Towards Self-Supervised Cross-Domain Fake News Detection

Carmela Comito[1], Francesco Sergio Pisani[1], Erica Coppolillo[2], Angelica Liguori[2,*], Massimo Guarascio[1] and Giuseppe Manco[1]

[1]Institute for High Performance Computing and Networking, Via P. Bucci 8-9/C, Rende, 87036, Italy

[1]University of Calabria, Via P. Bucci, Rende, 87036, Italy

Abstract

Twitter, Facebook, and Instagram are just some examples of social media currently used by people to share news with other users worldwide. However, the information widespread through these channels is typically unverified and/or interpreted according to the user's point of view. Accordingly, those means represent the perfect tool to hack user opinions with misleading or false news and make fake news viral. Identifying this malicious information is a crucial but challenging task since fake news can concern different topics. Indeed, the detection models learned against a specific domain will exhibit poor performances when tested on a different one. In this work, we propose a novel deep learning-based architecture able to mitigate this problem by yielding cross-domain high-level features for addressing this task. Preliminary experimentation conducted on two benchmarks demonstrated the validity of the proposed solution.

Keywords

Misinformation, Cross Domain Fake News Detection, Deep Learning

1. Introduction

Online Web sources and Social media represent the main means for news information dissemination and spreading. In particular, an exponential increase in the use of social media has accelerated information diffusion. The speed with which misinformation spreads, alongside social media's open access content production and dissemination, increases the potential damage, making online platforms major targets for fake news propagation.

Misinformation, in general, often spreads faster and more widely than true news, posing new risks for democracy and national security, weakening trust in public institutions, and putting at risk society's trust in information. As an example, disinformation is polarizing public debate on topics related to COVID-19; amplifying hate speech; heightening the risk of conflict, violence and human rights violations; and threatening long-term prospects for advancing democracy,

ITASEC 2023: The Italian Conference on CyberSecurity, May 03–05, 2023, Bari, Italy

[*] Corresponding author.

carmela.comito@icar.cnr.it (C. Comito); francescosergio.pisani@icar.cnr.it (F. S. Pisani); erica.coppolillo@unical.it (E. Coppolillo); angelica.liguori@dimes.unical.it (A. Liguori); massimo.guarascio@icar.cnr.it (M. Guarascio); giuseppe.manco@icar.cnr.it (G. Manco)

0000-0001-9116-4323 (C. Comito); 0000-0003-2922-0835 (F. S. Pisani); 0000-0002-4670-8157 (E. Coppolillo); 0000-0001-9402-7375 (A. Liguori); 0000-0001-7711-9833 (M. Guarascio); 0000-0001-9672-3833 (G. Manco)

CEUR Workshop Proceedings (CEUR-WS.org)

human rights, and social cohesion. It has been estimated that at least 800 people died and 5800 were admitted to hospital due to false information related to the COVID-19 pandemic, e.g., believing alcohol-based cleaning products are a cure for the virus [1]. As another example, a report estimated that over 1 million tweets were related to the fake news story "Pizzagate" by the end of the 2016 US presidential election[1].

While, in the last years, COVID-19 emergency provided a dramatic and pressing example of the paramount importance of increased effectiveness of fake news detection in the health field, recently other crucial issues that require proper communication have started attracting public attention, like the Russia-Ukraine war. Disinformation can be harmful in all these contexts and may lead public opinion to push forward detrimental decisions with huge social and economic costs.

It is, therefore, necessary to limit the impacts of misinformation, as well as develop specific tools and services to allow citizens and the professional community to access reliable and trustworthy information on the Web and Social media.

In this scenario, assessing the veracity and authenticity of news represents a crucial problem that can benefit from recent advances in Artificial Intelligence (AI) and Machine Learning (ML). As this task is time-consuming, expensive, and unfeasible on huge amounts of data produced on the Web, AI-Based tools represent an effective solution to automate the identification of deceptive information by limiting the need for intervention by specialized and trusted professionals.

In particular, the automatic detection of fake news is a relevant problem attracting great interest from the research community. This problem was traditionally addressed in the literature as a text classification problem [2] i.e., distinguishing between real and fake news documents.

However, learning reliable detection models able to identify misinformation requires coping with different complex issues. First, an effective solution should allow for handling low-level raw data frequently affected by noise, as the channels used to spread fake news typically allow for sharing only short text (e.g., Twitter). Moreover, the number of labelled training instances is limited; the labelling phase is a difficult and time-consuming task manually performed by domain experts. Finally, fake news can concern different topics; therefore, the features leveraged to perform the prediction should be domain independent to handle different topics.

Notably, most existing techniques [3, 4, 5, 6, 7, 8, 9] are limited to a single domain and perform poorly in cross-domain scenarios also because fake news usually emerge on novel events for which no labeled data is available. Despite the success of deep learning models with large amounts of labeled datasets, the algorithms still suffer in cases where fake news detection is needed on emergent events.

Contribution. To address the aforementioned issues, the proposed work introduces an end-to-end Deep Learning based framework for fake news detection tailored for cross-domain applications. The adoption of the Deep Learning (DL) paradigm [10] represents a natural solution to address the above issues, as DL techniques permit the learning of accurate classification models also from raw data (in our solution, the words composing the news) without requiring heavy intervention by data-science experts. Basically, these DL models are structured according to a hierarchical architecture (consisting of several layers of base computational units i.e., the

[1] https://www.bbc.com/news/blogs-trending-38156985

artificial neurons are stacked one upon the other), allowing for learning features at different abstraction levels to represent raw data.

In more detail, we explore and evaluate different deep learning based strategies (mainly Generative Networks) to learn transferable and discriminable feature representations for fake news detection. Moreover, we derive features that are domain-invariant and, thus, benefit the detection of fake news on newly arrived, emergent events for which only a few verified news are available.

The proposed solution is composed of three main components (i.e., neural models) that collaborate to solve two tasks simultaneously: the main one is to recognize fake information, and the second (auxiliary) task aims to produce domain invariant features. Preliminary experimentation conducted on two real datasets shows promising results and encourages further studies.

Organization of the paper. The remainder of the paper is structured as follows. In Section 2, we survey recent works concerning the cross-domain fake news detection problem; Section 3 introduces our approach and details the devised neural architecture, and Section 4 showcases numerical results. Finally, Section 5 concludes the paper and outlines possible future research directions.

2. Related Work

Numerous studies on automating fake news detection have been proposed in the recent few years. Most studies explore different supervised models with different modalities (e.g., *text*, *images*, and *propagation networks*) of news records to identify fake news. However, the performances of these existing state-of-the-art detection techniques significantly collapse if the news are coming from different domains (e.g., politics, gossip, medicine). In fact, while they perform well in the domain they were trained on (e.g., politics), they perform poorly in other domains (e.g., healthcare), especially for domains that are unseen or rarely seen during training.

News from different domains have significantly different word usage, specific communities of users involved in the news engagements, and also different propagation patterns. Furthermore, the models are biased toward event-specific features. Therefore, to address these challenges, it is key to learn models able to catch cross-domain information.

Cross-domain modeling refers to a model capable of learning from data in a certain source domain and being able to adapt and have a good performance on a different target one. A few previous works have attempted to perform fake news detection using cross-domain datasets. In this section, we overviewed some of the most significant approaches proposed in the literature.

In Wang et al. [11] an event discriminator has learned along with a multimodal fake news detector to overlook domain-specific information in news dataset. Specifically, the authors proposed a framework named Event Adversarial Neural Network (EANN), which can derive event-invariant features and thus benefit the detection of fake news on newly arrived events. In particular, an event discriminator measures the dissimilarities among different events, removes the event-specific features, and keeps shared features among events, setting up a minimax game with a multi-modal feature extractor to learn an event invariant representation, which can generalize well for the newly emerged events.

In [12] it is proposed an end-to-end model, named BERT-based domain adaptation neural network for multi-modal fake news detection (BDANN). BDANN comprises three main modules: a multi-modal feature extractor, a domain classifier, and a fake news detector. Specifically, the multi-modal feature extractor employs the pre-trained BERT model to extract text features and the pre-trained VGG-19 model to extract image features. The extracted features are then concatenated and fed to the detector to distinguish fake news. The role of the domain classifier is mainly to map the multi-modal features of different events to the same feature space, removing the event-specific dependency. The domain classifier actually performs domain adaptation. A limitation of this approach is that the domain classifier aims to classify the posts into one of a prefixed number of events, therefore it is not tailored for unseen events. Another drawback of the approach is that it does not account for the sequence or timing of the events.

In [13], crossdomain fake news detection is formulated 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 to address cross-domain fake news detection problems.

Silva et al. [14] proposed a multimodal fake news detection technique for cross-domain data that learns domain-specific and cross-domain information of news using two independent embedding spaces, which are then used to identify fake news. The framework consists of two main components: (1) unsupervised domain embedding learning and (2) supervised domain-agnostic news classification. These two components are integrated to identify fake news while exploiting domain-specific and cross-domain knowledge in the news. The unsupervised domain embedding learning technique exploits multimodal content (e.g., text, propagation network) to represent the domain of a news as a low-dimensional vector. The multimodal content is represented as a heterogeneous network that consists of both users tweeting the news items and words in the news title as a node. The unsupervised technique allows for selecting a set of unlabelled news, which can be used to train the fake news detection model that performs well for many domains while minimizing the labeling cost.

In [15] is proposed a deep architecture for cross-domain multimodal fake news detection, CrossFND (Cross-domain Fake News Detection). The approach exploits both cross-domain knowledge transfer and within-domain modeling of news content, user comments, and user-news interactions. The key idea is learning unsupervised feature representations and using them for domain adaptation. In the paper, the authors focused on feature-level domain adaptation and on learning a domain-independent textual representation such that a domain classifier is unable to detect the domain of the input text's latent representation. A key component of the approach is a domain classifier that detects the domain of a news content by finding feature-based differences among different domains. To train the domain classifier, a small portion of the target dataset is added on top of the source domain dataset. The domain classifier is a three-layer neural network similar to the fake news classifier, but with different weights and bias matrix. In order for the model to learn domain-independent features, the domain loss is maximized. Due to this, the model is forced to represent news information in a way that ignores domain-specific attributes in an effort to trick the classifier. The loss function of the fake news classifier is minimized to achieve the goal of accurately categorizing news as fake or authentic.

In [16] is proposed a multi-modal domain-adaptive approach that incorporates auxiliary information (e.g., user comments and user-news interactions) into a novel reinforcement learning-based model called REinforced Adaptive Learning Fake News Detection (REAL-FND). The approach extends the CrossFND framework proposed in [15] by modifying the domain classifier from a simple Multilayer Perceptron (MLP) into a more elaborated architecture based on reinforcement learning. Authors exploit reinforcement learning to transform the learned representation from the source to the target domain, ensuring that domain-specific features are obscured while domain-invariant components are maintained.

The majority of the approaches for cross-domain fake news detection mentioned above have two drawbacks: (1) it assumes that the news records from different domains arrive sequentially, even though this isn't always true for real-world streams; and (2) it necessitates knowing the domain of news records, which isn't always possible; (3) it is not able to capture newly emerging domains and handling temporal changes in domains. In contrast, our method learns cross-domain knowledge of news without being aware of the news' real domain.

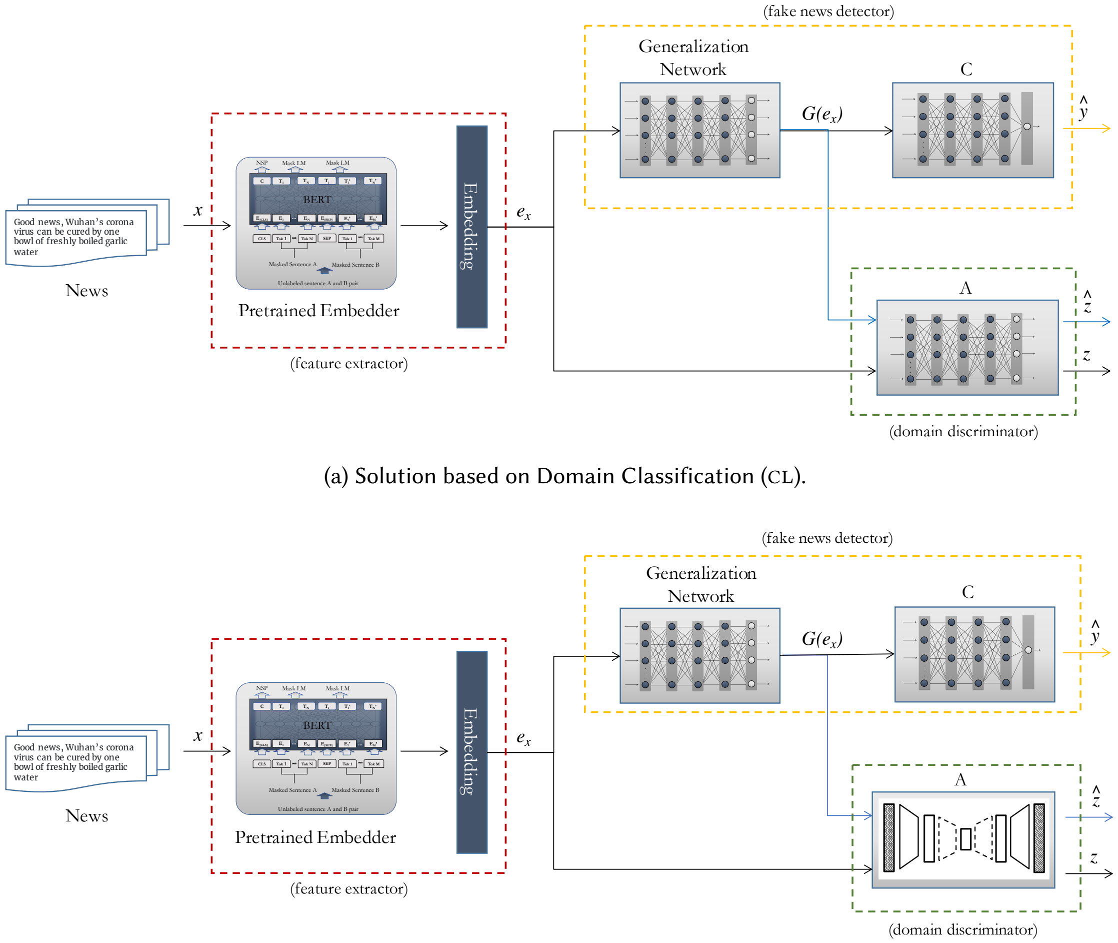
3. Self-Supervised Cross-Domain Generalization

In this section, we provide a detailed description of our approach based on DL to learn a detection model able to identify malicious/misleading information across different domains. The problem is particularly relevant in situations where fake news emerge in new contexts for which no prior evidence is available. Most of the existing approaches are designed for specific events and as a result, ineffective for emerging ones [3, 4, 5, 6, 7, 8, 9]. By contrast, we aim at devising solutions that are capable of generalizing discriminative signals for fake news, regardless of the underlying domain. The important challenge to address is how to enable feature representations that are discriminative when learning from a source domain and invariant with respect to the shift between the domains. For this purpose, we investigate feature-level adversarial learning combined with self-supervised information that can be obtained from the data, concerning domain characterization. The objective is to map the original data into a feature space where no boundaries can be detected among different domains, but boundaries can be sketched between fake and legit news.

Our proposal relies on a modular architecture based on feature extraction and embedding, domain generalization, and classification. Essentially, the general framework consists of three main components: *feature extractor*, *fake news detector* and *domain discriminator*.

The feature extractor maps an input x into a multidimensional representation e_x, which is a compressed mapping of different types of features (such as news content, images, user comments, propagation patterns, and user-news interactions). In our scheme, it is made of two components: the first component is an embedder used to encode the different modalities (e.g., Bidirectional Encoder Representations from Transformers (BERT), Graph Neural Networks (GNN), Video Graphics Array (VGA), etc.) for learning representations; while the second component combines all such representations.

In the current implementation, the embedder takes the form of a pre-trained BERT [17] instance. Essentially, BERT is a transformer-based neural architecture able to process natural language. It is trained through an algorithm including two main steps, named *Word Masking* and *Next Sentence Prediction* (NSP), respectively. In the former step, a percentage of the words



(a) Solution based on Domain Classification (CL).

(b) Solution based on latent space embedding (AE).

Figure 1: Learning architectures for cross-domain fake news detection.

composing a sentence is masked, and the model is trained to predict the missing terms by considering the word context, i.e., the terms that precede and follow the masked one. Then, the model is fine-tuned by considering a further task that allows for understanding the relations among the sentences. In our framework, we adopt a BERT instance pre-trained on Wikipedia pages. The fake news detector includes two modules integrated into a single architecture. The first module is a generalization network G that maps the combined embedding e_x into a generalized embedding $G(e_x)$. The second component is a binary classifier C that takes the generalized representation $G(e_x)$ as input and provides a classification response $\hat{y}$. The two networks work in a combined way: the purpose of the generalization network is to remove all domain-dependent features from e_x, in a way that allows the classification module to focus only on domain-invariant features.

The domain discriminator takes the form of the support network A, learned on an auxiliary task to characterize the source domain of the input e_x. The domain discriminator is exploited in an adversarial framework [18], for learning representations that enable domain adaption [11, 14]. Assuming that, for a generic input $\tilde{x}$, the response $A(\tilde{x})$ characterizes the source domain, the main idea is to exploit the generalization network $G(e_x)$ to deceive A. In practice, the characterization $A(e_x)$ should diverge substantially from the characterization $A(G(e_x))$. This should guarantee that domain information is obfuscated within the mapping G, while still maintaining the discriminating abilities through $C(G(e_x))$.

Specifically, we investigate two different architectural choices for A that enable such mechanisms. The main difference relies on how the self-supervision is combined with adversarial learning: in the first case, a classifier is adopted to model the discriminator while, in the second one, an autoencoder solution is used. Both these architectures are sketched in fig. 1. The first choice (devised in fig. 1a) exploits a pre-trained domain classifier. Here, we assume that given an input $\tilde{x}$, the response $A(\tilde{x})$ is a label characterizing the source domain. This requires that A is trained in order to be able to discriminate on the underlying topic characterizing $\tilde{x}$. In our framework, $\tilde{x}$ is either e_x or $G(e_x)$: in the former case, $A(\tilde{x}) = z$ and represents the actual source domain; in the latter, $A(\tilde{x}) = \hat{z}$ and is the predicted one. Then, given A, the whole network can be trained using the following loss:

$$\begin{aligned}\ell_{\mathsf{CL}}(x, y) =& BCE(\hat{y}, y) + \ell_o(z, \hat{z})\\ \hat{y} =& C(G(e_x))\\ \hat{z} =& A(G(e_x))\\ z =& A(e_x).\end{aligned}$$

Here, BCE represents the binary cross-entropy loss. The second term in the loss aims at penalizing the overlap between the distributions devised on both e_x and $G(e_x)$, through the A classifier. This can be achieved by exploiting a combination between Kullback-Leibler divergence and entropy on z and $\hat{z}$. In the following, we assume that A is a binary classifier and implement ℓ_o to penalize the certainty in classification on $G(e_x)$, as opposed to that on e_x:

$$\ell_o(z, \hat{z}) = \log(|.5 - \hat{z}|) + \log(z + \epsilon) + \log(1 - z + \epsilon).$$

The ϵ term in the formula bounds the contribution of z and avoids the discriminating capabilities on the original input embedding to make the contribution of the generalization network negligible. As we are considering a binary classifier, we use a constant value of 0.5 that represents the max uncertainty for the classification. In practice, the objective is to obtain a generalization that introduces maximum uncertainty in domain discrimination, as opposed to minimum uncertainty in the original embedding.

A major disadvantage of the above learning scheme is that the source domain's information must be known in advance (as a consequence, the domain discriminator A has to be pre-trained). In other words, the dataset has to be labeled with the domain from which the information was extracted. Although this is not necessarily an issue, an unsupervised approach that does not require prior knowledge of the source domain could be better suited. We explore this direction in the alternative architecture (shown in fig. 1b) that combines autoencoders with adversarial approaches for learning unsupervised feature representations about the domain and using them for domain adaptation. The main goal is to leverage more general latent feature representations characterizing a domain using an autoencoder that is trained by minimizing the reconstruction error within the source domain (without the need for the dataset to be labeled). The resulting latent feature is hence exploited in a minimax game between the domain discriminator and the fake news detector.

Formally, given input x and a label y, we tune the components of the network to map both e_x and $G(e_x)$ in the domain $[-1,1]^K$. Then, we alternatively train the components feature extractor/domain discriminator and fake news detector by means of the following losses:

$$\begin{aligned}
\ell_{\mathtt{AE},d}(x,y) =& BCE(\hat{y},y) + \log(2|e_x| - \|\hat{z} - G(e_x)\|^2)\\
\ell_{\mathtt{AE},g}(x,y) =& \log(\|e_x - z\|^2)\\
\hat{y} =& C(G(e_x))\\
\hat{z} =& A(G(e_x))\\
z =& A(e_x)
\end{aligned}$$

In practice, G learns to generalize the features of e_x by disregarding the components that characterize the source domain. The latter can be trained in an alternate fashion within the same learning framework, without the need to resort to supervised prior information about the domain. This guarantees full supervision and, at the same time, forces the detector to only focus on domain-invariant features. As regard the constant value 2, it represents the max difference between the outputs (e_x and $G(e_x)$) since we are adopting a `tanh` activation function.

4. Preliminary Experiments

This section describes preliminary experimentation conducted to test our solution. We introduce datasets, model parameters, and evaluation protocol, then we describe the adopted evaluation metrics, and finally, we discuss numerical results and show a qualitative analysis.

Datasets, parameters, and evaluation protocol. In our experimentation, we evaluated the quality of our solution by considering two real datasets extracted from the *FakeNewsNet* data repository [19, 20]. Basically, the gathered data focus on two main topics (i.e., politics and gossip) and are obtained from two fact-checking websites: *PolitiFact*[2] and *GossipCop*[3].

In Table 1, we report some relevant information concerning the two datasets: the overall number of articles, the vocabulary size, and some statistics on the number of words per article (i.e., average, median, first, and third quartile).

The neural architectures used in our experimentation mainly differ in the number of layers and neurons. Considering the solution depicted in Figure 1a, the Generalization Network is an MLP composed of three fully-connected layers (respectively, instantiated with 768, 256, and 768 neurons), whereas both the classifiers C and A are composed of two fully-connected

[2] https://www.politifact.com/
[3] https://www.gossipcop.com/

Table 1
Main features of the *PolitiFact* and *GossipCop* dataset.

Dataset	# Articles	# Classes	Vocabulary size	# Words per article			
				Avg.	Median	Q1	Q3
PolitiFact	814	2	60,870	817.5	199.5	66.5	530.7
GossipCop	4,719	2	149,557	349.0	205.0	109.5	323.0

layers (respectively, with 64 and 32 neurons) and an output layer equipped with a `sigmoid` activation function. As regards the solution shown in Figure 1b, the A model takes the form of a Variational Autoencoder (VAE). The encoder is composed of three layers, respectively, with sizes 768, 256, and 64. Symmetrically, the decoder includes the same number of layers and neurons. Also in this case, the classifier C includes two hidden layers but respectively, with 768 and 256 neurons and an output layer equipped with a `sigmoid` activation function. Finally, the Generalization Network is an MLP composed of three layers (instantiated, respectively, with size 768, 256, and 768). Unless otherwise specified, each layer is equipped with a `tanh` activation function.

The predictive performances exhibited by our solution are compared with a simple baseline neural model, initialized with the same architecture and parameters of the fake news detector component, and fed with the same embedding. In more detail, to evaluate the cross-domain generalization capability of our approach, we adopted the following protocol: each model is trained against a single domain training set and tested against the data from another one.

Lastly, the experiments were executed on an NVidia DGX Station equipped with 4 GPU V100 32GB. The model was learned by optimizing the weights in batches of 32 texts from the training set using the Adam optimizer with a learning rate $lr = 0.001$.

Evaluation Metrics. To assess the detection capabilities of the proposed approach, we computed different performance metrics for both test cases. First, let us define TP as the number of positive cases correctly classified, FP as the number of negative cases incorrectly classified as positive, FN as the number of positive cases incorrectly classified as negative, and TN as the number of negative cases correctly classified. Then, the following metrics have been considered:

- *Accuracy*: defined as the fraction of cases correctly classified, i.e., $\frac{TP+TN}{TP+FP+FN+TN}$.
- *Precision* and *Recall*: used to estimate a system's detection capability in identifying attacks and avoiding false alarms. Specifically, *Precision* $= \frac{TP}{TP+FP}$, while *Recall* $= \frac{TP}{TP+FN}$.
- *F-Measure*: summarizes the overall system performances and is calculated as the harmonic mean of *Precision* and *Recall.*

Notably, *F-Measure*, *Precision* and *Recall* are computed by using a Macro-Averaging strategy i.e., the metrics are computed for each class and then averaged. We adopted macro-averaging, which weights all the classes equally, to summarize with a single value the performances of both classes.

Numerical Results. In Table 2, we report the experimental results obtained by comparing our proposed framework with the baseline introduced above. The table shows the results obtained by using the evaluation protocol introduced in the previous paragraph, which consists in learning the model on a specific domain (Training Set) and validating it on another one (Test Set). `Our Solution (AE)` refers to the proposed architecture adopting an autoencoder as discriminator, while `Our Solution (CL)` indicates the setting in which the discriminator is a classifier.

Bold values represent the best results obtained on the test set. A cross-validation strategy [21] is used to select the best model in the learning stage.

Basically, the adoption of our solution allows for improving the predictive capability of the fake detector for each configuration and evaluation metric. These preliminary results highlight that the proposed framework permits the extraction of cross-domain features, which improves the generalization capabilities of the detector.

Table 2
Predictive performance metrics. Bold values represent the best results for each scenario.

Neural Model	Training Set	Test Set	Accuracy	Precision	Recall	F-Measure
`Baseline (MLP)`	*Politifact*	*GossipCop*	0.477	0.476	0.476	0.475
	GossipCop	*Politifact*	0.497	**0.568**	0.535	0.448
`Our Solution (CL)`	*Politifact*	*GossipCop*	**0.526**	**0.580**	**0.556**	**0.501**
	GossipCop	*Politifact*	0.518	0.520	0.520	0.518
`Our Solution (AE)`	*Politifact*	*GossipCop*	0.511	0.574	0.545	0.475
	GossipCop	*Politifact*	**0.571**	0.562	**0.557**	**0.553**

Although the preliminary results are promising, it is worth pointing out that the study has been carried out by exploiting only the textual content of social media posts. The obtained results are in line with the ones achieved by state-of-the-art models that rely only on the textual content of news and mainly address single-domain scenarios, while the proposed solutions generalize across domains.

However, the predictive performances of the detector could further benefit from using social media-rich data, and exploiting not only textual content but also temporal, geographical, and network information, enabling the development of novel fake news detection solutions by integrating data of different natures.

5. Conclusion and Future Work

In this work, we designed a framework for detecting fake news using the deep learning paradigm that effectively addresses a major issue in this field i.e., recognizing misinformation across different domains by exploiting one learning model able to generalize through the domains. The framework combines in a single architecture different neural components and tries to solve simultaneously two tasks: *(i)* extracting cross-domain features, and *(ii)* learning a detector able to distinguish real/fake news. Specifically, a BERT model pre-trained on Wikipedia data is used to generate an embedding from the news text, then this compressed representation feeds two other neural networks: the former performs the detection while the latter solves an auxiliary task so as to yield cross-domain invariant features.

Future works aim at refining the proposed framework. Part of our ongoing research is devoted to evaluating the proposed architecture on heterogeneous data. Multi-modality is a key approach to improving fake news detection. It is more difficult and more challenging to detect fake news on multi-modal inputs, as it requires not only the evaluation of each modality but also cross-modal connections and the effective combinations of the different inputs. This becomes even more challenging when each modality e.g., text or image is credible but the combination creates misinformative content.

We aim to exploit multi-modal data like images, video, news propagation patterns, social context representing the user engagements of news on social media (e.g. the number of followers, hashtags, friendship networks, retweets).

In fact, information such as “from who” and “how many times” the news has been replied, quoted or shared could represent further precious indicators to reveal the malicious nature of the news. In this respect, *Graph Neural Networks* (GNNs) have proven to be an effective tool to operate on the graph domain and summarize the social network properties into a single graph embedding.

Another method of strengthening multi-modal models for fake news detection is to augment them with features based on the behavior of news producers. News producers copy news stories from each other for easy engagement or to increase the perceived credibility of stories. The idea of exploiting features of media sources has been only partially explored in literature. We plan to address this gap by investigating alternative solutions like introducing a set of veracity features to characterize the sources based on content sharing behavior.

Acknowledgments

This work has been partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

References

[1] O. A. Aghababaeian H, Hamdanieh L, Alcohol intake in an attempt to fight covid-19: A medical myth in iran, Alcohol 88 (2020) 29–32.

[2] C. Liu, X. Wu, M. Yu, G. Li, J. Jiang, W. Huang, X. Lu, A two-stage model based on bert for short fake news detection, in: C. Douligeris, D. Karagiannis, D. Apostolou (Eds.), Knowledge Science, Engineering and Management, Springer International Publishing, Cham, 2019, pp. 172–183.

[3] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, Defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’19, 2019, p. 395–405.

[4] C. Raj, P. Meel, Arcnn framework for multimodal infodemic detection, Neural Networks 146 (2022) 36–68.

[5] T. Sachan, N. Pinnaparaju, M. Gupta, V. Varma, Scate: Shared cross attention transformer encoders for multimodal fake news detection, in: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '21, 2021, p. 399–406.

[6] R. Kumari, A. Ekbal, Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection, Expert Systems with Applications 184 (2021) 115412.

[7] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2017, p. MM '17.

[8] Q. Jing, D. Yao, X. Fan, B. Wang, H. Tan, X. Bu, J. Bi, Transfake: Multi-task transformer for multimodal enhanced fake news detection, in: IJCNN, 2021, pp. 1–8.

[9] J. Wang, H. Mao, H. Li, Fmfn: Fine-grained multimodal fusion networks for fake news detection, Applied Sciences 12 (2022).

[10] Y. Le Cun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[11] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, Association for Computing Machinery, 2018, p. 849–857. URL: https://doi.org/10.1145/3219819.3219903. doi:10.1145/3219819.3219903.

[12] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, L. Cui, Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8. doi:10.1109/IJCNN48605.2020.9206973.

[13] Y. Han, S. Karunasekera, C. Leckie, Graph neural networks with continual learning for fake news detection from social media, 2020.

[14] A. Silva, L. Luo, S. Karunasekera, C. Leckie, Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data, 2021. URL: https://arxiv.org/abs/2102.06314. doi:10.48550/ARXIV.2102.06314.

[15] K. Shu, A. Mosallanezhad, H. Liu, Cross-Domain Fake News Detection on Social Media: A Context-Aware Adversarial Approach, Springer Nature Singapore, Singapore, 2022, pp. 215–232. doi:10.1007/978-981-19-1524-6_9.

[16] A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido, H. Liu, Domain adaptive fake news detection via reinforcement learning, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, 2022, p. 3632–3640. URL: https://doi.org/10.1145/3485447.3512258. doi:10.1145/3485447.3512258.

[17] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144.

[19] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media, arXiv preprint arXiv:1809.01286 (2018).

[20] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[21] C. Schaffer, Selecting a classification method by cross-validation, Machine Learning 13 (1993) 135–143.