Unsupervised and Interpretable Domain Adaptation to Rapidly Filter Tweets for Emergency Services

Jitin Krishnan  
Department of Computer Science  
George Mason University  
Fairfax, VA, USA  
jkrishn2@gmu.edu

Hemant Purohit  
Department of Information Sciences & Technology  
George Mason University  
Fairfax, VA, USA  
hpurohit@gmu.edu

Huzefa Rangwala  
Department of Computer Science  
George Mason University  
Fairfax, VA, USA  
rangwala@gmu.edu

Abstract—During the onset of a natural or man-made crisis event, public often share relevant information for emergency services on social web platforms such as Twitter. However, filtering such relevant data in real-time at scale using social media mining is challenging due to the short noisy text, sparse availability of relevant data, and also, practical limitations in collecting large labeled data during an ongoing event. In this paper, we hypothesize that unsupervised domain adaptation through multi-task learning can be a useful framework to leverage data from past crisis events for training efficient information filtering models during the sudden onset of a new crisis. We present a novel method to classify relevant social posts during an ongoing crisis without seeing any new data from this event (fully unsupervised domain adaptation). Specifically, we construct a customized multi-task architecture with a multi-domain discriminator for crisis analytics: multi-task domain adversarial attention network (MT-DAAN). This model consists of dedicated attention layers for each task to provide model interpretability; critical for real-word applications. As deep networks struggle with sparse datasets, we show that this can be improved by sharing a base layer for multi-task learning and domain adversarial training. The framework is validated with the public datasets of TREC incident streams that provide labeled Twitter posts (tweets) with relevant classes (Priority, Factoid, Sentiment) across 10 different crisis events such as floods and earthquakes. Evaluation of domain adaptation for crisis events is performed by choosing one target event as the test set and training on the rest. Our results show that the multi-task model outperformed its single-task counterpart. For the qualitative evaluation of interpretability, we show that the attention layer can be used as a guide to explain the model predictions and empower emergency services for exploring accountability of the model, by showing the words in a tweet that are deemed important in the classification process. Finally, we show a practical implication of our work by providing a use-case for the COVID-19 pandemic.

Index Terms—Social Media, Crisis Analytics, Text Classification, Unsupervised Domain Adaptation, Interpretability

I. INTRODUCTION

During the sudden onset of a crisis situation, social media platforms such as Twitter provide valuable information to aid crisis response organizations in gaining real-time situational awareness [1]. Effective analysis of important information such as affected individuals, infrastructure damage, medical emergencies, or food and shelter needs can help emergency responders make time-critical decisions and allocate resources in the most effective manner [2]–[5].

Several machine learning systems have been deployed to help towards this humanitarian goal of converting real-time social media streams into actionable knowledge. Classification being the most common task, researchers have designed models [4], [9]–[10] that classify tweets into various crisis-dependent categories such as priority, affected individuals, type of damage, type of assistance needed, usefulness of the tweet, etc. Social media streams contain short, informal, and abbreviated content; with potential linguistic errors and sometimes contextually ambiguous. These inherently challenging properties of tweets make their classification task and formulation less trivial when compared to traditional text classification tasks. In this paper, we address two practically important and under-developed aspects of current research in social media mining for crisis analytics to classify relevant social web posts: a) a fully unsupervised domain adaptation, and b) interpretability of predictions. A fully unsupervised domain adaptation uses no data from the ongoing crisis to train the model. Nguyen et al., 2016 [6] showed that their convolutional neural network (CNN) model does not require feature engineering and performed better than the state-of-the-art methods; one of their...
models being completely unsupervised [6]. Similarly, Alam et al., 2018 [7] designed a CNN architecture with adversarial training on graph embeddings, but utilizing unlabeled target data. Our goal is to construct an unsupervised model that does not require any unlabeled target data with the capability of being interpretable. We specifically address the problem of data sparsity and limited labels by designing a multi-task classification model with domain adversarial training; which, to the best of our knowledge, is not explored in social media mining for crisis analytics. Another crucial component of our model is interpretability. In prior works, when a top performing model produces an accuracy of 78%, for instance, it is unclear how trustworthy it is and what features are deemed important in the model’s decision-making process. An interpretable model like ours can present with a convincing evidence of which words the classifier deems important when making a certain prediction, and helps ensure reliability for domain users, e.g., emergency managers.

**Contributions:**

a) To address the problems of data sparsity and limited labels, we construct a customized multi-task learning architecture (MT-DAAN) to filter tweets for crisis analytics by training four different classification tasks (c.f. examples in Fig. 3) across ten different crisis events under domain shift. This multi-task domain adversarial model consists of dedicated attention layers for each task for interpretability and a domain classifier branch to promote the model to be domain-agnostic.  

b) We demonstrate that the attention layers provide interpretability for the predictions made by the classifiers; with the goal to aid emergency services in a more meaningful way.  
c) We empirically validate the performance of the underlying single-task attention-based neural network architecture by comparing it to the state-of-the-art methods, for improving generalizability and interpretability for domain adaptation in unsupervised tweet classification tasks in general.  
d) Additionally, through experiments, we show that deep networks struggle with small datasets, and that this can be improved by sharing the base layer for multi-task learning and domain adversarial training.

II. RELATED WORK AND BACKGROUND

A. Domain Adaptation

Domain Adaptation in text classification tasks has a long line of fruitful research [11]–[13] that try to minimize the difference between the domains so that a model trained solely on one domain is generalizable to unseen test data from a completely different domain. With the introduction of Domain-Adversarial training of Neural Networks (DANN) [14], many state-of-the-art models now utilize unlabeled target data to train classifiers that are indiscriminate toward different domains. The speciality of this architecture is that it consists of an extra branch, which performs domain classification using unlabeled data from different domains. Thus, both task and domain classifiers share some bottom layers but have separate layers towards the top. A negative gradient from the domain classifier branch is back-propagated to promote the features at the lower layers of the network incapable of discriminating the domains. Recent works such as Adversarial Memory Network (AMN) [15] utilizes attention, in addition to DANN, to bring interpretability to capture the pivots for sentiment classification. Hierarchical Attention Network (HATN) [16] expands upon AMN by first extracting pivots and then jointly training networks for both pivots and non-pivots.

For filtering social web data for crisis analytics, these models do not suffice and need customized expansions due to the following reasons:  
a) Collecting and using large unlabeled target data from the new/ongoing crisis event may not be practically viable, thus, we aim for a fully unsupervised modeling.  
b) Having access to unlabeled data from multiple crisis events can alleviate the above problem to an extent by using it to train the domain classifier branch to push the model to be domain independent.  
c) Due to the low-resource nature of the dataset, binary classifiers may miss important lower level features that can be potentially improved by a multi-task model that shares the lower layers of the network for all the tasks. This is also evident from our results in Table III and [14] which show that deep models that perform much better than simple models on Amazon reviews do not significantly outperform them on TREC tweet dataset for crises.

B. Multi-Task Learning

Multi-Task Learning (MTL) solves multiple tasks at the same time with a goal to improve the overall generalization capability of the model [17]. Within the context of Deep Learning, MTL is performed by sharing (or constraining) lower level layers and using dedicated upper level layers for various tasks. A rich overview of MTL in Deep Neural Networks is presented by Ruder (2017) [18]. MTL has been a successful strategy over the past few years for many research explorations such as relationship networks [19] in computer vision and Sluice networks [20] in natural language processing. Similar problems in domain adaptation of semantic classification and information retrieval were addressed by jointly learning to leverage large amounts of cross-task data [21]. In low resource datasets such as for crises, the chance of overfitting is very high. Thus, it seems intuitively better for the model to find a shared representation capturing different tasks and not just one, such that feature commonalities across tasks can be exploited.

C. Attention Mechanism

Attention mechanism [22], originally designed for machine translation problems, has become one of the most successful and widely used methods in deep learning that can look at a part of a sentence at a time like humans. This is particularly useful because of its ability to construct a context vector by weighing on the entire input sequence unlike previous sequence-to-sequence models [23] that used only the last hidden state of the encoder network (typically BiLSTM [24], LSTM [25], or GRU [26]). For example, in a sentence, the context vector is a dot product of the word activations and weights associated with each word; thus leading to an improved contextual memorization, especially for long sen-
tences. Our method incorporates such attention mechanisms to enhance interpretability of the classifier.

![Diagram](image_url)

**Fig. 2.** Fully Unsupervised Domain Adaptation Set-up for Multi-Task Crisis Tweet Classification.

### III. METHODOLOGY

#### A. Problem Statement: Unsupervised Domain Adaptation for Crisis Tweet Classification

Using notations in Table I, consider a set $C$ of all crisis events such as Guatemala Earthquake or Typhoon Yolanda. The task of unsupervised domain adaptation for crisis analytics is to train a classifier for a specific target crisis ($c_t$) using labeled ($L_{C-c_t}$) and unlabeled ($U_{C-c_t}$) data from all other crises, where $C - c_t$ denotes the set of all crisis events minus the target crisis. We assume that no data record from the target crisis is available for training. Following the traditional domain adaptation terminology, $X_s = L_{C-c_t}$ represents the labeled data from the source domain $S$ and $Y_s = y_{C-c_t}$ represents the ground truth labels on which the classifier is trained. And, $X_t = L_{c_t}$ represents the labeled data from the target domain $T$ and $Y_t = y_{c_t}$ represents the ground truth labels; both of which are only used for testing the classifier. $X_d = U_{C-c_t}$ represents the unlabeled data from different domains minus the target. To summarize:

**Input:** $X_s, Y_s, X_d$

**Output:** $Y_{t,pred} \leftarrow predict(X_t)$

#### B. Overview

In the following sections, we describe three models: Single-Task Attention Network (ST), Single-Task Domain Adversarial Attention Network (ST-DAAN), and Multi-Task Domain Adversarial Attention Network (MT-DAAN). ST is the model we adopt from [27] to build the single-task attention based baseline. ST-DAAN is constructed on top of ST to make the model domain agnostic by performing adversarial training using gradient reversal. Finally, MT-DAAN is constructed on top of ST-DAAN with dedicated attention layers for each task on a shared BiLSTM layer. This is shown in Figure 2.

### C. Single-Task Attention Network (ST)

We first describe the single-task attention network [27] on top of which we build our models. This model aligns with our goals of interpretability and unsupervised domain adaptation. This BiLSTM based model with Attention gives us three main advantages:

1. Unlike several existing domain adaptation methods that use unlabeled target data to train the domain adversarial component via gradient reversal, this method is a fully unsupervised baseline which also can be customized for multi-task learning.
2. The method uses attention mechanism which in turn weighs each word in a sentence based on its importance. This can be directly utilized for interpretability.
3. The method also runs much faster (only a few minutes), i.e. highly useful in crisis times, as compared to the top performing semi-supervised models such as HATN [16] (hours).

This model [27] consists of a BiLSTM layer which produces $T_x$ activations, each corresponding to a word in the sentence. These activations are passed through dense and softmax layers and are combined by dot product to produce the context vector $\sum_{k=1}^{T_x} a^{<k>} a^{<k>}$, where $a^{<k>}$ is the BiLSTM activation from $k$-th word and $a^{<k>}$ is the attention weight of $k$-th word. Sentences with words greater than $T_x$ are stripped and those with words lower than $T_x$ are padded. This single-task ($m = 1$) attention network is the building block with which rest of the following models are constructed. The single-task binary cross entropy loss function is shown below.

$$L_T = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$  \hspace{1cm} (1)

where $T$ represents the task, $y$ is the true label, and $\hat{y}$ is the predicted label.

#### D. Single-Task Domain Adversarial Attention Network (ST-DAAN)

To study the specific contribution of domain adversarial training, we construct a secondary baseline over the ST architecture by constructing an additional branch with gradient reversal layer which is represented by the green blocks in Figure 2. This is a single-task binary classifier with $m = 1$. Domain Adversarial Training of Neural Networks (DANN) [14] was introduced with a goal to confuse the classifier by back-propagating a negative gradient from a separate domain clas-
sifier branch (right-most branch, as shown in Figure 2). This makes the classifier agnostic to difference in domains. This back-propagation is implemented using a gradient reversal layer [14] which does nothing during the forward pass but pushes a negative gradient (−λ ∂L/∂f) during the backward (gradient update) pass. Ld is the domain classification loss. λ is the strength of the reversal, and f represents the lower level layers or features over which the negative gradient update is performed. In our architecture, the goal is to make the BiLSTM layer indiscriminate towards various crisis domains such that the multi-task classification does not depend on the domain from which the tweet/sentence is coming from. The ST-DAAN loss function is shown below.

$$L'_F = L_F + w_dL_d$$

where $w_d$ is the domain adversarial loss weight. $L_d$ represents the categorical cross entropy loss for multi-domain discriminator shown below.

$$L_d = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C-c_t} [y_{ij} \log y_{ij}]$$

where $C - c_t$ is the set of all crisis events without the target event.

E. Multi-Task Domain Adversarial Attention Network (MT-DAAN)

Building on top of ST-DAAN, we construct MT-DAAN, which is intended to classify problems with multiple tasks or labels. For each task, a dedicated attention layer is allocated from which it predicts binary labels. The BiLSTM layer remains exactly the same as in the single-task model but multiple attention blocks are added for each task along with a domain classifier. In the architecture decision process, we first investigated a multi-label classifier where all layers are shared with the final softmax layer making multi-label predictions. In low resource settings, constructing a multi-label classifier using a shared architecture is challenging for two reasons: a) joint balancing positive and negative samples across all classes is not trivial and potentially challenging to make it extensible when new classes need to be added, and b) attention layer may not always produce class-specific insights as the weights are assigned to train for the combination of labels. On the other hand, in the multi-task architecture with separate attention layers, it is easy to add more classes. If some classes require more training, it is trivial to further tune a model specific to that class. More importantly, context<tk> vector for j-th task identifies the influential words from each sentence for that specific task. The complete architecture is shown in Figure 2. MT-DAAN loss function is shown below:

$$L_{MT-DAAN} = \sum_{k=1}^{m} (wKL_{Tk} + w_dL_d)$$

where m is the number of tasks, $w_k$ is the loss weight and $L_{Tk}$ is the loss term for each task, $w_d$ is the domain adversarial loss weight, and $L_d$ is the domain adversarial loss term.

---

**TABLE II**

TREC DATASET STATISTICS; SHOWING THE NUMBER OF POSITIVE SAMPLES FOR EACH OF THE 4 CLASSES. P=Priority, F=Factoid, S=Sentiment, AND I=Irrelevant.

<table>
<thead>
<tr>
<th>CRISIS EVENTS</th>
<th>Total Tweets</th>
<th>Vocab</th>
<th>Avg #words</th>
<th>P</th>
<th>F</th>
<th>S</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012 Guatemala Earthquake</td>
<td>154</td>
<td>422</td>
<td>18.74</td>
<td>104</td>
<td>108</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>2013 Typhoon Yolanda</td>
<td>564</td>
<td>1746</td>
<td>19.47</td>
<td>249</td>
<td>46</td>
<td>119</td>
<td>51</td>
</tr>
<tr>
<td>2013 Australia Bushfire</td>
<td>677</td>
<td>2102</td>
<td>20.21</td>
<td>152</td>
<td>213</td>
<td>167</td>
<td>36</td>
</tr>
<tr>
<td>2013 Boston Bombings</td>
<td>535</td>
<td>1755</td>
<td>19.30</td>
<td>117</td>
<td>28</td>
<td>234</td>
<td>198</td>
</tr>
<tr>
<td>2013 Queensland Floods</td>
<td>713</td>
<td>2301</td>
<td>19.08</td>
<td>233</td>
<td>54</td>
<td>173</td>
<td>215</td>
</tr>
<tr>
<td>2014 Chile Earthquake</td>
<td>311</td>
<td>919</td>
<td>16.54</td>
<td>48</td>
<td>26</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>2014 Typhoon Hagupit</td>
<td>1470</td>
<td>2893</td>
<td>15.36</td>
<td>469</td>
<td>375</td>
<td>741</td>
<td>133</td>
</tr>
<tr>
<td>2015 Nepal Earthquake</td>
<td>2048</td>
<td>564</td>
<td>20.06</td>
<td>329</td>
<td>64</td>
<td>206</td>
<td>70</td>
</tr>
<tr>
<td>2015 Paris Attacks</td>
<td>2066</td>
<td>4152</td>
<td>18.62</td>
<td>306</td>
<td>183</td>
<td>782</td>
<td>429</td>
</tr>
<tr>
<td>2018 Florida School Shooting</td>
<td>1118</td>
<td>2940</td>
<td>21.40</td>
<td>329</td>
<td>64</td>
<td>206</td>
<td>70</td>
</tr>
</tbody>
</table>

---

F. Model Interpretability

The output (α) of the attention layer (ATT) of each task, is a $T_x$-dimensional vector; $T_x$ being the number of words in the sentence. The context vector ($\sum_{k=1}^{T_x} \alpha^{<k>}$) is the product of these attention weights and the $T_x$-dimensional activation (α) from the BiLSTM layer. α essentially weights how much each word in the sentence contributes to the classification result. Thus, α is the component that is evaluated for model interpretability.

IV. DATASETS

A. TREC Dataset

TREC-IS [1] (Text Retrieval Conference - Incident Streams) is a program that encourages research in information retrieval from social media posts with the goal to improve the state-of-the-art social media based crisis analytics solutions. We use the dataset from 2018 track proposal. Statistics of this curated dataset of Twitter downloaded from TREC is shown in Table II. The original dataset consisted of 15 crisis events. However, due to very low data, we trimmed the events and tasks such that there are at least 10 positive samples for each task.

The four tasks used in our experiments are shown below:

1) **Priority**: Different priority levels are assigned for each tweet: *low, medium, high, critical*. We convert this into a binary classification problem where *low* = 0 and {*medium, high, critical*} = 1.

2) **Factoid**: ‘Factoid’ is a categorical label that represents if a tweet is stating a fact. E.g. ‘death toll rises ..’

3) **Sentiment**: ‘Sentiment’ is a categorical label that represents if a tweet represents a sentiment. E.g. ‘Worried.. Thoughts and prayers.’

4) **Irrelevant**: ‘Irrelevant’ is a categorical label for tweets that do not provide any relevant information.

B. Amazon Reviews Dataset

The standard benchmark dataset of Amazon reviews [28] is widely used for cross-domain sentiment analysis. We chose four domains: Books (B), Kitchen (K), DVD (D), and Electronics (E). The raw data [18] part of Blitzer’s original raw dataset, used in this work is from HATN [16]. This dataset

http://dcs.gla.ac.uk/~richardm/TREC/IS/
http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
https://github.com/hssqml/HATN/tree/master/raw_data
<table>
<thead>
<tr>
<th>S → T</th>
<th>LR</th>
<th>SVM</th>
<th>CNN</th>
<th>BiLSTM</th>
<th>AMN</th>
<th>HATN</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>B → K</td>
<td>76.40</td>
<td>75.95</td>
<td>81.20</td>
<td>84.45</td>
<td>81.88</td>
<td>87.03</td>
<td>81.22</td>
</tr>
<tr>
<td>B → E</td>
<td>75.53</td>
<td>74.05</td>
<td>80.44</td>
<td>84.61</td>
<td>80.55</td>
<td>85.75</td>
<td>85.51</td>
</tr>
<tr>
<td>B → D</td>
<td>81.08</td>
<td>81.43</td>
<td>82.94</td>
<td>83.52</td>
<td>85.62</td>
<td>87.07</td>
<td>86.32</td>
</tr>
<tr>
<td>K → B</td>
<td>76.12</td>
<td>75.78</td>
<td>78.78</td>
<td>80.67</td>
<td>79.05</td>
<td>84.88</td>
<td>81.85</td>
</tr>
<tr>
<td>K → E</td>
<td>80.37</td>
<td>81.20</td>
<td>85.17</td>
<td>87.37</td>
<td>86.68</td>
<td>89.00</td>
<td>87.09</td>
</tr>
<tr>
<td>K → D</td>
<td>73.72</td>
<td>74.98</td>
<td>76.41</td>
<td>78.49</td>
<td>79.50</td>
<td>84.72</td>
<td>81.13</td>
</tr>
<tr>
<td>E → B</td>
<td>74.85</td>
<td>74.18</td>
<td>78.08</td>
<td>80.18</td>
<td>77.52</td>
<td>84.03</td>
<td>81.50</td>
</tr>
<tr>
<td>E → K</td>
<td>81.85</td>
<td>81.85</td>
<td>86.59</td>
<td>89.00</td>
<td>87.83</td>
<td>90.08</td>
<td>89.21</td>
</tr>
<tr>
<td>E → D</td>
<td>75.82</td>
<td>75.83</td>
<td>78.35</td>
<td>84.86</td>
<td>85.05</td>
<td>84.32</td>
<td>81.37</td>
</tr>
<tr>
<td>D → B</td>
<td>81.17</td>
<td>82.20</td>
<td>82.26</td>
<td>84.83</td>
<td>84.53</td>
<td>87.78</td>
<td>87.02</td>
</tr>
<tr>
<td>D → K</td>
<td>76.42</td>
<td>77.58</td>
<td>81.09</td>
<td>85.21</td>
<td>81.67</td>
<td>87.47</td>
<td>86.37</td>
</tr>
<tr>
<td>D → E</td>
<td>72.74</td>
<td>73.68</td>
<td>79.56</td>
<td>83.66</td>
<td>80.42</td>
<td>86.32</td>
<td>85.63</td>
</tr>
<tr>
<td>AVG</td>
<td>77.12</td>
<td>79.39</td>
<td>80.91</td>
<td>83.45</td>
<td>82.54</td>
<td>86.84</td>
<td>85.02</td>
</tr>
</tbody>
</table>

TABLE III

Performance comparison (accuracy) of various models on the standard benchmark dataset of Amazon reviews. Methods in blue do not use any unlabeled target data; hence relevant in our context. Each reported score is an average of 10 independent runs of each experiment.

C. COVID-19 Tweet Dataset

For the COVID-19 use-case, we use Twitter posts collected using CitizenHelper [29] system in March 2020, for the geobounding box of the Washington D.C. Metro region. These tweets were annotated by volunteers of regional Community Emergency Response Teams (CERTs), with ‘Relevant’ label denoting how relevant a tweet is for crisis response operations. The label values range on a scale of −4 to +4. We convert them into binary classes by considering values 1 and 2 as −ve (0) class and values 3 and 4 as +ve (1) class. This dataset consists of 4911 tweets with −ve (Relevant=0) and 637 tweets with +ve (Relevant=1) classes. Following unsupervised domain adaptation criteria, the filtering models are trained using only the TREC dataset and evaluated on the COVID-19 tweets. For each independent run of the experiment, a balanced subset of size 637 for both classes is selected for testing.

V. RESULTS & DISCUSSION

We first validate the performance of the adopted unsupervised ST model [27] by comparing it with the following standard neural network architectures and state-of-the-art models used for domain adaption in text. We use the standard benchmark dataset of Amazon reviews. Following the traditional domain adaptation experimental setup, each experiment represented as S → T consists of a source domain (S) on which the model is trained and a target domain (T) on which the model is tested. We use Keras deep learning library for our implementations; with \( T_{c}=200 \) for Amazon reviews and 30 for Tweets. We use Adam optimizer with a dropout of 0.4, maximum epoch of 50, early stopping patience of 3, batch size of 32, and validation split of 0.15.

1) Simple Baselines: We construct simple baseline classifiers [30]: Logistic Regression (LR) and Support Vector Machines (SVM). The input to these models are constructed by aggregating the 300-dimensional word embeddings of words in each review.

2) CNN: A standard Convolutional Neural Network inspired by Kim, 2014 [31] is constructed with the following architecture:

\[
\text{Word Embeddings}(T_{c}, 300) \rightarrow \text{Conv1D}(128, 5) \rightarrow \text{MaxPooling1D}(5) \rightarrow \text{Conv1D}(128, 5) \rightarrow \text{MaxPooling1D}(5) \rightarrow \text{Conv1D}(128, 5) \rightarrow \text{GlobalMaxPooling1D}() \rightarrow \text{Dense}(128) \rightarrow \text{Dense}(2) \rightarrow y
\]

This is combined with dropouts, relu activations, and ending with softmax activation producing labels for binary classification. State-of-the-art deep learning methods for existing social media mining approaches of crisis analytics [6, 7] use a similar architecture.

3) BiLSTM: This is the bottom-most layer in Figure 2 with the activation \( a_{<T_{c}>} \) passed through the following:

\[
\text{Dense}(10) \rightarrow \text{Dense}(2) \rightarrow y \text{ also including dropouts, relu activation, and ending with softmax.}
\]

4) AMN and HATN: AMN [15] and HATN [16] are attention-based methods which use gradient reversal to perform domain adversarial training on the unlabeled data from source and target domains. HATN is an extension to AMN by adding the hierarchical component and jointly training pivot and non-pivot networks.

Input to all the models are word vectors [32]. The evaluation on Amazon reviews shows how well the single-task (ST) model perform when compared to the existing top-performing domain adaptation models on benchmark dataset. Table III shows accuracy scores on the Amazon cross-domain sentiment analysis dataset. HATN uses unlabeled target data, gradient reversal, explicit pivot extraction, and joint training making it a computationally expensive method. As shown in the experimental evaluation, we use the same Amazon dataset and GoogleNews word vectors for our experiments. ST, being unsupervised with no need of unlabeled target data,
performed competitively with an overall accuracy of 85.02%; thus establishing a strong fully unsupervised building block for us to build upon.

A. Crisis Tweets vs Amazon Reviews

Table [III] and [IV] show that deep models struggle with small datasets such as TREC-IS tweets. When ST model outperformed Logistic Regression by ~ 8% on the Amazon reviews dataset, the difference was only less than 1% with no statistical significance on the TREC-Priority dataset. Note that we conduct experiments with various parameter combinations on the deep models when using tweets. For example, that we conduct experiments with various parameter combinations on the TREC-Priority dataset resulting in poor generalization. Multi-task learning and domain adversarial training try to alleviate this problem by training the shared BiLSTM layer with much more data from different tasks and unlabeled data.

B. MT-DAAN Performance Evaluation

The primary purpose of the MT-DAAN model is to show that sharing the bottom layer of the model (i.e., shared representation) for different tasks along with domain adversarial training can help improve the generalizability of some of the tasks that are otherwise trained alone in the single-task model. The experiments for MT-DAAN are setup in the same unsupervised way as for single-task. No data from the test crisis is used for training. For example, if we are testing our model for the event 'Typhoon Yolanda', no data from this crisis is used for training. Note that the domain classifier component uses unlabeled data only from rest of the crisis; making it a fully unsupervised domain adaptation approach. Performance scores of the four tasks ('Priority, Factoid, Sentiment, and Irrelevant') are shown in Table[V]. The results show clear performance improvement for 'Priority, Factoid, and Irrelevant' tasks. However, 'Sentiment' task did not show significant improvement. We speculate that this is because other tasks do not generalize the bottom layer enough to boost the sentiment classification performance. These results show the usefulness of multi-task learning as well as domain adversarial training where different tasks in multiple domains help each other when the data is sparse and labels are limited.

C. Word Vectors

We use fastText [33] as our word embeddings for tweets because of its sub-word usage and the ability to create vectors for arbitrary and out-of-vocabulary words. Although there

<table>
<thead>
<tr>
<th>TARGET</th>
<th>Sentiment</th>
<th>ST</th>
<th>Acc</th>
<th>F1</th>
<th>ST-DAAN</th>
<th>Acc</th>
<th>F1</th>
<th>MT-DAAN</th>
<th>Acc</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td></td>
<td></td>
<td></td>
<td>ST-DAAN</td>
<td>Acc</td>
<td></td>
<td>MT-DAAN</td>
<td>Acc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc</td>
<td></td>
<td></td>
<td></td>
<td>Acc</td>
<td></td>
<td></td>
<td>Acc</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>F1</td>
<td></td>
<td></td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COVID-19</td>
<td></td>
<td>73.25</td>
<td>77.36</td>
<td>74.55</td>
<td>77.51</td>
<td>77.00</td>
<td>78.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE VI

UNSUPervised DOMAIN ADAPTATION RESULTS FOR COVID-19 TWEETS USING ONLY THE TREC DATASET FOR TRAINING. EACH REPORTED SCORE IS AN AVERAGE OF 10 INDEPENDENT RUNS OF EACH EXPERIMENT.
intuitive indicators of +ve predictions. These examples show
example, ‘death toll rises’ in the Factoid example, ‘worried, prayers’ in the Sentiment example, and ‘thoughts with people’ in the Irrelevant example are clear intuitive indicators of +ve predictions. These examples show

D. Interpretability: Attention Visualization

The attention weights used to create the context vector by the dot product operation with word activations represent the interpretable layer in our architecture. These weights represent the importance of each word in the classification process. Some examples are shown in Figures 3 and 4. Stronger the color intensity stronger the word attention. In the first example, ‘boston police urging’ is the reason why the tweet is classified as +ve priority. Similarly, ‘death toll rises’ in the Factoid example, ‘worried, prayers’ in the Sentiment example, and ‘thoughts with people’ in the Irrelevant example are clear intuitive indicators of +ve predictions. These examples show

\[ \text{Precision} = \frac{\text{Number of Relevant Tweets Predicted Correctly}}{\text{Total Number of Tweets Predicted as Relevant}} \]

\[ \text{Recall} = \frac{\text{Number of Relevant Tweets Predicted Correctly}}{\text{Total Number of Relevant Tweets}} \]

\[ \text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ \text{AUC-ROC} = \frac{\text{Area Under the Curve of the ROC}}{\text{Maximum Area Under the Curve}} \]

\[ \text{Accuracy} = \frac{\text{Number of Correctly Predicted Tweets}}{\text{Total Number of Tweets}} \]

The attention weights used to create the context vector by the dot product operation with word activations represent the interpretable layer in our architecture. These weights represent the importance of each word in the classification process. Some examples are shown in Figures 3 and 4. Stronger the color intensity stronger the word attention. In the first example, ‘boston police urging’ is the reason why the tweet is classified as +ve priority. Similarly, ‘death toll rises’ in the Factoid example, ‘worried, prayers’ in the Sentiment example, and ‘thoughts with people’ in the Irrelevant example are clear intuitive indicators of +ve predictions. These examples show

\[ \text{Precision} = \frac{\text{Number of Relevant Tweets Predicted Correctly}}{\text{Total Number of Tweets Predicted as Relevant}} \]

\[ \text{Recall} = \frac{\text{Number of Relevant Tweets Predicted Correctly}}{\text{Total Number of Relevant Tweets}} \]

\[ \text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ \text{AUC-ROC} = \frac{\text{Area Under the Curve of the ROC}}{\text{Maximum Area Under the Curve}} \]

\[ \text{Accuracy} = \frac{\text{Number of Correctly Predicted Tweets}}{\text{Total Number of Tweets}} \]

the importance of having interpretability as a key criterion in crisis domain adaptation tasks for social media.

To the best of our knowledge, in social media mining for crisis analytics, there does not exist a ground truth dataset that highlights the words that explain the labels for tweets. Using our model as a guide, we hope to build a robust evaluation dataset as our immediate next step so that the models can be quantitatively evaluated using robust trust-evaluation methods such as LIME [35]. It is also crucial to note that binary classification tasks such as sentiment analysis of Amazon reviews has a clear class divide that produces intuitive keywords such as ‘good’, ‘excellent’, or ‘great’ for +ve reviews and ‘bad’, ‘poor’, or ‘horrible’ for -ve reviews. However, for short texts such as tweets shown in Figure 4 ‘relevancy’ can depend on the context and it is unclear which keywords truly represent the examples in the ‘irrelevant’ class.

VI. COVID-19 USE-CASE

We show a practical implication of our work by applying it to COVID-19 tweets described in Section 4.3. Our goal is to interpretably predict if a COVID-19 tweet is relevant or not; a binary classification task. The models are trained using only the TREC dataset and evaluated on the COVID-19 tweets (a balanced subset of size 637 for +ve and -ve labels). We found that a combination of ‘Priority’ and ‘Irrelevant’ labels from TREC performs better to predict COVID-19’s ‘Relevant’ label (this can be trivially verified by constructing two binary classifiers). We augment all three methods (ST, ST-DAAN, and MT-DAAN) with an additional condition before label prediction: \( R_c = P_t \cap T_c \), which means that a COVID-19 tweet is ‘Relevant’ only if it is predicted both ‘Priority’ = 1 and ‘Irrelevant’ = 0. The scores are reported in Table VI and the attention results are shown in Figure 4 demonstrating the effectiveness of our proposed method.
We presented a novel approach of unsupervised domain adaptation with multi-task learning to classify relevant information from Twitter streams for crisis management, while addressing the problems of data sparsity and limited labels. We showed that a multi-task learning model that shares the lower layers of the neural network with dedicated attention layers for each task along with a domain classifier branch can help improve generalizability and performance of deep models in the settings of limited data. Furthermore, we showed that using an attention-based architecture can help in interpreting the classifier’s predictions by highlighting the important words that justify the predictions. We also presented an in-depth empirical analysis of the state-of-the-art models on both benchmark dataset of Amazon reviews and TREC dataset of crisis events. The application of our generic approach for interpretable and unsupervised domain adaptation within a multi-task learning framework can benefit social media mining systems in diverse domains beyond crisis management.

VIII. ACKNOWLEDGEMENT

We thank the U.S. National Science Foundation grant IIS-1815459 for partially supporting this research, and labmates Yasas Senarath and Rahul Pandey for COVID-19 data annotation along with the Montgomery County CERT volunteers led by Steve Peterson. We also acknowledge ARGO team as the experiments were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University.

REFERENCES