Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data

Amila Silva, Ling Luo, Shanika Karunasekera, Christopher Leckie

School of Computing and Information Systems
The University of Melbourne
Parkville, Victoria, Australia
{amila.silva@student., ling.luo@, karus@, caleckie@}unimelb.edu.au

Abstract

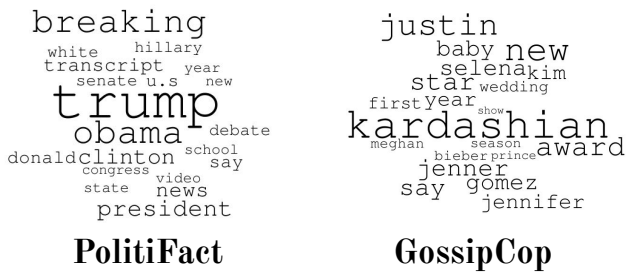
With the rapid evolution of social media, fake news has become a significant social problem, which cannot be addressed in a timely manner using manual investigation. This has motivated numerous studies on automating fake news detection. Most studies explore supervised training models with different modalities (e.g., text, images, and propagation networks) of news records to identify fake news. However, the performance of such techniques generally drops if news records are coming from different domains (e.g., politics, entertainment), especially for domains that are unseen or rarely-seen during training. As motivation, we empirically show that news records from different domains have significantly different word usage and propagation patterns. Furthermore, due to the sheer volume of unlabelled news records, it is challenging to select news records for manual labelling so that the domain-coverage of the labelled dataset is maximized. Hence, this work: (1) proposes a novel framework that jointly preserves domain-specific and cross-domain knowledge in news records to detect fake news from different domains; and (2) introduces an unsupervised technique to select a set of unlabelled informative news records for manual labelling, which can be ultimately used to train a fake news detection model that performs well for many domains while minimizing the labelling cost. Our experiments show that the integration of the proposed fake news model and the selective annotation approach achieves state-of-the-art performance for cross-domain news datasets, while yielding notable improvements for rarely-appearing domains in news datasets.

Introduction

Motivation. Today, social media is considered as one of the leading and fastest media to seek news information online.Thus, social media platforms provide an ideal environment to spread fake news (i.e., disinformation). Many times the cost and damage due to fake news are high and early detection to stop spreading such information is of importance. For example, it has been estimated that at least 800 people died and 5800 were admitted to hospital as a result of false information related to the COVID-19 pandemic, e.g., believing alcohol-based cleaning products are a cure for the virus[1]. Due to the high volumes of news generated on a daily basis, it is not practical to identify fake news using manual fact checking. Therefore, automatic detection of fake news has recently become a significant problem attracting immense research effort.

Challenges. Nevertheless, most existing fake news detection techniques fail to identify fake news in a real-world news stream for the following reasons. First, most existing techniques (Silva et al. 2020; Zhou et al. 2020; Shu et al. 2019, 2020b; Ruchansky et al. 2017) are trained and evaluated using datasets (Shu et al. 2020a; Cui et al. 2020) that are limited to a single domain such as politics, entertainment, healthcare. However, a real news stream typically covers a wide variety of domains. We have empirically found that existing fake news detection techniques perform poorly for such a cross-domain news dataset despite yielding good results for domain-specific news datasets. This observation may be due to two reasons: (1) domain-specific word usage; and (2) domain-specific propagation patterns. For example, Figure 1 adopts two datasets from different domains, PolitiFact for politics and GossipCop for entertainment, which are two widely used labelled datasets to train fake news detection models. Fig. 1 shows that there are significant differences in the frequently used words and propagation patterns of these two datasets. To address this challenge, some previous works (Wang et al. 2018; Castelo et al. 2019) learned models to overlook such domain-specific information and only rely on cross-domain information (e.g., webmarkup and readability features) for fake news detection. However, domain-specific knowledge could be useful for accurate identification of fake news. As a solution, this work aims to address ***how to preserve domain-specific and cross-domain knowledge in news records to detect fake news in cross-domain news datasets.*** Second, the studies in (Han et al. 2020; Janicka et al. 2019) show that most fake news detection techniques are not good at identifying fake news records from unseen or rarely-seen domains during training. As a solution, fake news detection models can be learned using a dataset that covers as many domains as possible. Here we assume that the fake news detection model requires supervision as supervised techniques are known to be substantially better at identifying fake news compared to the unsupervised methods (Yang et al. 2019a). In such a supervised learning setting, each training (i.e., labelled) data point has an associated labelling cost. Thus, the total labelling budget

[1]https://www.bbc.com/news/world-53755067



(a)

Feature	Weiner Index	Network Depth	Maximum Outdegree	Propagation Speed
p-value	1.81e-2	5.81e-19	4.11e-4	3.42e-29

(b)

Figure 1: (a) Word clouds for the top 20 words in PolitiFact and GossipCop. (b) Two-sample t-test results conducted using different graph-level features extracted from the propagation networks in PolitiFact and GossipCop.

constrains the number of data instances that can be selected for manual labelling. Due to the sheer volume of unlabelled news records available, there is a need to ***identify informative news records to annotate such that the labelled dataset ultimately covers many domains while avoiding any selection biases.***

Contribution. To address the aforementioned challenges, this work makes the following contributions:

• We propose a multimodal[2] fake news detection technique for cross-domain news datasets that learns domain-specific and cross-domain information of news records using two independent embedding spaces, which are subsequently used to identify fake news records. Our experiments show that the proposed framework outperforms state-of-the-art fake news detection models by as much as 7.55% in F1-score.

• We propose an unsupervised technique to select a given number of news records from a large data pool such that the selected dataset maximizes the domain coverage. By using such a dataset to train a fake news detection model, we show that the model achieves around 25% F1-score improvements for rarely-appearing domains in news datasets.

Related Work

Fake news detection methods mainly rely on different attributes (text, image, social context) of news records to determine their veracity. Text content-based approaches (Yang et al. 2016; Volkova et al. 2017; Pérez-Rosas et al. 2018; Pennebaker et al. 2015) mainly explore word usage and linguistic styles in the headline and body of news records to identify fake news. Some works analyse the images in news records along with the text content for fake news detection. For example, the studies in (Jin et al. 2017; Wang et al. 2018; Khattar et al. 2019) use pre-trained image models (e.g., VGG-19, ResNet) to extract features from images, which are integrated with text features to identify fake news. Also, some works consider the social context of a news record, i.e., how the record is propagated across social media, as another modality to differentiate fake news records from real ones. Existing work in this line mostly applies various machine learning techniques to extract features from propagation patterns, including Propagation Tree Kernels (Ma et al. 2017), Recurrent Neural Networks (Wu et al. 2018; Liu et al. 2018), and Graph Neural Networks (Monti et al. 2019). However, all these modalities (i.e., text, propagation patterns) generally show notable differences (see Figure 1) for news records in different domains. Thus, most existing techniques perform poorly for cross-domain news datasets due to their inability to capture such domain-specific variations. Our model also relies on the text content and social context of news. However, the main objective of our model is to capture such domain-specific variations of news records.

Domain-agnostic Fake News Detection. Several previous works have attempted to perform fake news detection using cross-domain datasets. In (Wang et al. 2018), an event discriminator is learned along with a multimodal fake news detector to overlook domain-specific information in news records. The study in (Castelo et al. 2019) carefully selects a set of features (e.g., psychological features, readability features) from news records that are domain-invariant. These techniques rely only on cross-domain information in news records. In contrast, Han et al. (2020) consider cross-domain fake news detection as a continual learning task, which learns a model for a large number of tasks sequentially. This work adopts Graph Neural Networks to detect fake news using their propagation patterns and applies well-known continual learning approaches Elastic Weight Consolidation (Kirkpatrick et al. 2017) and Gradient Episodic Memory (Lopez-Paz et al. 2017) to address cross-domain fake news detection problem. This approach has two limitations: (1) it assumes that the news records from different domains arrive sequentially, though this is not always true for real-world streams; and (2) it requires the domain of news records to be known, which is not generally available. In contrast, our approach exploits both domain-specific and cross-domain knowledge of news records without knowing the actual domain of news records.

Active Learning for Fake News Detection. Almost all the aforementioned models are supervised. Although there are unsupervised fake news detection techniques (Yang et al. 2019b; Hosseinimotlagh et al. 2018), they are generally inferior to the supervised approaches in terms of accuracy. However, the training of supervised models requires large labelled datasets, which are costly to collect. Therefore, how to obtain fresh and high-quality labelled samples for a given labelling budget is challenging. Some works (Wang et al. 2020; Bhattacharjee et al. 2017) adopt conventional active learning frameworks to select high-quality samples, in which the model is initially trained using a small randomly selected dataset. Then, the beliefs derived from the initial model are used to select subsequent instances to annotate. This approach has two limitations: (1) it requires a

[2]We define multimodality as information acquired from different sources/attributes following (Zhang et al. 2017), instead of restricting just for sensory media (e.g., text, image).

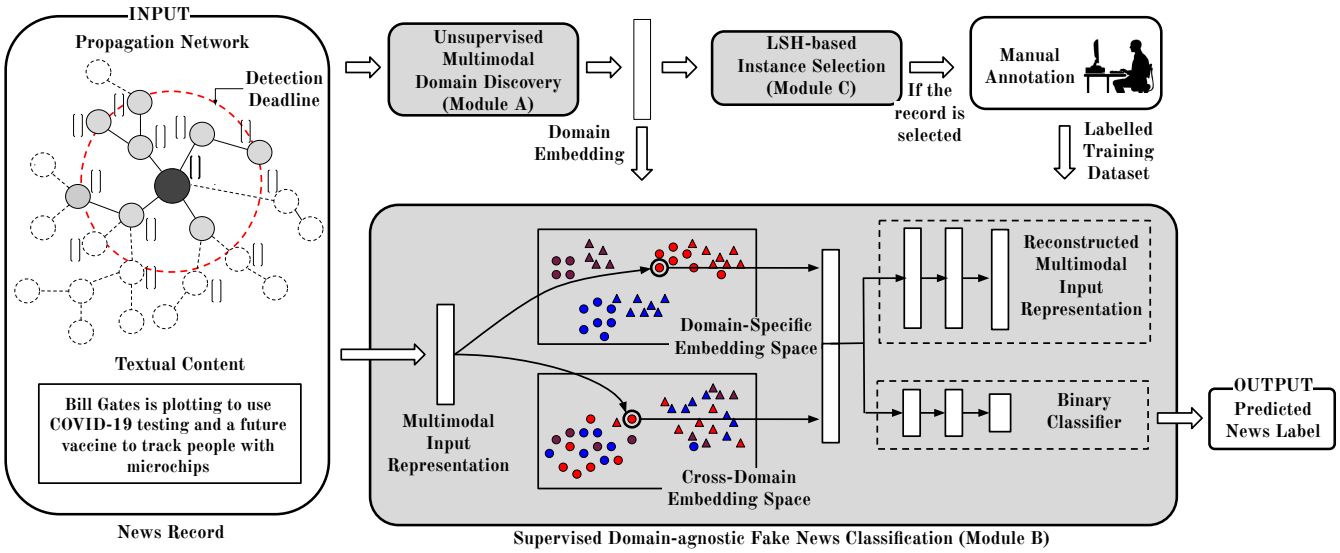


Figure 2: Overview of the proposed framework. In the illustrated embedding spaces, each data point's colour and shape denote its domain label and veracity label (i.e., triangle for fake news and circle otherwise) respectively.

Table 1: Descriptive statistics of PolitiFact, GossipCop and CoAID datasets.

Dataset	PolitiFact	GossipCop	CoAID
# Fake News	269	1269	135
# Real News	230	2466	1568

pre-trained model to select instances; and (2) it is known to be highly vulnerable to the biases introduced by the initial model. In contrast, our instance selection approach does not depend on such an initial model. Also, none of the previous works attempted to explicitly maximize the domain-coverage of the labelled dataset, which is vital to train a model that perform equally well for multiple domains.

Problem Statement

Let R be a set of news records. Each record $r \in R$ is represented as a tuple $\langle t^r, W^r, G^r \rangle$, where (1) t^r is the timestamp when r is published online; (2) W^r is the text content of r; and (3) G^r is the propagation network of r for time bound ΔT. We keep ΔT low (= five hours) for our experiments to evaluate early detection performance. Each propagation network G^r is an attributed directed graph (V^r, E^r, X^r), where nodes V^r represent the tweets/retweets of r and the edges E^r represent the retweet relationships among them. X^r is the set of attributes of the nodes (i.e., tweets) in G^r. More details about E^r and G^r are given in (Silva et al. 2021).

Our problem consists of two sub-tasks: (1) select a set of instances R^L from R to label while adhering to the given labelling budget B, which constrains the number of instances in R^L. The labelling process assigns a binary label y^r for each record r: y^r is 1 if r is false and 0 otherwise; (2) learn an effective model using R^L to predict the label y^r for unlabelled news records $r \in R^U$ as false or real news records. In this work, R $(R^L \cup R^U)$ is not constrained to a specific domain. To emulate such a domain-agnostic dataset, we combine three publicly available datasets: (1) PolitiFact (Shu et al. 2020a), which consists of news related to politics; (2) GossipCop (Shu et al. 2020a), a set of news related to entertainment stories; and (3) CoAID (Cui et al. 2020), a news collection related to COVID-19. All three datasets provide labelled news records and all the tweets related to each news item. The statistics of the datasets are shown in Table 1.

Our Approach

As shown in Fig. 2, the proposed fake news detection model consists of two main components: (1) unsupervised domain embedding learning (Module A); and (2) supervised domain-agnostic news classification (Module B). These two components are integrated to identify fake news while exploiting domain-specific and cross-domain knowledge in the news records. In addition, the proposed instance selection approach (Module C) adopts the same domain embedding learning component to select informative news records for labelling, which eventually yields a labelled dataset that maximizes the domain-coverage.

Unsupervised Domain Discovery

For a given news record r, assume that its domain label is not available. The proposed unsupervised domain embedding learning technique exploits multimodal content (e.g., text, propagation network) of r to represent the domain of r as a low-dimensional vector $f_{domain}(r)$. Our approach is motivated by: (1) the tendency of users to form groups containing people with similar interests (i.e., homophily) (McPherson et al. 2001), which results in different domains having distinct user bases; and (2) the significant differences in domain-specific word usage as shown in Figure 1a.

We exploit the aforementioned motivations by constructing a heterogeneous network which consists of both users tweeting the news items and words in the news title as nodes, using the following steps (Line 1-9 in Algo. 1): (1) create a

Algorithm 1: Domain Embedding Learning

Input: A collection of news records R
Output: Domain embeddings $f_{domain}(r)$ of $r \in R$

```
// Network construction
```
Initialize an empty graph G;
for $r \in R$ **do**
  $S^r \leftarrow X^r \cup U^r$
  for *each pair* $(s_1, s_2) \in S$ **do**
    $e \leftarrow (\{s_1, s_2\}, 1)$;
    if *edge* e *exists in graph* G **then**
      Increment edge e in graph G by 1;
    else
      Add edge e to graph G;
```
// Community Detection
```
$C \leftarrow$ Find communities in G using Louvain;
```
// Embedding Learning
```
for $r \in R$ **do**
  Compute $f_{domain}(r)$ using Eq. 2
Return $f_{domain}(r)$ of $r \in R$.

set S^r for each news record r by adding all the users U^r in the propagation network G^r and all the words appearing in the news title W^r (tokenized using whitespaces); (2) for each pair of items in S^r, build a weighted edge e linking the two items in the graph; and (3) repeat Steps 1 and 2 for all the news records, until we obtain the final network G. Then, we adopt the Louvain algorithm[3] (Blondel et al. 2008) to identify communities in G. Here, we select the Louvain algorithm as it was shown to be one of the best performing parameter-free community detection algorithms in (Fortunato 2010). At the end of this step, we obtain a set of communities/clusters C, each having either a highly connected set of users or words. As the nodes of G contain both users and words, such communities may have formed either due to a set of users engaging with similar news records or a set of words only appearing within a fraction of news records. Following the aforementioned motivations, this work assumes each community in C belongs to a single domain.

In the next step, we compute the soft membership $p(r \in c)$ of r in a cluster c using the following equation:

$$p(r \in c) = \sum_{v \in c \cap r} v_{deg} / \sum_{c \in C} \sum_{v \in r} v_{deg} \quad (1)$$

Here $p(r \in c)$ is proportional to the number of common users or words that r and c have. Each node (i.e., user or word) v is weighted using the degree v_{deg} in G (i.e., number of occurrences) to reflect their varying importance for the corresponding community. Finally, we produce the domain embedding $f_{domain}(r) \in \mathbb{R}^{|C|}$ of r as the concatenation of r's likelihood belonging to communities in C:

$$f_{domain}(r) = p(r \in c_1) \oplus p(r \in c_2) \oplus \dots p(r \in c_{|C|}) \quad (2)$$

where $\oplus$ denotes concatenation.

[3]Please see Supplementary Material in (Silva et al. 2021) for detailed pseudo code of the Louvain algorithm

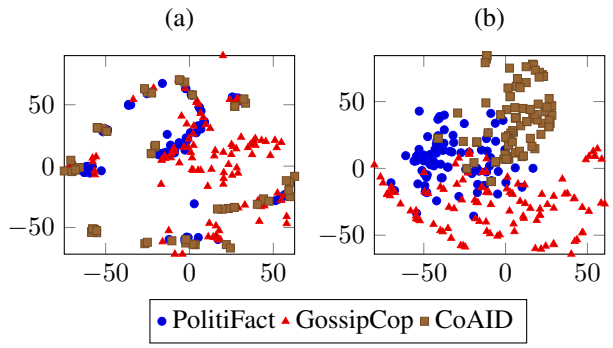


Figure 3: t-SNE visualization of domain embeddings from: (a) user-based domain discovery algorithm in (Chen et al. 2020) and (b) multimodal domain discovery approach proposed in this work.

In Figure 3, we adopt t-SNE (Maaten et al. 2008) to visualize the domain embedding space of the proposed approach and the user-based domain discovery algorithm proposed in (Chen et al. 2020). Due to space limitations, we present more details about the baseline in (Silva et al. 2021). As can be seen in Figure 3, the proposed approach yields a clear separation between the domains compared to the baseline. This may be mainly due to the ability of our approach to jointly exploit multimodalities, both users and text of news records to discover their domains. In addition, most previous works on domain discovery ultimately assign hard domain labels for news records, which could lead to substantial information loss. For example, some news records may belong to multiple domains, which cannot be captured using hard domain labels. Hence, by having a low-dimensional vector to represent embedding, our approach could preserve such knowledge related to the domains of news records.

Domain-agnostic News Classification

In our news classification model, each news record r is represented as a vector $f_{input}(r)$ using the textual content W^r and the propagation network G^r of r (elaborated in the section Experiments). Then, our classification model maps $f_{input}(r)$ into two different subspaces such that one preserves the domain-specific knowledge, $f_{specific} : f_{input}(r) \rightarrow \mathbb{R}^d$, and the other preserves the cross-domain knowledge $f_{shared} : f_{input}(r) \rightarrow \mathbb{R}^d$, of r. Here d is the dimension of the subspaces. Then, the concatenation $f_{specific}(r)$ and $f_{shared}(r)$ is used to recover the label y^r and the input representation $f_{input}(r)$ of r during training via two decoder functions g_{pred} and g_{recons} respectively.

$$\overline{y^r} = g_{pred}(f_{specific}(r) \oplus f_{shared}(r))$$
$$\overline{f_{input}(r)} = g_{recon}(f_{specific}(r) \oplus f_{shared}(r))$$
$$L_{pred} = BCE(y^r, \overline{y^r}) \quad (3)$$
$$L_{recon} = ||f_{input}(r) - \overline{f_{input}(r)}||^2 \quad (4)$$

where $\overline{y^r}$ and $\overline{f_{input}(r)}$ denote the predicted label and the predicted input representation respectively. BCE stands for the Binary Cross-Entropy loss function. We mini-

mize L_{pred} and L_{recon} to find the optimal parameters of $(f_{specific}, f_{shared}, g_{pred}, g_{recon})$.

However, L_{pred} and L_{recon} do not leverage domain differences in news records. Hence, we now discuss how the mapping functions for subspaces, $f_{specific}$ and f_{shared}, are further learned to preserve the domain-specific and cross-domain knowledge in news records.

Leveraging Domain-specific Knowledge To preserve the domain-specific knowledge, we introduce an auxiliary loss term $L_{specific}$ to learn a new decoder function $g_{specific}$ to recover the domain embedding $f_{domain}(r)$ of r using the domain-specific representation $f_{specific}(r)$. We minimize $L_{specific}$ to find the optimal parameters for $(f_{specific}, g_{specific})$ to capture the domain-specific knowledge by $f_{specific}$, and this process can be defined as follows:

$$L_{specific} = ||f_{domain}(r) - g_{specific}(f_{specific}(r))||^2$$
$$(\hat{g}_{specific}, \hat{f}_{specific}) = \underset{(g_{specific}, f_{specific})}{\operatorname{argmin}} (L_{specific}) \quad (5)$$

Leveraging Cross-domain Knowledge In contrast, we learn f_{shared} to overlook domain-specific knowledge of the news records. Consequently, f_{shared} preserves the cross-domain knowledge in the news records. Here, we train a decoder function g_{shared} to accurately predict the domain of r using $f_{shared}(r)$. Meanwhile, we learn f_{shared} to fool the decoder g_{shared} by maximizing the loss of g_{shared}. Such a formulation forces f_{shared} to only rely on cross-domain knowledge, which are useful to transfer the knowledge across domains. This process can be defined as a minimax game between g_{shared} and f_{shared} as follows:

$$L_{shared} = ||g_{shared}(f_{shared}(r)) - f_{domain}(r)||^2$$
$$(\hat{g}_{shared}, \hat{f}_{shared}) = \underset{f_{shared}}{\operatorname{argmin}} \underset{g_{shared}}{\operatorname{argmax}}(-L_{shared}) \quad (6)$$

Integrated Model Then the final loss function of the model is formulated as:

$$L_{final} = L_{pred} + \lambda_1 L_{recon} + \lambda_2 L_{specific} - \lambda_3 L_{shared} \quad (7)$$

where λ_1, λ_2 and λ_3 controls the importance given to each loss term compared to L_{pred} (i.e., main task).

To learn the minimax game in L_{shared}, the final loss function L_{final} is sequentially optimized using the following two steps:

$$(\widehat{\theta_1}) = \underset{\theta_1}{\operatorname{argmin}} L_{final}(\theta_1, \theta_2) \quad (8)$$

$$(\widehat{\theta_2}) = \underset{\theta_2}{\operatorname{argmax}} L_{final}(\widehat{\theta_1}, \theta_2) \quad (9)$$

where θ_1 and θ_2 denote the parameters in ($f_{specific}$, f_{shared}, $g_{specific}$, g_{pred}, g_{recon}) and g_{shared} respectively. The empirically studied convergence properties of the proposed optimization scheme are presented in (Silva et al. 2021).

LSH-based Instance Selection

The aforementioned model is able to exploit the domain-specific and cross-domain knowledge in news records to identify their veracity. Nevertheless, if the model is used to identify fake news records in unseen or rarely appearing domains during training, we empirically observe that the performance of the model substantially drops. This observation is expected and is consistent with the findings in (Castelo et al. 2019), which could be due to the domain-specific word usage and propagation patterns as shown in Fig. 1. Hence, we propose an unsupervised technique to come up with a labelled training dataset for a given labelling budget B such that it covers as many domains as possible. The ultimate objective of this technique is to learn a model using such a dataset that performs well for many domains.

Our approach initially represents each news record $r \in R$ using its domain embedding $f_{domain}(r)$. Then, we propose a Locality-Sensitive Hashing (LSH) algorithm based on random projection to select a set of records in R that are distant in the domain embedding space, which can be elaborated using the following steps:

1. Create $|H|$ different hash functions such as $H_i(r) = sgn(h_i \cdot f_{domain}(r))$, where $i \in \{0, 1, \ldots, |H|-1\}$ and h_i is a random vector, and $sgn(.)$ is the sign function. The random vectors h_i are generated using the following probability distribution, as such a distribution was shown to perform well for random projection-based techniques (Achlioptas 2001):

$$h_{i,j} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/3 \\ -1 & \text{with probability } 1/6 \end{cases} \quad (10)$$

2. Construct an $|H|$-dimensional hash value for each news record r as $H_0(r) \oplus H_1(r) \oplus \ldots \oplus H_{|H|-1}(r)$, where $\oplus$ defines the concatenation operation. According to the Johnson-Lindenstrauss lemma (Johnson et al. 1984), such hash values approximately preserve the distances between the news records in the original embedding space with high probability. Hence, neighbouring records in the domain embedding space are mapped to similar hash values.
3. Group the news records with similar hash values to construct a hash table.
4. Randomly pick a record from each bin in the hash table and add to the selected dataset pool.
5. Repeat steps (1), (2), (3) and (4) until the size of the dataset pool reaches the labelling budget B.

In Figure 4a, we compare 10% of the original dataset selected using the proposed approach and random selection. As can be seen, random selection follows the empirical distribution of the datasets in Table 1 and picks few instances from rarely appearing domains (e.g., fake/real news in PolitiFact, fake news in CoAID). Thus, the model trained on such a dataset may poorly perform on rarely appearing domains. In contrast, the proposed approach provides a significant number of samples from even rarely occurring domains.

In addition, the proposed approach is efficient ($O(|H||R|)$ complexity) compared to the naive farthest point selection algorithms (e.g., k-Means (Lloyd 1982) with $O(|R|^2)$ complexity, where $|R| >> |H|$). To measure the domain coverage of the instances selected from the proposed instance selection approach, we adopt the metric introduced in (Laib et al. 2017), which can be computed as follows for a given set of records $r_1, r_2, ..., r_n$ that are represented using their

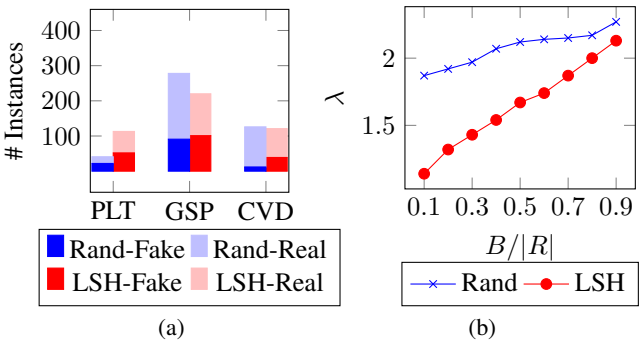


Figure 4: Statistics of datasets selected using random selection (Rand) and the proposed LSH-based technique (LSH). (a) Number of fake and real news records selected from each domain when $B/|R| = 0.1$ and (b) domain-coverage measure λ (lower λ is better) for different $B/|R|$ values.

domain embeddings: $\lambda = \frac{1}{\bar{\delta}}(\frac{1}{n}\sum_{i=1}^{n}(\delta_i - \bar{\delta})^2)^{\frac{1}{2}}$, where $\delta_i = min_k(L2\ norm(f_{domain}(r_i), f_{domain}(r_k)))$ and $\bar{\delta} = \sum \delta_i/n$. If the coverage is high, λ is small. Hence, the proposed approach yields a better domain-coverage compared to random instance selection as shown in Figure 4b.

Experiments

Experimental Setup

Encoding and Decoding Functions In our model, each record r is initially represented as a low-dimensional vector $f_{input}(r)$ using its text content and propagation network. We adopt RoBERTa-base, a robustly optimized BERT pretraining model (Liu et al. 2019) to learn the text-based representation $f_{text}(r)$ of r. The propagation network-based representation $f_{network}(r)$ of r is represented using the unsupervised network representation learning technique proposed in (Silva et al. 2020). Then, the final input representation $f_{input}(r)$ is constructed as $f_{text}(r) \oplus f_{network}(r)$, where $\oplus$ denotes concatenation. All the other encoding and decoding functions, ($f_{specific}$, f_{shared}, $g_{specific}$, g_{shared}, g_{pred}, g_{recon}), are modelled as 2-layer feed-forward networks with sigmoid activation[4].

Dataset We combine three disinformation datasets: (1) PolitiFact; (2) GossipCop; and (3) CoAID, to produce a cross-domain news dataset[5]. Then, we randomly choose 75% of the dataset as the candidate data pool R_{pool} for training and the remaining 25% for testing. For a given labelling budget B, we select B instances from R_{pool} to train the model. The same process is performed for 3 different training and test splits and the average performance is reported. We evaluate the performance for each domain separately using the testing instances from each domain. For the evaluation, we adopt four metrics: (1) Accuracy (Acc); (2) Precision (Prec); (3) Recall (Rec); and (4) F1 Score (F1).

Baselines In Table 2, we compare our approach with seven widely used fake detection techniques and their variants[4].

Parameter Settings After performing a grid search, we have set the hyper-parameters in our model as[4]: $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 5$, $d = 512$. To satisfy the Johnson–Lindenstrauss lemma, we set $|H| = 10$ ($>> log(|R|)$). For the specific parameters of the baselines, we use the default parameters mentioned in their original papers.

Results

Quantitative Results for Fake News Detection As shown in Table 2, the proposed approach yields substantially better results for all three domains, outperforming the best baseline by as much as 7.55% in F1-score. The best baseline, EANN-Multimodal, also adopts domain-information when determining fake news. This observation shows the importance of having domain-knowledge of news records when identifying fake news in cross-domain datasets. In addition to the architectural differences of the model, EANN-Multimodal is different from our approach for two reasons: (1) EANN-Multimodal only preserves cross-domain knowledge in news records. Thus, it overlooks domain-specific knowledge, which is shown to be useful in our ablation study in Table 2; and (2) EANN-Multimodal adopts a hard label (i.e., exclusive membership) to represent the domain of a news record. Our approach conversely uses a vector to represent the domain of a news record. Thus, our approach can accurately represent the likelihood of each record for different domains. These differences may explain the importance of our approach compared to the best baseline.

Out of the baselines, the multimodal approaches (except HPNF+LIWC) generally achieve better results compared to the uni-modal approaches. Thus, we can conclude that each modality (i.e., propagation network and text) of news records provides unique knowledge for fake news detection. In HPNF+LIWC, each news record is represented using a set of hand-crafted features. In contrast, other multimodal approaches including our approach learn data-driven latent representations for news records, which may be able to capture latent and complex information in news records that are useful to determine fake news. These observations further support two main design decisions in our model: (1) to exploit multimodalities of news records; and (2) to adopt a representation learning-based technique.

Ablation Study Our ablation study in Table 2 shows that without the domain-specific loss (Eq. 5) and the cross-domain loss (Eq. 6), the F1-score of the model substantially drops by around 6% and 3% for the PolitiFact dataset, which is the smallest domain of the training dataset. Hence, it is important to have a domain-specific layer to preserve the domain-specific knowledge and a separate cross-domain layer to transfer common knowledge between domains.

To check whether our model actually learns the aforementioned intuition behind each embedding layer, we visualize each embedding layer using t-SNE in Figure 5. As can be

[4]We present more details about implementations and parameter selections in the Supplementary Material in (Silva et al. 2021)

[5]Here we do not consider the existing datasets on rumour detection (Kochkina et al. 2018; Ma et al. 2017) as they are not consistent with the fake news definition (i.e., disinformation).

Table 2: Results for fake news detection of different methods, which are classified under three categories: (1) text content-based approaches (T); (2) social context-based approaches (S); and (3) multimodal approaches (M).

Method	Type			Politifact				Gossipcop				CoAID			
	T	S	M	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
LIWC (Pennebaker et al. 2015)	✓			0.488	0.680	0.565	0.432	0.662	0.550	0.516	0.472	0.903	0.586	0.531	0.538
text-CNN (Kim 2014)	✓			0.608	0.621	0.623	0.608	0.733	0.698	0.703	0.701	0.903	0.679	0.674	0.677
HAN (Yang et al. 2016)	✓			0.632	0.672	0.651	0.648	0.716	0.703	0.709	0.706	0.919	0.698	0.682	0.688
EANN-Unimodal (Wang et al. 2018)	✓			0.794	0.811	0.790	0.791	0.765	0.732	0.738	0.734	0.925	0.842	0.763	0.792
HPNF (Shu et al. 2020b)		✓		0.697	0.692	0.683	0.687	0.721	0.703	0.689	0.695	0.902	0.652	0.693	0.672
AE (Silva et al. 2020)		✓		0.784	0.783	0.774	0.779	0.834	0.828	0.802	0.812	0.928	0.686	0.673	0.677
HPNF + LIWC (Shu et al. 2020b)			✓	0.704	0.723	0.708	0.716	0.734	0.715	0.706	0.708	0.911	0.682	0.709	0.690
SAFE (Zhou et al. 2020)			✓	0.793	0.782	0.771	0.775	0.831	0.822	0.798	0.806	0.931	0.754	0.744	0.748
EANN-Multimodal (Wang et al. 2018)			✓	0.804	0.808	0.794	0.798	0.836	0.812	0.815	0.813	0.944	0.849	0.803	0.808
Our Approach ($B = 100\%\|R_{pool}\|$)			✓	**0.840**	**0.836**	**0.831**	**0.835**	**0.877**	**0.840**	**0.832**	**0.836**	**0.970**	**0.876**	**0.863**	**0.869**
Our Approach ($B = 50\%\|R_{pool}\|$)			✓	**0.838**	**0.836**	**0.828**	**0.833**	**0.848**	**0.822**	**0.797**	**0.808**	**0.963**	**0.870**	**0.854**	**0.862**
Ablation Study ($B = 100\%\|R_{pool}\|$)															
(-) *Domain-shared loss*				0.823	0.821	0.812	0.815	0.864	0.832	0.828	0.829	0.956	0.857	0.861	0.858
(-) *Domain-specific loss*				0.792	0.800	0.783	0.786	0.858	0.832	0.821	0.828	0.934	0.850	0.857	0.853
(-) *Network modality*				0.816	0.815	0.817	0.815	0.765	0.749	0.745	0.746	0.945	0.803	0.855	0.827
(-) *Text modality*				0.804	0.798	0.793	0.795	0.837	0.835	0.815	0.817	0.932	0.711	0.704	0.707

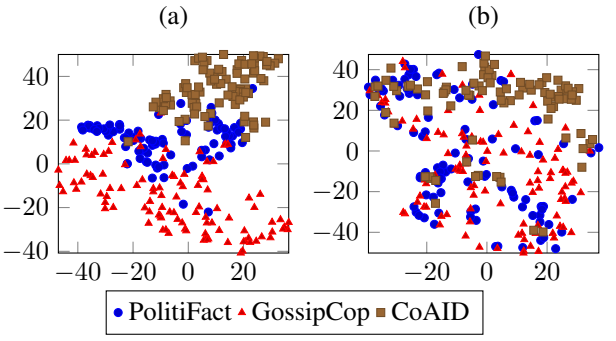


Figure 5: t-SNE visualization of the (a) domain-specific and (b) cross-domain embeddding spaces.

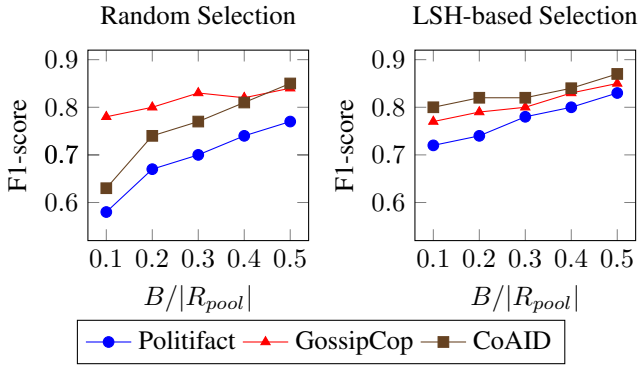


Figure 6: F1-scores for the fake news detection task with different instance selection strategies.

seen, the domain-specific embedding layer preserves the domain of the news records by mapping different domains into different clusters. In contrast, we cannot identify the domain labels of news records from the cross-domain embedding space. Hence, this embedding space is useful to share common knowledge between records from different domains.

Furthermore, we analyse the contribution of each modality. It can be seen that *network modality* is more useful to determine fake news in GossipCop, while *text modality* is the most informative one for CoAID. This observation further signifies the importance of multimodal approaches to train models that generalize for multiple domains.

Evaluation of LSH-based Instance Selection As shown in Table 2, our model outperforms the baselines even with a constrained budget B ($50\%|R_{pool}|$) to select training data using the LSH-based instance selection technique. To verify its significance further, Figure 6 compares the proposed LSH-based instance selection approach with random instance selection for different B values. The proposed approach substantially outperforms the random instance selection for the rarely-appearing or highly imbalanced domains. It increases F1-score by 24% for PolitiFact and 27% for CoAID, when $B/|R_{pool}| = 0.1$. This may be due to the ability of our approach to maximize the coverage of domains when selecting instances (see Figure 4), instead of biasing towards a domain with larger number of records.

Conclusion

In this work, we proposed a novel fake news detection framework, which exploits domain-specific and cross-domain knowledge in news records to determine fake news from different domains. Also, we introduced a novel unsupervised approach to select informative instances for manual labelling from a large pool of unlabelled news records. The selected data pool is subsequently used to train a model that can perform equally for different domains. The integration of the aforementioned two contributions yields a model with low labelling budgets that outperforms existing fake news detection techniques by as much as 7.55% in F1-score.

For future work, we intend to extend our model as an online learning framework to determine fake news in a real-world news stream, which typically covers a large number of domains. This setting introduces new challenges such as capturing newly emerging domains and handling temporal

changes in domains. Also, how to use the alignment in multimodal information to weakly guide the learning process of the proposed model is another interesting direction to explore, which may further reduce the labelling cost in a conventional supervised learning setting.

Acknowledgments

This research was financially supported by Melbourne Graduate Research Scholarship and Rowden White Scholarship. We would like to specially thank Yi Han for his insightful comments and suggestions for this work. We are also grateful for the time and effort of the reviewers in providing valuable feedback on our manuscript.

References

Achlioptas, D. 2001. Database-friendly Random Projections. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 274–281.

Bhattacharjee, S. D.; Talukder, A.; and Balantrapu, B. V. 2017. Active Learning Based News Veracity Detection with Feature Weighting and Deep-shallow Fusion. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, 556–565.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10).

Castelo, S.; Almeida, T.; Elghafari, A.; Santos, A.; Pham, K.; Nakamura, E.; and Freire, J. 2019. A Topic-agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of the World Wide Web Conference*, 975–980.

Chen, Z.; and Freire, J. 2020. Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds. In *Companion Proceedings of the World Wide Web Conference*, 584–592.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv e-prints* arXiv:2006.00885.

Fortunato, S. 2010. Community Detection in Graphs. *Physics Reports* 486(3-5): 75–174.

Han, Y.; Karunasekera, S.; and Leckie, C. 2020. Graph Neural Networks with Continual Learning for Fake News Detection from Social Media. *arXiv e-prints* arXiv:12007.03316.

Hosseinimotlagh, S.; and Papalexakis, E. E. 2018. Unsupervised Content-based Identification of Fake News Articles with Tensor Decomposition Ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Janicka, M.; Pszona, M.; and Wawer, A. 2019. Cross-Domain Failures of Fake News Detection. *Computación y Sistemas* 23(3).

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the ACM International Conference on Multimedia*, 795–816.

Johnson, W. B.; and Lindenstrauss, J. 1984. Extensions of Lipschitz Mappings into a Hilbert Space. *Contemporary Mathematics* 26(189-206): 1.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of The World Wide Web Conference*, 2915–2921.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences* 114(13): 3521.

Kochkina, E.; Liakata, M.; and Zubiaga, A. 2018. All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the International Conference on Computational Linguistics*, 3402–3413.

Laib, M.; and Kanevski, M. 2017. Unsupervised Feature Selection Based on Space Filling Concept. *arXiv preprint arXiv:1706.08894* .

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* arXiv:1907.11692.

Liu, Y.; and Wu, Y.-f. B. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 354–361.

Lloyd, S. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory* 28(2): 129–137.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient Episodic Memory for Continual Learning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, 6467–6476.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 708–717.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27(1): 415–444.

Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; and Bronstein, M. M. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *arXiv e-prints* arXiv:1902.06673.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The Development and Psychometric Properties of

LIWC2015. Technical report. URL https://repositories.lib.utexas.edu/handle/2152/31333.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *Proceedings of the International Conference on Computational Linguistics*, 3391–3401.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the ACM on Conference on Information and Knowledge Management*, 797–806.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. DEFEND: Explainable Fake News Detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020a. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 171–188.

Shu, K.; Mahudeswaran, D.; Wang, S.; and Liu, H. 2020b. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, 626–637.

Silva, A.; Han, Y.; Luo, L.; Karunasekera, S.; and Leckie, C. 2020. Embedding Partial Propagation Network for Fake News Early Detection. *Proceedings of the International workshop on Mining Actionable Insights from Social Networks (MAISoN 2020) co-located with CIKM2020* .

Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Supplementary Materials for Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data URL https://drive.google.com/drive/folders/1JRWxtAwd52Uibw0AHYWwcIAdNaWK813?usp=sharing.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 647–653.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.

Wang, Y.; Yang, W.; Ma, F.; Xu, J.; Zhong, B.; Deng, Q.; and Gao, J. 2020. Weak Supervision for Fake News Detection via Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, 516–523.

Wu, L.; and Liu, H. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 637–645.

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019a. Unsupervised Fake News Detection on Social Media: A Generative Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5644–5651.

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019b. Unsupervised Fake News Detection on Social Media: A Generative Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 33: 5644–5651.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Zhang, C.; Zhang, K.; Yuan, Q.; Tao, F.; Zhang, L.; Hanratty, T.; and Han, J. 2017. React: Online multimodal embedding for recency-aware spatiotemporal activity modeling. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 245–254.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Proceedings of Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 354–367.

Supplementary Material for Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data

Amila Silva, Ling Luo, Shanika Karunasekera, Christopher Leckie

School of Computing and Information Systems
The University of Melbourne
Parkville, Victoria, Australia
{amila.silva@student., ling.luo@, karus@, caleckie@}unimelb.edu.au

Abstract

This is the supplementary material for the paper titled "Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data".

Louvain Algorithm for Community Detection

This section presents more details about the Louvain algorithm (Blondel et al. 2008), which is used in the proposed domain embedding learning approach to identify communities in a network.

As shown in Algorithm 1, the Louvain algorithm identifies the communities in a network using the following steps:

1. Each vertex is placed in their own community (Line 1 in Algo. 1);
2. Each vertex is retained in its own cluster or merge with an immediate neighbour such that the modularity scores of the network is maximised (Line 3-15 in Algo. 1). The modularity score is computed as:

$$Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

   where $\sum_{in}$ and $\sum_{tot}$ represents the total weight of all links inside a community/cluster and total weight of all links to a community/cluster, respectively. Similarly, the terms k_i and $k_{i,in}$ denote the total weight of all links to i and total weight of links to i within the community/cluster. Lastly, m denotes the total weight of all links in the network graph;
3. Build a new network where vertices in the same community are combined as a single vertex.
4. Repeat Steps 2 and 3 until there are no more mergings between communities.

At the end of this algorithm, we will obtain a set of communities of the provided network such that the modularity score of the network is maximised. We selected this algorithm in our model because it is known (Lim, Karunasekera, and Harwood 2017) to generate a relatively small number of communities compared to other parameter-free community detection algorithms such as Infomap (Rosvall and Bergstrom 2008) and Label Propagation (Raghavan, Albert, and Kumara 2007).

Algorithm 1: Louvain Algorithm

Input: $G = (V, E)$ where V and E are the vertices and edges of the network G
Output: $A = (V, C)$: Assignment of vertices V into communities C

```
1  Assign all vertices v into their own community;
2  do
3      for v ∈ V do
4          MaxModularity ← −1;
5          MaxModNeighbour ← NULL;
6          for each neighbour v_n of v do
7              ShiftMod ← Modularity score of
                shifting v to v_n's community;
8              if ShiftMod > MaxModularity then
9                  MaxModularity ← ShiftMod;
10                 MaxModNeighbour ← v_n;
11         OriginalMod ←Modularity score of v in
            its original community;
12         if OriginalMod > MaxModularity then
13             Shift v to the community of
                MaxModNeighbour;
14         else
15             Keep v in its original community;
16 while A stabilises (i.e., no more shifts);
```

Multimodal Input Representation

In our model, each news record r is inputted as a low-dimensional vector $f_{input}(r)$ using its text content (i.e., news title) and propagation network (i.e., social context). Initially, we construct two independent representations for r using its text content $f_{text}(r)$ and propagation network $f_{network}(r)$. Then, these two representations are concatenated to produce the final representation of r: $f_{input}(r) =$

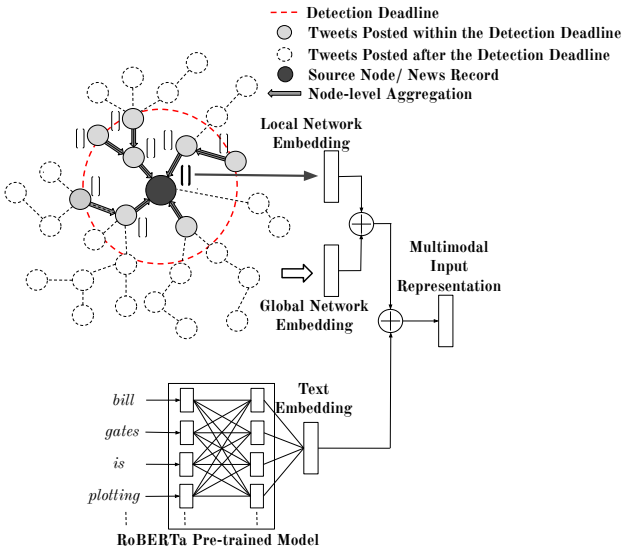


Figure 1: Multimodal Input Representation

$f_{text}(r) \oplus f_{network}(r)$. This process is elaborated in this section.

Text Representation

In this work, the text content of a news record is represented using RoBERTa (Liu et al. 2019), a robustly optimized BERT pre-training model. For a given textual content $\{w_1 w_2 w_3 ....w_n\}$ of a news record r, the RoBERTa model returns the text-based latent representation $f_{text}(r) \in \mathbb{R}^{d_t}$ of r. Out of the different variants of pretrained RoBERTa models, we adopt the roberta-large model available in https://pytorch.org/hub/pytorch_fairseq_roberta/, where $d_t = 1024$.

Propagation Network Representation

We explore two types of features: global-level features (global); and node-level features (local), of the propagation network $G^r = (V^r, E^r, X^r)$ to generate the network-based representation $f_{network}(r)$ of a record r.

Propagation Network Construction We consider all the tweets/retweets related to r as the nodes V^r of G^r. There is an extra node (i.e., source node) in G^r to represent the news, which links different information cascades of r. The edges E^r of G^r represent how a news item spreads from one person to another as shown in Fig 1. Specifically, there is an edge from node i to node j if (1) the user of tweet i mentions the user of tweet j; or (2) tweet i is public and tweet j is posted within the detection deadline (= five hours) after tweet i.

Global Representation We use the following features as global-level features: (1) Wiener Index (g_1); (2) Number of nodes (g_2); (3) Network depth (g_3); (4) Number of nodes at different hops (g_5); and (5) Branching factor at different levels (g_6). Finally, all these features are concatenated together to formulate the global-level network representation $f_{global}(r)$ of a record r.

Table 1: Node-level Features

Type	Features
user	whether the user is verified (n_1), the number of followers (n_2), the number of friends (n_3), the number of lists (n_4), and the number of favourites (n_5)
text	the sentiment scores computed using VADER with the text content in the tweet (n_6), the proportion of positive words (n_7), the proportion of negative words (n_8), the number of mentions (n_9), and the number of hashtags(n_{10})
temporal	the time difference with the source node (n_{11}); the time difference with the immediate predecessor (n_{12}); and the average time difference with the immediate successors (n_{13}); user account timestamp (n_{14})

Algorithm 2: Local Network Representation

Input: propagation network $G^r = (V^r, E^r, X^r)$
source node of r $v_s \in V^r$
Output: The local representation $f_{local}(r)$

$h_v^0 \leftarrow x_v \quad \forall v \in V^r$
for t *in* $1, 2, ..., k$ **do**
  for v *in* V **do**
    $h_v^t \leftarrow \frac{1}{2} h_v^{t-1} + \frac{1}{2} \frac{\sum_{\forall (v,u) \in E_t^r} h_u^{t-1}}{\sum_{\forall (v,u) \in E_t^r} 1}$
$f_{local}(r) \leftarrow h_{v_s}^k$
$return\ f_{local}(r)$

Local Representation For the node-level features, we extract three types of features: (1) text-based; (2) user-based; and (3) temporal-based, which are listed in Table 1. For a given propagation network G^r of a record r, all the features in Table 1 are extracted to represent each vertex (i.e., tweet) in G^r. Then, we adopt the node-level aggregation approach proposed in (Silva et al. 2020) to propagate the aforementioned node-level features to the source node as elaborated in Algo. 2. This algorithm returns the final representation of the source node (see Fig. 1) of G^r as the local representation $f_{local}(r)$ of r.

Finally, the network-based representation is formulated as:

$$f_{network}(r) = f_{global}(r) \oplus f_{local}(r) \quad (1)$$

where $\oplus$ denotes concatenation.

Note: We standardise[1] each dimension of $f_{network}(r)$ before inputting to the model to stabilise the learning process of our model.

Encoding and Decoding Functions

In our fake news detection classifier, we have six encoding and decoding functions, ($f_{specific}$, f_{shared}, $g_{specific}$, g_{shared}, g_{pred}, g_{recon}). In this work, all these functions are

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

modelled as 2-layer feed-forward networks with sigmoid activation. Formally, we can define a encoding/decoding function f that maps an input $x \in \mathbb{R}^{d_{input}}$ to an output $z \in \mathbb{R}^{d_{output}}$ as:

$$z = \sigma(A_2(\sigma(A_1 x + b_1)) + b_2)$$

where $A_1 \in \mathbb{R}^{(d_{hidden}, d_{input})}$, $A_1 \in \mathbb{R}^{(d_{output}, d_{hidden})}$, $b_1 \in \mathbb{R}^{d_{hidden}}$, and $b_2 \in \mathbb{R}^{d_{output}}$ are trainable parameters. σ denotes sigmoid activation. We set d_{hidden} as $\max(d_{input}, d_{output})/2$. For example, assume that f takes inputs of 1024 dimensions and outputs of 128 dimensions. Then, the size of the hidden layer is 512. We leave the optimal neural architecture search for each encoding and decoding function in our model as future work.

Domain Discovery Baseline

We compare our domain discovery approach with the baseline proposed in (Chen and Freire 2020), which assigns hard domain labels for news records based on the users engaged with each news record. For the visualization purpose, we convert these hard domain labels (i.e., one-hot vector) to domain embeddings as they preserve pairwise domain similarity between records (Shu, Wang, and Liu 2019). We elaborate the steps that we followed to generate the domain embeddings using this baseline as follows:

1. Initially, we construct a network by considering each news record as a node.
2. Each news record r (i.e., node) is represented using the list of the users U^r tweeting the the particular news record.
3. The pairwise similarity of nodes is computed for a given two nodes r_1 and r_2 as:

   $$similarity(r_1, r_2) = \frac{|U^{r_1} \cap U^{r_2}|}{|U^{r_1} \cup U^{r_2}|}$$

   Then r_1 and r_2 are connected in the graph if $similarity(r_1, r_2) > \alpha$. α is set to 0.4 following the original paper (Chen and Freire 2020).
4. The Louvain algorithms is used to identify the communities $C = c_1, c_2, ...$ in the constructed graph, which yields hard cluster (considered as domains) assignment for each node.
5. Then each node r can be represented as an one-hot vector $\mathbb{I}^r \in \mathbb{R}^{|C|}$, in which $\mathbb{I}^r_i := \{1\ if\ r \in c_i; 0\ otherwise\}$
6. Finally, we construct the domain embedding $f_{domain}(r) \in \mathbb{R}^{|R|}$ of r by concatenating the cosine similarity scores of $\mathbb{I}^r$ with other news records:

   $$f_{domain}(r) = (\mathbb{I}^r \cdot \mathbb{I}^{r_0}) \oplus (\mathbb{I}^r \cdot \mathbb{I}^{r_1}) ... \oplus (\mathbb{I}^r \cdot \mathbb{I}^{r_{|R|-1}})$$

   where $\oplus$ denotes concatenation operation.

Since this approach considers news records as the nodes of the constructed graph, it is difficult to extend such an approach to learn domain embeddings for new records. In contrast, the proposed approach in this paper constructs its knowledge network using words and users as nodes. Thus, we can generate the domain embeddings for a new record using the words and users related to the new record. Also, our approach considers both text and user information of news records to identify their domain labels.

Fake News Detection Baselines

We compare our fake news detection model with seven widely used baselines and their variants:

- LIWC (Pennebaker et al. 2015) ((i.e., Linguistic Inquiry and Word Count)) learns feature vectors from the text content of news records by counting the number of lexicons falling into different psycho-linguistic categories[2]. Then, a logistic regression model[3] is used as the classifier to predict fake news using LIWC feature vectors.
- text-CNN (Kim 2014) uses Convolution Neural Networks (CNN) to model the text content of news records at different granularity levels with the help of multiple convolutional filters and multiple CNN layers[4].
- HAN (Yang et al. 2016) adopts a hierarchical attention neural network framework to model the text content of news records, which can assign varying importance to words and sentences when making final predictions by word-level and sentence-level attention[5].
- EANN (Wang et al. 2018) produces a latent representation for each news record using its different modalities (e.g., text, network) such that the domain-specific knowledge in news records are ignored in the latent space. Subsequently, the latent representation is used to predict the label of the news record. We compare our model with two variants of EANN:
  - EANN-Unimodal only considers the text modality of a news record to generate the latent representation; and
  - EANN-Multimodal considers both text and network modalities of a news record to produce the latent embedding.

  For a fair comparison of the models, we adopt the same text and network representation techniques in our model to encode the input modalities of EANN.
- HPNF (Shu et al. 2019) extracts various features (e.g., structural features, temporal features) from the propagation network of a news record to generate its feature representation. Then, a Logistic Regression is used to classify news records using the extracted propagation network-based model. In HPNF+LIWC, we concatenate the features vectors from HPNF and LIWC together to construct the feature representation for news records.
- AE (Silva et al. 2020) adopts an Auto-encoder architecture to learn latent representation for each news record based on its propagation network. Subsequently, the latent representations are used to determine fake news records.
- SAFE (Zhou, Wu, and Zafarani 2020) proposes a multimodal approach for fake news detection. For a given news record, this model learns separate latent representations for each modality. Also, it jointly learns another representation to represent cross-modality knowledge, which

[2] https://liwc.wpengine.com/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] https://github.com/yoonkim/CNN_sentence

[5] https://github.com/tqtg/hierarchical-attention-networks

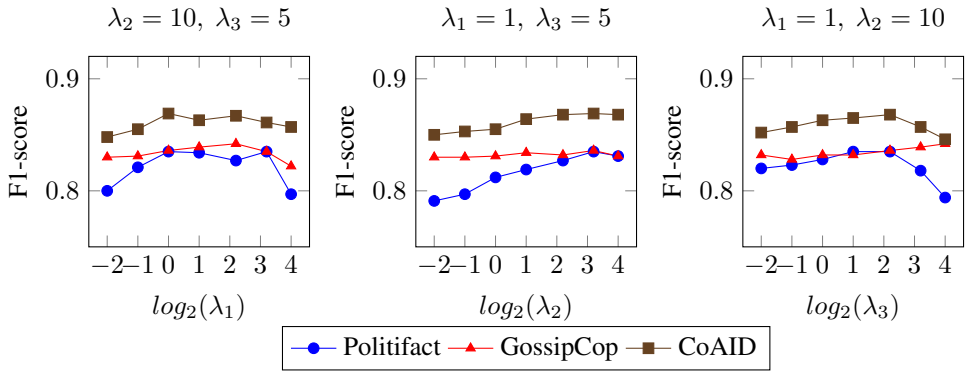


Figure 2: F1-scores for the fake news detection task with different hyper-parameters: λ_1; λ_2; and λ_3.

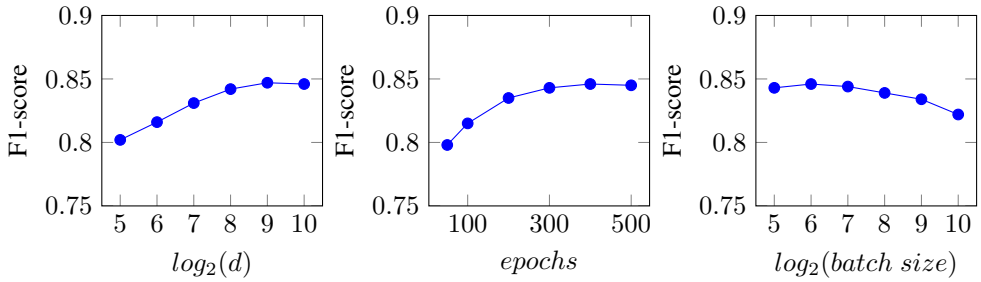


Figure 3: F1-scores (overall for the all three datasets) for the fake news detection task with different hyper-parameters: d; $epochs$; and $batch\ size$.

is consistent across modalities. Finally, all three representations are concatenated and fed to a classifier to predict the label of the record. The original work of this model considers the text and image modalities of news records. For a fair comparison with our model, here we use the text and network modality of news records for this baseline too. We adopt the same text and network representation techniques in our model to encode the input modalities in this baseline too.

Parameter Sensitivity

This section evaluates how changes to the hyper-parameters of the model affect its performance on the fake news detection tasks.

In Figure 2, we analyse the performance of our model for different λ_1, λ_2 and λ_3 values (see Eq. 7 in the paper), which varies the importance assign to each loss term in our model. By setting a very high value ($> 2^2$) or a very low value ($< 2^{-1}$) for λ_1 tends to drop the performance consistently for all three datasets. It means that L_{recon} loss term should be included in our model with moderate importance compared to the other loss terms. The performance of the model for PolitiFact and CoAID domains drop substantially for low $\lambda_2 < 5$ and high $\lambda_3 > 5$ values. By setting a low $\lambda_2 < 5$ or a high $\lambda_3 > 5$ value, our model assigns more importance to the cross-domain embedding space. The cross-domain embedding space could be dominated by frequently appearing domains (GossipCop in this dataset). Thus, assigning more importance for cross-domain embedding space, the model could poorly perform for small domains e.g., PolitiFact and CoAID in this dataset as shown in Fig. 2. This observation further signifies the importance of having domain-specific knowledge of news items to identify fake news.

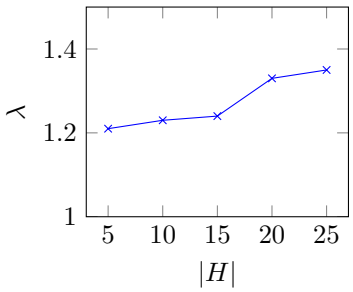


Figure 4: Domain-coverage measure λ (lower λ is better) of the dataset selected using the LSH-based instance selection with different $|H|$ (number of hash functions) values, when $B/|R| = 0.1$.

We examine the sensitivity of the model's performance for other parameters: latent dimension (d); number of epochs; and batch size. Overall, the model yields consistent performance for $d > 256$, $epochs > 300$, and $batch\ size < 128$ values.

There is only one hyper-parameter in the proposed LSH-based instance selection approach, which is the number of hash functions ($|H|$) used for the random projections. As shown in Figure 4, domain coverage of the proposed ap-

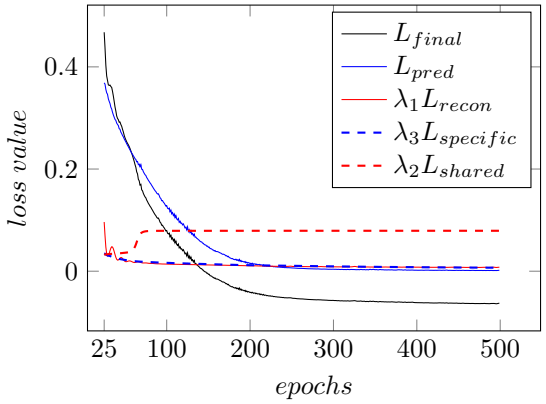


Figure 5: Convergence properties of the loss function.

proach reduces (increases λ measure) for high $|H|(> 20)$ values. This is intuitive because high $|H|$(lengthy hash codes) value could map even very close neighbours in the embedding space into different bins. Thus, the selected instance from different bins could be close-neighbours. In contrast, low $|H|$ values increases the domain coverage. Nevertheless, having a very low $|H|$ value increases the time complexity as it requires many iterations of the hashing step to meet a given labelling budget.

In summary, we adopt the following hyper-parameter values for the results reported in the paper: (1) $lambda_1 = 1$; (2) $lambda_2 = 10$; (3) $lambda_3 = 5$; (4) $|H| = 10$; (5) $d = 512$; (6) $epochs = 300$; (7) $batch\ size = 64$. We use the Adam optimizer for the optimization. For the parameters of the optimizer (e.g., learning rate, moments), the default parameters in Keras[6] are used. Due to the randomness involved in the training and testing datasets splitting process, we conducted all our experiments using three random state value: $\{0, 1, 2\}$, and the average performance is reported in the paper.

Convergence Analysis

In Figure 5, we examine the convergence properties of the loss function of our model. Our loss function consists of four terms: prediction loss (L_{pred}); reconstruction loss (L_{recon}); domain-specific loss ($L_{specific}$); and cross-domain loss (L_{shared}). As can be seen in Fig. 5, each loss term converges around 250 epochs. Since L_{shared} is trained as a minimax game, the converging L_{shared} in Fig. 5 empirically verifies the convergence of the proposed minimax game to exploit cross-domain knowledge in news records. Moreover, L_{recon}, $L_{specific}$ and L_{shared} are mean-squared error based loss terms and L_{pred} is based on binary cross-entropy. Hence, the typical value range for the non-converged L_{pred} differs from the other loss terms. This also shows the importance of having λ_1, λ_2, and λ_3 to penalise such differences due to different loss functions.

[6]https://keras.io/api/optimizers/adam/

References

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10).

Chen, Z.; and Freire, J. 2020. Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds. In *Companion Proceedings of the World Wide Web Conference*, 584–592.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Lim, K. H.; Karunasekera, S.; and Harwood, A. 2017. Clustop: A Clustering-based Topic Modelling Algorithm for Twitter using Word Networks. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, 2009–2018. IEEE.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* arXiv:1907.11692.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The Development and Psychometric Properties of LIWC2015. Technical report. URL https://repositories.lib.utexas.edu/handle/2152/31333.

Raghavan, U. N.; Albert, R.; and Kumara, S. 2007. Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks. *Physical review E* 76(3): 036106.

Rosvall, M.; and Bergstrom, C. T. 2008. Maps of Random Walks on Complex Networks Reveal Community Structure. *Proceedings of the National Academy of Sciences* 105(4): 1118–1123.

Shu, K.; Mahudeswaran, D.; Wang, S.; and Liu, H. 2019. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. *arXiv e-prints* arXiv:1903.09196.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 312–320.

Silva, A.; Han, Y.; Luo, L.; Karunasekera, S.; and Leckie, C. 2020. Embedding Partial Propagation Network for Fake News Early Detection. *Proceedings of the International workshop on Mining Actionable Insights from Social Networks (MAISoN 2020) co-located with CIKM2020* .

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1480–1489.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Proceedings of Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 354–367.