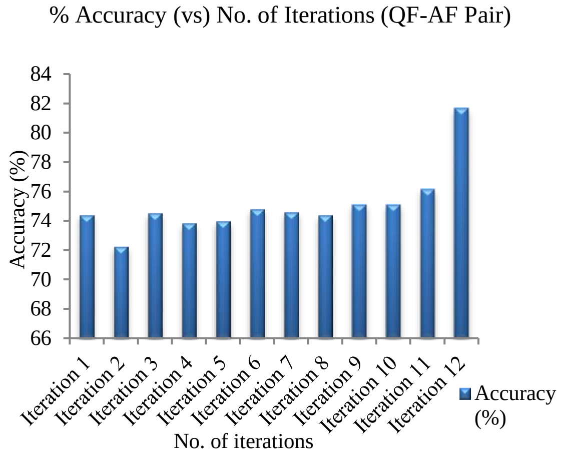

Figure 4. Accuracy (%) plot for QF-AF pair of datasets derived based upon Table 4 results.

Table 5. Weighted auROC results and accuracy (%) (over 3-fold and 5-fold CV) for QF-BB pair of disaster event using proposed GADA framework across eight iterations (using 1000 instances from Queensland Flood labeled dataset as source).

GADA Iterations (QF-BB)	Weighted Avg. ROC		Accuracy (%)	
	3-fold CV	5-fold CV	3-fold CV	5-fold CV
Iteration 1	0.793	0.818	57.76	58.24
Iteration 2	0.761	0.783	68.93	67.38
Iteration 3	0.742	0.745	70.78	70.75
Iteration 4	0.939	0.938	73.61	73.14
Iteration 5	0.880	0.883	76.61	76.43
Iteration 6	0.942	0.941	86.72	86.56
Iteration 7	0.956	0.954	88.96	88.91
Iteration 8	0.948	0.949	90.401	90.41

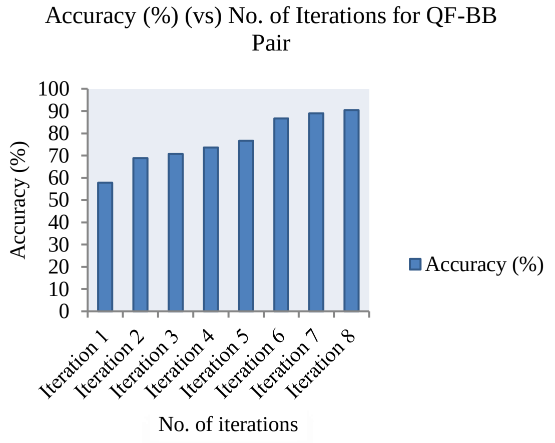

Figure 5. Accuracy (%) plot for QF-BB pair of datasets derived based Table 5 results.

Table 6. Weighted auROC results and accuracy (%) (over 3-fold and 5-fold Cross Validation) for QF-OKT pair of disaster events using proposed GADA framework across seven iterations (using 1000 instances from Queensland Floods labeled dataset as source).

GADA Iterations (QF-OKT)	Weighted Avg. ROC		Accuracy (%)	
	3-fold CV	5-fold CV	3-fold CV	5-fold CV
Iteration 1	0.782	0.821	61.33	62.34
Iteration 2	0.913	0.914	84.06	82.56
Iteration 3	0.913	0.928	87.39	87.18
Iteration 4	0.915	0.936	88.24	88.87
Iteration 5	0.934	0.923	89.15	88.08
Iteration 6	0.941	0.935	89.71	89.71
Iteration 7	0.937	0.933	89.94	88.93

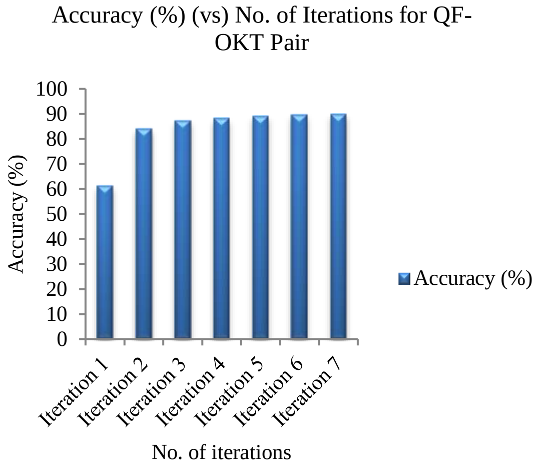

Figure 6. Accuracy (%) plot for QF-OKT pair of datasets derived based upon Table 6 results.

4.1. Evaluation by Comparison of Results with Previous Works

This section compares the experimental results with baselines proposed previously in domain adaptation disaster research. The accuracy (%) is compared between the following approaches-Supervised Naïve Bayes, Supervised Random Forest, Supervised SVM, Supervised Logistic Regression, Naïve Bayes Domain Adaptation with Expectation Maximization, Naïve Bayes Domain Adaptation with Self Training, Random Forest Domain Adaptation with Self Training and the proposed GADA Framework. Tables 7 and 8 compare the accuracy between previous works and GADA.

Table 7. Accuracy (%) comparison between previous works and proposed GADA framework.

Approaches	BB-WTE	QF-AF	QF-BB	QF-OKT
NB-Supervised	94.77	78.87	74.97	84.13
LR-Supervised	87.85	72.06	58.00	79.01
RF-Supervised	92.15	74.49	71.65	81.56
SVM-Supervised	84.29	76.69	65.76	83.97
NB-Domain Adaptation-EM	95.79	82.43	76.69	86.63
NB-Domain Adaptation-ST	94.82	86.01	81.86	85.48
RF-Domain Adaptation-ST	Not Reported	Not Reported	Not Reported	Not Reported
GADA Framework	**96.11**	**81.68**	**90.401**	**89.94**

Table 8. Weighted AUROC comparison between previous works and proposed GADA framework.

Approaches	BB-WTE	QF-AF	QF-BB	QF-OKT
NB-Supervised	0.983	0.860	0.820	0.880
LR-Supervised	0.919	0.714	0.472	0.775
RF-Supervised	0.977	0.860	0.833	0.899
SVM-Supervised	0.835	0.733	0.661	0.824
NB-Domain Adaptation-EM	0.989	0.882	0.832	0.925
NB-Domain Adaptation-ST	0.984`	0.922	0.890	0.924
RF-Domain Adaptation-ST	0.870	0.874	0.875	0.873
GADA Framework	**0.983**	**0.896**	**0.948**	**0.937**

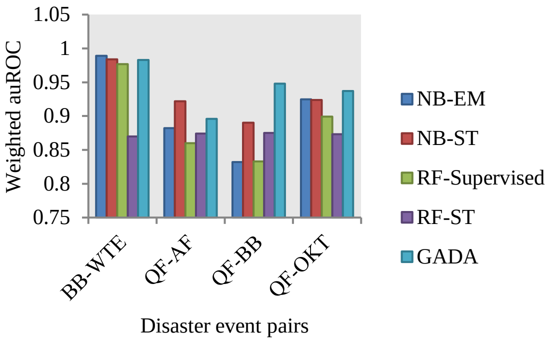

Figure 7. Weighted auROC (averaged over 3-folds) results for the four pairs of disasters (BB-WTE, QF-AF, QF-BB, QF-OKT) using benchmark approaches and GADA approach.

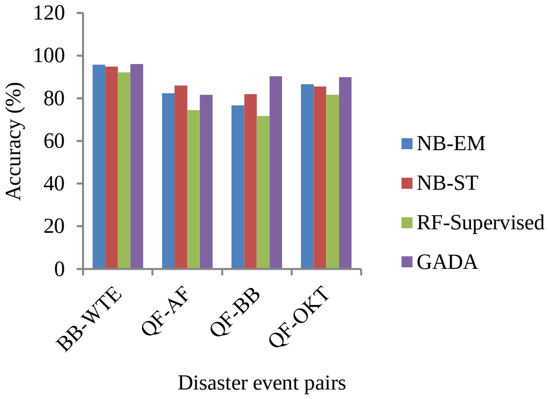

Figure 8. Accuracy (%) (averaged over 3-folds) results for the four pairs of disasters (BB-WTE, QF-AF, QF-BB, QF-OKT) using benchmark approaches and GADA approach.

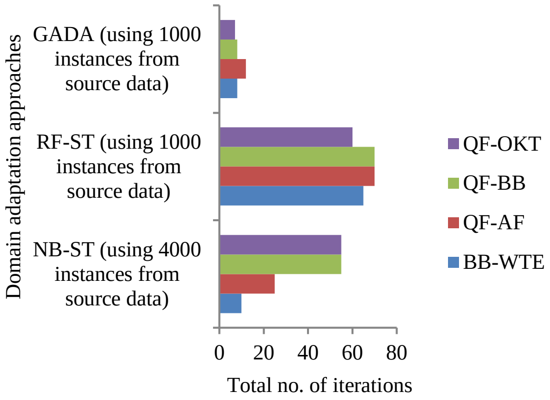

Figure 9. No. of iterations for which the various domain adaptation approaches attain maximum weighted AUROC.

5. Discussion

As shown in Figures 7, 8, and 9, the proposed GADA framework for classification of disaster tweets significantly improves over the baselines (previous supervised approaches and existing domain adaptation approaches) in 86% of the cases for four dataset pairs of experiments from the CrisisLex dataset. Among domain adaptation approaches, in 91% of cases, the number of iterations to achieve the best performance tends to be lower than previous domain adaptation approaches are described in Mazloom *et al*. [16]. In terms of the number of iterations, GADA achieved a 20% to 80% reduction in the number of iterations of the experiment over previous works, which in turn can significantly reduce the overall training time. This will have a significant impact in the immediate aftermath of a disaster event as the target disaster event datasets get classified with the highest accuracy much faster and help the disaster response team to act faster and provide support to the disaster victims much more efficiently. Unlike previous works [14, 15, 16], in this study the "Re-Tweet" or "RT" feature is not discarded. Our experiments show that utilizing 'RT' feature without discarding helps significantly in transfer learning from one domain to another, especially in the initial stages of the experiment. This results in a significant reduction in the number of iterations in the experiment. The use of GA helps in increasing or reducing the features by tuning GA parameters, thereby reducing the training time of iterations.

6. Conclusions

In the aftermath of a disaster event, automated approaches to separate useful disaster-related social media posts offer a better solution for emergency teams and disaster victims for help and support. While the accuracy of classification of social media posts is important, what is more important is quick response and support. It can be hard to have human label data while a disaster event is happening.

To achieve reasonably good performance of classification and also without the need for laborious human labelling, the domain adaptation approach developed in previous works is useful. This study improved the previous domain adaptation approaches by proposing a Genetic Algorithm for feature selection GADA during domain adaptation. It is expected that the proposed GADA framework will improve the performance of existing approaches and achieve the best performance.

The proposed GADA framework significantly improves upon the supervised approach by offering superlative performance improvement of accuracy. Over the existing domain adaptation approaches, GADA helps in reducing the training time of experiments through robust feature engineering. One of the limitations of the proposed framework is that it is not tested on datasets other than English language datasets. Also, the GADA has experimented in binary classification settings and not on a multi-label setting. Thus, as future work, GADA can be improved by experimenting in a multi-label classification setting and can be expanded to other language datasets.

References

[1] Andre J., Siarry P., and Dognon T., "An Improvement of the Standard Genetic Algorithm Fighting Premature Convergence in Continuous Optimization," *Advances in Engineering Software*, vol. 32, no.1, pp.49-60, 2001.

[2] Babatunde O., Armstrong L., Leng L., and Diepeveen D., "A Genetic Algorithm-Based Feature Selection," *International Journal of Electronics Communication*

and Computer Engineeringm, vol. 5, no. 4, pp. 899-905, 2014.

[3] Bermejo P., Gámez J., and Puerta J., "Speeding up Incremental Wrapper Feature Subset Selection with Naive Bayes Classifier," *Knowledge-Based Systems*, vol. 55, pp. 140-147, 2014.

[4] Blum A. and Mitchell T., "Combining Labelled and Unlabelled Data with Co-Training," *in Proceedings of the 11th annual conference on Computational Learning Theory*, Madison, pp. 92-100, 1998.

[5] Catak F. and Bilgem T., "Genetic Algorithm Based Feature Selection in High Dimensional Text Dataset Classification," *WSEAS Transactions on Information Science and Applications*, vol. 12, no. 28, pp. 290-296, 2015.

[6] Chinnaiah V. and Kiliroor C., "Heterogeneous Feature Analysis on Twitter Data Set for Identification of Spam Messages," *The International Arab Journal of Information Technology*, vol. 19, no. 1, pp. 38-44, 2022.

[7] Günal S., "Hybrid Feature Selection for Text Classification," *Turkish Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, pp. 1296-1311, 2012.

[8] Hassanat A., Almohammadi K., Alkafaween E., Abunawas E., Hammouri A., and Prasath V., "Choosing Mutation And Crossover Ratios for Genetic Algorithms-A Review with A New Dynamic Approach," *Information*, vol. 10, no. 12, pp. 390, 2019.

[9] Huang C. and Dun J., "A Distributed PSO-SVM Hybrid System with Feature Selection and Parameter Optimization," *Applied Soft Computing*, vol. 8, no. 4, pp.1381-1391, 2008.

[10] Jin X., Xu A., Bie R., and Guo P., "Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles," *in Proceedings of the International Workshop on Data Mining for Biomedical Applications*, Singapore, pp. 106-115, 2006.

[11] Lanzi P., "Fast Feature Selection with Genetic Algorithms: A Filter Approach," *in Proceedings of the IEEE International Conference on Evolutionary Computation*, Indianapolis, pp. 537-540, 1997.

[12] Leardi R., "Application of a Genetic Algorithm to Feature Selection under Full Validation Conditions and to Outlier Detection," *Journal of Chemometrics*, vol. 8, no. 1, pp. 65-79, 1994.

[13] Li H., Guevara N., Herndon N., Caragea D., Neppalli K., Caragea C., Squicciarini A., Tapia A., "Twitter Mining for Disaster Response: A Domain Adaptation Approach," *in Proceedings ISCRAM*, Krystiansand, 2015.

[14] Li H., Caragea D., Caragea C., and Herndon N., "Disaster Response Aided By Tweet Classification With A Domain Adaptation Approach," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 16-27, 2018.

[15] Li X. and Caragea D., "Domain Adaptation with Reconstruction for Disaster Tweet Classification," *in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, China, pp. 1561-1564, 2020.

[16] Mazloom R., Li H., Caragea D., Caragea C., and Imran M., "A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets," *International Journal of Information Systems for Crisis Response and Management*, vol. 11, no. 2, pp. 1-19, 2019.

[17] Mohammed T. Bayat O., Ucan O., and Alhyali S., "Hybrid Efficient Genetic Algorithm for Big Data Feature Selection Problems," *Foundations of Science*, vol. 25, no. 21, pp. 1-17, 2019.

[18] Olteanu A., Castillo C., Diaz F., and Vieweg S., "Crisislex: A lexicon for Collecting and Filtering Microblogged Communications in Crises," *in Proceedings of the 8th International AAAI Conference on Web and Social Media*, Ann Arbor, 2014.

[19] Pan S. and Yang Q., "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp.1345-1359, 2010.

[20] Panchal G. and Panchal D., "Solving np Hard Problems Using Genetic Algorithm," *Transportation*, vol.106, pp. 6-2, 2015.

[21] Parilla-Ferrer B., Fernandez P., BallenaIV J., "Automatic Classification of Disaster-Related Tweets," *in Proceedings of the International conference on Innovative Engineering Technologies*, Bangkok, pp. 62-69, 2014.

[22] Rudra K. Ghosh S., Ganguly N., Goyal P., "Extracting Situational Information from Microblogs during Disaster Events: A Classification-Summarization Approach," *in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Melbourne, pp. 583-592, 2015.

[23] Schulz A., Guckelsberger C., and Janssen F., "Semantic Abstraction for Generalization of Tweet Classification: An Evaluation of Incident-Related Tweets," *Semantic Web*, vol. 8, no. 3, pp. 353-372, 2017.

[24] Stowe K., Paul M., Palmer M., Palen L., and Anderson K., "Identifying and Categorizing Disaster-Related Tweets," *in Proceedings of the 4th International Workshop on Natural Language Processing for Social Media*, Austin, pp. 1-6, 2016.

[25] Tiwari R. and Singh M., "Correlation-based Attribute Selection Using Genetic Algorithm," *International Journal of Computer Applications*, vol. 4, no. 8, pp. 28-

34, 2010.

[26] Umbarkar A. and Sheth P., "Crossover Operators in Genetic Algorithms: A Review," *ICTACT Journal on Soft Computing*, vol. 6, no. 1, 2015.

[27] Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Elsvier, 2005.

[28] Yarowsky D., "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *in Proceedings 33rd Annual Meeting of The Association for Computational Linguistics*, USA, pp. 189-196, 1995.

[29] Zhai Y., Song W., Liu X., Liu L., and Zhao X., "A chi-Square Statistics Based Feature Selection Method in Text Classification," *in Proceedings of the IEEE 9th International Conference on Software Engineering and Service Science*, Beijing, pp. 160-163, 2018.

[30] Zhou Y., De S., and Moessner K., "Real World City Event Extraction from Twitter Data Streams," *Procedia Computer Science*, pp. 443-448, 2016.

Lokabhiram Dwarakanath received the B.Engg degree in Electronics and Communication Engineering from Dr.MGR Engg College, University of Madras, India, and the M.Sc. degree in Enterprise Business Systems from the Brunel University, West London, U.K. He is currently pursuing the Ph.D. degree in the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. His research interests include data science, natural language processing, information systems, cloud computing, machine learning, big data, and social media analytics.

Amirrudin Kamsin is a Senior Lecturer at the Faculty of Computer Science and Information Technology, and the Acting Director and Deputy Director (ODL and Professional Programme) at the University of Malaya Centre for Continuing Education (UMCCed), University of Malaya, Malaysia. He received his BIT (Management) in 2001 and MSc in Computer Animation in 2002 from University of Malaya and Bournemouth University, UK respectively. He obtained his PhD in Computer Science from University College London (UCL) in 2014. His research areas include human-computer interaction (HCI), authentication systems, e-learning, mobile applications, serious game, augmented reality and mobile health services.

Liyana Shuib obtained her Master of Information System (Data Mining) from Universiti Kebangsaan Malaysia in 2005 and a Ph.D. from the University of Malaya, Malaysia in 2013 respectively. She is an Associate Professor at the Department of Information Systems, Faculty of Computer Science & Information Technology and the Deputy Director of Analytics at Academic Strategic Planning Centre, Deputy Vice Chancellor (Academic & International), University of Malaya, Malaysia. She has published a number of journal papers and proceedings locally and internationally. Her research interests include personalization, e-learning, recommender system, data science, data mining, artificial intelligence application, and educational technology. She has won more than 20 awards from reputable innovation competition internationally. She is also a senior member of IEEE computing society, an active blogger and presently, the principal investigator of multiple research grant in the Faculty.