IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 February 2023, accepted 25 February 2023, date of publication 8 March 2023, date of current version 13 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251892

RESEARCH ARTICLE

Explainable Misinformation Detection Across Multiple Social Media Platforms

GARGI JOSHI[1], ANANYA SRIVASTAVA[1], BHARGAV YAGNIK[1], MOHAMMED HASAN[1], ZAINUDDIN SAIYED[1], LUBNA A. GABRALLA[2], AJITH ABRAHAM[3], (Senior Member, IEEE), RAHEE WALAMBE[1,4], AND KETAN KOTECHA[1,4]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2]Department of Computer Science and Information Technology, College of Applied, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[3]Faculty of Computing and Data Sciences, FLAME University, Lavale, Pune, Maharashtra 412115, India
[4]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India

Corresponding authors: Rahee Walambe (rahee.walambe@sitpune.edu.in) and Ketan Kotecha (director@sitpune.edu.in)

This work was supported in part by Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, through the Researchers Supporting Project PNURSP2023R178; and in part by the Scheme for Promotion of Research Collaboration (SPARC) of the Ministry of Human Resources Development (MHRD), Government of India, under Sanction ID. P571.

ABSTRACT Web Information Processing (W.I.P.) has enormously impacted modern society since a huge percentage of the population relies on the internet to acquire information. Social Media platforms provide a channel for disseminating information and a breeding ground for spreading misinformation, creating confusion and fear among the population. One of the techniques for the detection of misinformation is machine learning-based models. However, due to the availability of multiple social media platforms, developing and training AI-based models has become a tedious job. Despite multiple efforts to develop machine learning-based methods for identifying misinformation, more work must be done on developing an explainable generalized detector capable of robust detection and generating explanations beyond black-box outcomes. Knowing the reasoning behind the outcomes is essential to make the detector trustworthy. Hence employing explainable A.I. techniques is of utmost importance. In this work, the integration of two machine learning approaches, namely domain adaptation and explainable A.I., is proposed to address these two issues of generalized detection and explainability. Firstly the Domain Adversarial Neural Network (DANN) develops a generalized misinformation detector across multiple social media platforms. DANN generates the classification results for test domains with relevant but unseen data. The DANN-based, traditional black-box model cannot justify and explain its outcome, i.e., the labels for the target domain. Hence a Local Interpretable Model-Agnostic Explanations (LIME) explainable A.I. model is applied to explain the outcome of the DANN model. To demonstrate these two approaches and their integration for effective explainable generalized detection, COVID-19 misinformation is considered a case study. We experimented with two datasets and compared results with and without DANN implementation. It is observed that using DANN significantly improves the F1 score of classification and increases the accuracy by 3% and A.U.C. by 9%. The results show that the proposed framework performs well in the case of domain shift and can learn domain-invariant features while explaining the target labels with LIME implementation. This can enable trustworthy information processing and extraction to combat misinformation effectively.

INDEX TERMS Covid 19, DANN, lime, misinformation detection, social media, text processing, web information processing, XAI.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

I. INTRODUCTION

Information and data are primarily stored on various social networks and internet platforms, and web information processing offers opportunities to modify and extract the data

[1]. Web Information Processing (W.I.P.) manipulates data available from various internet/web sources to produce useful information. One of the specific aspects of web information processing includes extracting and using the information available on social media, including the internet and social networking websites. A considerable percentage of the population relies on the internet to acquire information. Social Networking platforms are growing daily, with 4.2 billion users in January 2021 and a 13.2% increase in 2020 alone. The rate doubled compared to 2019-2020, only 7.2% [2]. Due to this spike, the enormous amount of misinformation communicated on social media platforms poses an unprecedented challenge causing significant harm and deleterious to many people worldwide. Misinformation can be any inadvertent premise not backed by facts or scientific data and leads to misconceptions. This is highly important in the case of healthcare-related misinformation, which can lead to fatal consequences. Furthermore, the circulation of misinformation during a health crisis induces trepidation among the population. For example, detail about COVID-19 from these social media posts and news sources can include opinionated ''natural remedies,'' the origin of COVID-19, or about the vaccines and their side effects cascade hesitancy in getting vaccinated even if vaccines are available.

The type of language and ideas on a specific topic portrayed by the users from one social media platform to another significantly differs, starting from vocabulary, grammar, etc. [3]. This also differs from news data sources. Annotating data from multiple information pools can take time and effort. This is a major reason many researchers focus on one specific data source when classifying misinformation/fake news. In the case of a pandemic, dependency on a particular source drastically affects the robustness of a classification model to be applied on other social networks when spreading misinformation across all platforms is gradually increasing and equally important. The motivation of the work is drawn from the wide use and spread of social media platforms, which are a huge source of influence for spreading misinformation as the information hails from multiple sources, in other words, multimodal homogeneous data sources providing rich contextual information. Detection of misinformation is crucial to human mental and social well-being to reduce the confusion and ambiguity arising from misinformation spread. Despite the use of different A.I. and machine learning-based techniques for the detection of misinformation, there is a huge need to have a generalized detector for misinformation detection across multiple social media platforms and have an explainer for explaining the justifications for the predictions to ensure trust and adoption of the system for combating misinformation effectively in the healthcare domain.

Creating an adaptive model over diverse media platforms can be done effectively by learning information from one domain and utilizing it to learn in another domain. Domain Adaptation [4] deals with such generalization beyond training distribution and comes under the purview of out-of-distribution generalization [5]. This is accomplished using Domain Adversaries to learn domain invariant feature representations. DANN applies the gradient reversal layer to make the feature distribution of source and target domains similar. Data distribution differs among social media platforms due to user behavior; data from different platforms constitute different domains. Here, the domain doesn't refer to a particular industry vertical but to different social media data sources. For example, comments posted on Instagram represent a domain, while news articles published on various platforms represent different domains. The dataset of comments from the different social platforms and news sources are considered as different domains because the interacting group of audience with the platform differ in thought processes and user behavioral patterns. In this work, we have considered data across multiple social media platforms such as YouTube, Instagram, news, and Reddit for covid 19 misinformation detection. Data distribution differs across social media platforms due to user behavior, as data from different platforms constitute different domains. Creating an adaptive model over diverse media platforms can be done effectively by learning information from one domain and utilizing it to learn in another domain. So this is a cross-domain adaptation, where the features of the domains are the same, but the underlying distribution is different. In this case, the domain shift is based on two domains having the same input feature space ($\chi s = \chi$), but the shift is because of different data distributions, i.e., $P(Xs) \neq P(Xt)$. Domain dimensions are also the same. Where a domain is defined as the same input features but coming from a different distribution. So the overall work focuses on curating data from multiple sources of social media platforms and adaptation to different domains due to data distribution differences and building a generalized detector with the DAAN model, i.e., Domain adversarial neural network, and explaining the predictions class/target labels of the DANN model with Explainable AI-based techniques.

However, it is important to note that the DANN-based approach is a traditional black-box model, and the outcome, i.e., the reasons for the generated target labels, are unknown. Despite multiple efforts to develop machine learning-based methods for identifying misinformation, little work has focused on providing explanations beyond a black-box decision [6]. Explores must be provided for the target labels to develop a trustworthy generalized detector for social media misinformation. Hence in this work, we employ a DANN and implement it to detect misinformation in coronavirus-related posts across these diverse domains, followed by a Local Interpretable Model-Agnostic Explanations (LIME) [7] framework to generate the reasoning behind the outcomes. So the objectives of the study are

1. Can we train a classifier on one source domain and use it to test a similar but unseen target domain using machine learning?
2. Can we develop an explainable model integrated with the generic detector to explain the target domain labels?

For coming to Section II will discuss the related work on detecting misinformation using artificial intelligence and machine learning, domain adaptation, and explainable artificial intelligence for web information processing. The datasets used for the study, namely Misovac, and CoAid, are described in section III. The methods, such as preprocessing, data augmentation, and DANN on the CoAID and MiSoVac data sets, are described in section IV. Section V describes the system design and architecture, especially for DANN and LIME. The results and experiments section VI describes the performed experiments, corresponding results, and evaluation metrics, followed by the result discussion section. The article ends with the conclusion section and a discussion on the future scope of the study.

II. RELATED WORK

This work reviews the literature concerning three aspects—A.I. and machine learning-based misinformation detection, domain adaptation, and explainability through the subsequent subsections. The W.I.P. consists of information extraction and making it available for internet users. The internet has become a tool for generating and free flow of information worldwide in today's world. Information generates ideas and drives decisions. However, the internet generates false information since no regulatory body moderates the content, especially on social media and W.I.P., impacting society [8]. Air and sound pollution has been a major concern for the last few decades; however, now is the time to worry about information pollution. Information pollution's direct reasons and impacts are difficult to identify and explain and even more challenging to quantify. In [9], the information disorder phenomenon is examined comprehensively. In many instances, there is no malicious intent to generate and spread misinformation; however, in other cases, it might be just very selfish, e.g., increasing the sale of a certain drug or purely being business-oriented, to spread wrong information to create panic about certain drugs or treatments. Understanding the ethical and moral status of the people involved is interesting. When people cannot tell what is credible and what is not and act on that information, poor decisions can impact our lives and financial well-being [10]. We must recognize that communication and information sharing plays a significant role in representing shared beliefs. Since social platforms are designed to express through likes, comments, and shares, all the efforts towards fact-checking and debunking false information are ineffective since the emotional aspect of sharing information is impossible to control. The mining of misinformation in social media spreads uncontrollably and tremendously fast and, in recent times, has been responsible for causing harm to the social fabric of our world [11]. Misinformation can be disinformation, rumors spam, fake news, etc. The world economic forum considers the rampant spread of misinformation online one of the ten global risks [12].

Regarding misinformation, ''an era of fake news'' is occurring rapidly, where misinformation is transmitted speedily, intentionally, or unintentionally. Various formats of non-textual media, such as bit-mapped pictures, are being used, contributing to the diffusion of misinformation and disinformation. Misinformation affects communities in various ways, for instance, racist hostility and exclusion. Therefore, preventing such emerging behaviors is an important area of research and study. Pizzagate, the Anti-vaccine movement, Russian scientists discovered a cure for homosexuality, etc., are some of the popular fake news items of 2017 [13], [14], suggesting that curbing social bots may be an effective strategy for mitigating the spread of online misinformation. It is also important to understand that although much misinformation is focused on the political domain, medical misinformation has threatened countries worldwide. Research has demonstrated how inaccurate advice from a person who has no medical knowledge is proliferated through hoaxes, tweets [15], online Q&A forums, Pinterest [16], Yahoo, and Google [17]. The context of medical information encompasses the broader aspects of trustworthiness, reliability, dependability, integrity, and reputation of the medical practitioner and the A.I. developer in the high-risk health and safety domain. Due to the explosion and wide acceptance of social media reporting, the issue of non-credible news has become extremely relevant. In the context of medical information, this problem is even more serious because it directly relates to the health and well-being of people. Medical misinformation is an obvious concern as the information can be shared without rigorous review. The first step toward handling the spread of misinformation is to detect it automatically. Machine learning and artificial intelligence methods are employed for this task, especially for automatic misinformation detection on social media platforms.

A. DETECTION OF MISINFORMATION USING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Earlier efforts on misinformation detection go back to the start of the internet revolution Kinchla, and Atkinson [18] have studied the effect of false information on psychophysical judgments. Their experimental results show that false information reduces the probability of a correct response. Various applications, including credibility assessment of microblogs [19], have been reported recently. The credibility assessment algorithms are automated, human-based, and hybrid approaches. Automated approaches include various machine learning approaches. Human-based can be the voting, cognitive, and manual verification approaches. Hybrid approaches combine these. Shao et al. [19] have developed Hoaxy, an open platform for dealing with misinformation. The huge scale of information (data), dynamic nature, and homophily are primary challenges in detecting misinformation. To this end, various researchers have studied and developed methods to see false information. With the flood of information from social websites, it is impossible to check, analyze and vet the potential deception. Various methods have been reported, from simple classifiers to state-of-the-art attention networks [20], [21], [22], [23], [24]. A recent EEG-based emotion recognition study using a hybrid CNN and LSTM classification approach was conducted in [25]. A line of research models social media problems as an optimization problem for sentiment analysis on social networks [26]. A text-based deep learning model for hate speech detection for covid 19 is developed in [27]. However, these approaches are explicitly reported for a single social media platform and information classification. No method can be used across various platforms. This work considers a state-of-the-art domain adaption paradigm for developing a misinformation detector trained on one source domain and adapted to multiple similar domains considering the domain difference with explainable predictions.

B. DOMAIN ADAPTATION

Domain adaptation is classified as a transductive learning procedure [28] under transfer learning [29]. It is applied in problems where the task stays the same across domains; however, data may not be available in the target domain. Initially, domain adaptation setups considered this problem a supervised or semi-supervised approach with substantial sources and scant data in the target domain [30], [31] proposed solutions to the domain shift problem using such setups. However, with the difficulties in annotated data, unsupervised domain adaptation techniques emerged that alleviated the domain shift problem. Domain adaptation techniques are classified into data-centric, model-centric, and hybrid methods [32]. Data-centric approaches include methods like pseudo labelling [33], [34], [35] and Data Selection [36], [37].The model-centric approach focuses on modifications in the architecture that may include the use of pivot-based methods for feature augmentation [29], [38], [39] that have now been developed to learn from neural networks [40], [41] via attention [42], [43]. Feature generalization techniques are capable of learning common hidden features between domains, and such approaches include stacked de noising auto-encoders [44] and marginalized stacked de noising auto-encoders [45], [46], and Dual representation based auto-encoder [47]. Reference [48] proposed considering features from either source or target using domain separation networks; however, they lacked domain-specific features in the classifier as they were only used in the decoder. Reference [49] Proposed a gradient reversal technique to maximize domain confusion and minimize task error. DANN-based approaches have widely been applied to applications like sentiment classification [49], [50], [51], language identification [52], duplicate question detection [53], etc. [54] applied domain adversarial training leveraging the knowledge distillation [55] with an extra loss during adaptation. Hybrid approaches like [51], [56], [57], [58] combine data-centric and model-centric approaches. Unlike DANN, a loss-centric approach, feature-centric and data-centric techniques are used in text classification but do not use context-dependency and linguistic information and are multi-shot training procedures [31]. Also, there has been very little research tackling domain differences across social media platforms. In this work, we employ DANN to develop a generic misinformation classifier across multiple platforms. We consider one or more social media platform(s) as the source domain and other multiple platforms as target domains and apply DANN to demonstrate our approach. The most relevant case study of COVID-19 misinformation is considered for the demonstration. However, due to the black-box nature of the DANN, it isn't easy to generate explanations for the target labels. Hence to that end, the explainable A.I. is essentially employed.

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR WEB INFORMATION PROCESSING

Machine learning models have recently achieved tremendous progress in performance and accuracy. Still, the complex non-intuitive hidden layer processing makes them an opaque black box with a lack of insights on how and why a model generated a certain decision/outcome [59]. This black-box nature results in little or no understanding of the model's internal logic, adversely affecting these models' trust, usability, and adoption in real-world applications [60]. Model interpretability or explainability is of paramount importance when it comes to debugging the model for flaws, ensuring trust, transparency, accountability, and ethics in the model's outcomes, and complying with the governance such as EU-GDPR, which allows the end-user to seek the right to explain automated algorithmic decisions. XAI improves model performance by performing internal audits and bias detection [61]. The natural language processing (NLP) text models typically deployed for detecting misinformation on large-scale social media platforms demand accurate introspection and justification of the model's underlying predictions to ensure trust, transparency, and fair decision-making from different stakeholders [62]. A prime concern is to generate human-understandable comprehensive model explanations to interpret the model predictions and underlying mechanics [63]. State-of-the-art machine learning models are complex black boxes facing accuracy interpretability trade-offs raising concerns about models' reliable and reasonable behavior in the real world [64]. Reliability on mere performance and accuracy metrics needs to be improved. Interpretability is a prime metric for establishing trust in the inner workings and logic for accurate model predictions and decision-making [65]. Model interpretability can be achieved intrinsically by designing faithful and consistent explanations of the model or post hoc, i.e., explaining the predictions after model building without compromising the accuracy interpretability trade-off.

The model's scope is local to a specific instance or globally applicable to the entire model [66]. Model explainability techniques are broadly classified into intrinsic InDesign and post hoc explainability techniques that are irrespective of the underlying model and are derived from post-model development without impacting the accuracy and design of the model [67]. The most widely used techniques for post hoc

TABLE 1. Key comparison metrics with existing techniques.

Existing Approaches	Method Used	Limited to	Pitfall
[20]-[24]	Attention-based	Applicable to the single source social media platform	No generalization across multiple social media platforms The prediction remains a black box
[26]	Optimization problem		
Our Approach	DANN based	Explainable and generalizes across multiple social media platforms	

explainability in the literature are LIME (Local Interpretable Model Agnostic Explanations) and SHAP (Shapely Additive Explanations), Deep Shap, Deep Lift, and Cxplain). Explainability techniques are crucial in overcoming the black-box nature of classifiers for unimodal and multimodal deep neural nets for vision and language processing [68].LIME is a model-agnostic explainability technique applied to any classifier. The model is learned by perturbing the input data samples and understanding how the predictions change based on the input changes. LIME modifies a single data instance by tweaking the individual feature values and observing the effect on the model outcome. This provides insights into why a specific prediction was made or which feature contributed to the prediction. Feature relevance depicts the importance of individual features on the model predictions and the influence of different features over the model outcomes.

To interpret which model captures linguistic knowledge and semantic details and why a certain prediction is made, explanations can be derived for the predictions focusing on perturbed inputs [69]. This approximates the underlying classifier model with a second model learned by perturbing the original instance. This enables one to identify input components with the most significant influence or impact on predictions. This approach is model agnostic, and it is easier to learn explanations on a locally weighted dataset than approximate a model globally. Local Interpretable Model-Agnostic Explanations (LIME) [70] generate locally faithful explanations and learn an interpretable model locally around the specific prediction. LIME provides interpretable data representations for non-expert users representing the presence and absence of faithful and consistent words to the local model without impacting model performance. Table 1 lists the comparison metrics with existing work based on the survey.

In summary, the novelty and contribution of this work are:

1. Development of a classifier that can effectively be used across multiple social media platforms for misinformation classification with limited training data. Development of economic approach regarding time, processing, and efficiency avoids training models on individual platforms.
2. Implementation of the DANN-based architecture that outperforms the state-of-the-art results on the CoAID dataset.
3. The development of a novel misinformation dataset (MiSoVac) related to COVID-19 vaccination from social media platforms.
4. An explainable, trustworthy adoption paradigm addresses domain adaptation and explainability in integrating multiple social media platforms.

III. DATASET DESCRIPTION

Misinformation is spread in many forms, including newspapers, television, and the internet. However, most misinformation is spread across social media domains like Twitter, Instagram, Facebook, YouTube, Reddit, and Whats App. For this research, we focused on developing a generic misinformation detection with a specific topic of COVID-19; curation of the MiSoVac dataset, which includes data from multiple social media platforms related to vaccines with CoAID dataset (openly available) to demonstrate the DANN-based explainable approach for classification. The use case in this work is for covid 19 misinformation detection across multiple social media platforms. Since the CoAid dataset is the largest dataset for covid 19 misinformation available in the public domain, it is considered the source domain to learn the features from source to target. The MiSoVac dataset is created by curating data from multiple sources and is referred to as a target domain for domain adaptation.

A. CoAID

COVID-19 healthcare misinformation Dataset (CoAID) [71] includes 4251 news articles, 296000 related user engagements, and 926 social platform posts fact-checked by verified fact-checking sites. The data set is collected from articles published in December 2019 to September 1, 2020. Topics like COVID-19, coronavirus, pneumonia, flu, lockdown, stay home, quarantine, and ventilator were covered for the data set.

B. MiSoVac

This dataset was explicitly developed to focus on the case study of COVID-19 vaccine-related misinformation. Therefore, we collected the COVID-19 vaccine-related misinformation data (MiSoVac) from social media sites like Twitter, Instagram, YouTube, and Reddit from November 2020 to February 2021. We used the Selenium library in python to create a custom web scraper for Twitter, Instagram, YouTube, and Reddit, which can scrape comments/textual content based on account I.D., Hashtags, and date time. For scraping comments on posts related to Covid-19 vaccine engagement, we fetched all posts with Covid-19 vaccine engagement maintaining a list. Next, we iterate over each link to fetch the comments using selenium and export the fetched comments into a CSV file for further analysis and processing. Samples of the MiSoVac dataset are enlisted in Table 2. The 'None' class of the MiSoVac dataset is not used for Training or testing as it is insignificant. Only the Misinformation classes 'True' and 'False' were used for training and testing purposes;

TABLE 2. Samples from the MiSoVac dataset.

Misinformation type (True)	Misinformation type (False)	Misinformation type (None)
Twitter		
@WHO Solidarity Trials are also underway in many countries. Once these projects are complete, controlling #COVID19 will be more accessible due to the availability of proper medicines, vaccines, etc., to control infectious diseases worldwide.	An article claims that "Bill Gates' vaccine" would modify human D.N.A.	-
Reddit		
The antibodies decrease by a factor of six for S.A. Does anyone know what this means/does	If youâ€™ve had the initial variant, are the antibodies no longer effective against new variants?	And if it's suitable for Moderna, it's likely good for Pfizer. The good news today. See you in the next variant story
News		
People who are pregnant, breastfeeding, or want to become pregnant can get vaccinated against COVID-19. But they should talk to their medical provider.	COVID vaccines made with R.N.A. are not vaccines, but a gene therapy that could turn us into transgenic beings or cause us diseases	
YouTube		
All vaccines are experimental. With no long-term effects established, one should understand that the Pharmaceutical company is not liable if severe complications or death occurs. It's a take-at-your-risk option. …	The Vaccine has been weaponized!	Great video, Vox! The animation, the music, and the narration were all well-made.
Instagram		
@jamesr.french: I do not think so, but I know the symptoms may last a month or two after being infected with the COVID-19 virus.	😂😂😂😂 biggest scam in history. No flu deaths, no pneumonia deaths, no influenza deaths... just "COVID." 😂😂😂 wtfe! It takes 5 to 10 years to come up with a vaccine before it ever goes to market, yet all of a sudden, in just 5 to 7 months, we have a life-saving vaccine. 😂 the same Vaccine that has killed people….	Sick and tired of this already, till when most of us continue living like this

hence, the explainable A.I. model is built to justify the target labels of the predictions for two classes, i.e., true and false. And Table 3 shows the data distribution of the MiSoVac dataset.

IV. METHODS

A. PREPROCESSING

The information on social media is usually easygoing and casual, leading to a decrease in the ability of a language model to comprehend the corpus; consequently, performing broad preprocessing on the information became necessary [72]. Fig 1 depicts the diagrammatic flow for preprocessing. The underlying pipeline expanded contractions and truncations into their standard form. Basic pronouns, conjunctions, articles, and relational words in English vocabulary usually add no logical importance to a sentence and are disregarded by search engines; hence were eliminated from the corpus. The expulsion of URLs, hashtags, mentions, and punctuations was completed as a part of preprocessing. The emoji's were supplanted with the content indicating its importance.

TABLE 3. MiSoVac data distribution of various social media platforms.

Source	No. of Samples Labelled True (contain correct information)	No. of Samples Labelled False (contain misinformation)	No. of Samples labeled None. (contain unimportant information)	Total
Twitter	182	346	0	528
Reddit	39	7	322	368
News	149	185	0	334
YouTube	4	10	484	498
Instagram	30	59	2336	2425

B. DATA AUGMENTATION

The data and target labels in the MiSoVac dataset were imbalanced, causing the model to be incapable of generalizing on both classes. Due to this, augmentation had to be performed to balance the data and increase the data diversity. Table 4 shows an example of a sentence augmented from the MiSoVac data set. The following operations were performed using [73]:

1. Augmenting words by feeding neighboring words to the BERT language model leverages contextual word embedding.
2. Translation of text into other languages and then translating back to English sentences. The following languages were used in this approach: French, Japanese, German, and Urdu.

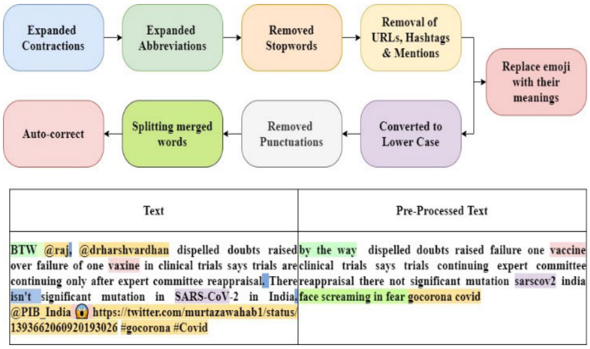
The details of the model components, such as embedding, layers, and the DANN model, are listed below.

1) GloVe

Global Vectors or GloVe [74] is an unsupervised learning algorithm for acquiring word vector portrayals. This is accomplished by projecting words into a significant space where the distance between words is identified with semantic similarity. Training is performed on a collected worldwide word-word co-occurrence matrix from a corpus, with features portraying fascinating linear substructures of the word vector space.

2) LSTM

LSTMs or Long Short-Term Memory [75] networks are derived from the Recurrent Neural Networks (R.N.N.s) that



Text	Pre-Processed Text
BTW @raj, @drharshvardhan dispelled doubts raised over failure of one vaxine in clinical trials says trials are continuing only after expert committee reappraisal. There isn't significant mutation in SARS-CoV-2 in India. @PIB_India 😱 https://twitter.com/murtazawahab1/status/1393662060920193026 #gocorona #Covid	by the way dispelled doubts raised failure one vaccine clinical trials says trials continuing expert committee reappraisal there not significant mutation sarscov2 india face screaming in fear gocorona covid

FIGURE 1. Preprocessing pipeline and an example.

TABLE 4. Example of data augmentation.

Data	Augmented Data
Last night I closed my eyes and saw that people in the hospital room had been vaccinated with cod.	I closed my eyes last night. I've seen people get a COVID vaccine in the hospital room.

can process data sequences. While R.N.N.s can learn the context, due to backpropagation, there is an issue of vanishing or exploding gradients, which LSTM can overcome. LSTM unit comprises multiple gates. These gates regulate the flow of information and allow some relevant information to pass through the sequence. So this way, the model remembers the information it has learned in the beginning. Therefore, 128 units of LSTMs were used in the architecture.

3) DANN

Domain Adversarial Neural Network (DANN) is a representation learning approach where the training and testing data sets are similar but belong to different distributions. The main inspiration behind the working of DANN is that for effective domain transfer to be achieved, predictions must be made based on overlapping features. Due to this, the network cannot discriminate between the training (source) and testing (target) domains [49]. Consequently, it does not require a labeled target domain dataset when using DANN but requires a labeled source domain data set. The DANN architecture consists of mainly three parts – feature extractor, label predictor, and domain classifier. The feature extractor consists of a deep neural network for performing feature-based learning. The label predictor and the feature extractor form a standard feed-forward neural architecture responsible for classifying the class label of the training data sample. Lastly, the domain classifier is accountable for achieving the unsupervised nature of DANN, in which it is connected to the feature extractor through the Gradient Reversal Layer (G.R.L.). G.R.L. has no parameters to be updated and acts as an identity transform during the forward propagation. Backpropagation multiplies the gradient obtained from the next layer by −1 and passes it to the previous layer, as mentioned in Eqn. 1. This layer ensures that the feature distributions over the source and target domains are similar as possible to obtain the domain invariant features. During training, those parameters of feature mapping are sought, which maximizes the loss of the domain classifier by making the source and target distributions as similar as possible and simultaneously minimizing the loss of the label predictor. DANN focuses on learning features that combine discriminative ness and domain invariance by optimizing the underlying features jointly.

$$\theta_f = \theta_f - \mu \left(\frac{dL_y}{d\theta_y} + (-1)(\lambda)\frac{dL_d}{d\theta_d} \right) \tag{1}$$

In Eqn 1, ⊖f is the gradient of the feature extraction layer, μ is the learning rate, λ is a hyper parameter for gradient reversal, Ly is label predictor loss, Ld is domain predictor loss, ⊖y is the parameters of label predictor, and Θd is the parameters of domain predictor. This equation represents the update of feature parameters during backpropagation using G.R.L.

V. SYSTEM DESIGN AND ARCHITECTURAL DETAILS

The data collected from different sources was passed through a preprocessing pipeline and augmented to make it model-ready. The parameters for the model are set to default settings. Text vectorization was carried out max_features = 20000, i.e., the maximum vocab size, max_len = 200, i.e., sequence length to pad the outputs to embedding_ dims = 200 This data was then vectorized using Glove embedding with Adam optimizer and binary cross entropy loss before feeding it to a feature extractor. The Feature Extractor (F.E.) block consisted of sequential Conv1d and Max Pool layers, followed by an LSTM and Dense layer. Finally, the procured vectors from the feature extractor block were parallelly passed into Label Predictor and Domain Classifier (D.C.) blocks. The Label Predictor (LP) was used to classify either of the labels, "True" or "False," using a Dense and a sigmoid layer. The domain classifier also consisted of a Dense and a sigmoid layer that predicted the input data's domain (source/target). During backpropagation, the gradient of the D.C. block would pass through a gradient reversal layer, as explained in section IV-A. Fig.2 depicts the flow diagram and steps in the proposed DANN-based approach. Fig 3. gives a visual representation of the model architecture explained here.

A. EXPLAINABLE POST HOC MODEL USING LIME

The different explainability methods for text classification in NLP include gradient-based saliency, integrated gradients IG., layer wise relevance propagation LRP. and perturbation methods, feature importance-based techniques such as Shap, and attention-based heat maps showing model attention on specific words at a particular instance. For the misinformation detection application, we have used the Local Interpretable Model agnostic Explanation method to have a peek inside the black box DANN model to derive the target label prediction explainability of the DANN model as LIME is a post hoc model agnostic and can be applied to any classifier without change in model intrinsic and provide instance level local explanations on different multiple source modalities such as

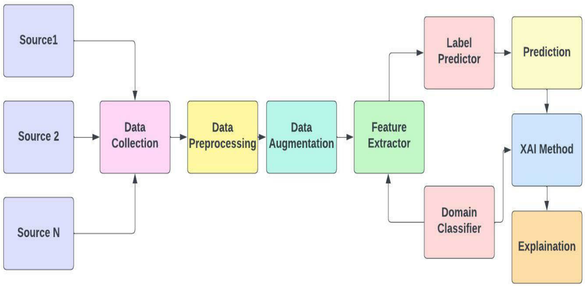


FIGURE 2. Flow diagram of the proposed DANN-based model.

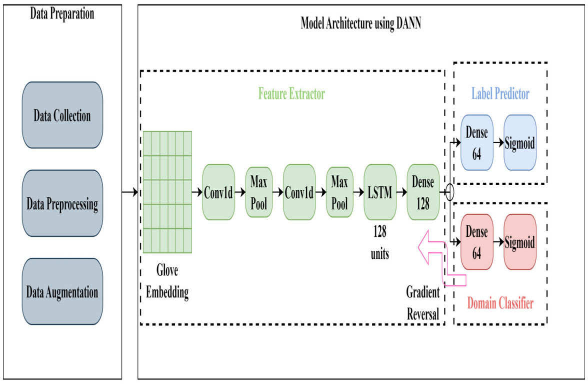


FIGURE 3. Model architecture for DANN.

image, text, and tabular data. Lime trains a linear model to approximate the local decision boundary of a particular instance. The technique can be applied to any classifier with text image and structured data providing a high degree of flexibility compared to the other methods giving it a significant weightage over others. The explanations are post hoc and are derived by locally approximating the underlying black box model by a linear interpretable model such as a random forest [76]. This generates sparse explanations that are short and human-understandable. LIME is trained on small perturbations of original instances as local approximations around the predictions by sampling and obtaining a surrogate dataset for the input instance whose decision is to be explained. The weighting of features is based on how close they are to the original instance, and the top significant features influencing the predictions are extracted. A random forest surrogate model with 500 trees is used for LIME implementation.

B. MODEL EVALUATION AND COMPARISON PROCEDURE

A comprehensive testing methodology was implemented to test the effectiveness of the Domain adaptation approach. The first method, where the source dataset training was done using a FE+LP model and target data, was used as a testing component only. This would be referred to as "Without DANN" in the paper. Method 2 (With DANN) consisted of using DANN (FE+(L.P., DC)) on the source and target data. Comparing "Without DANN' and "With DANN" approaches would help us demonstrate the approach's effectiveness over normal unsupervised techniques. News data comprised of the source domain and target domain would be the social media platforms like Twitter, Reddit, Instagram, etc. Metrics like Accuracy, Precision, Recall, and F1 score are used for the evaluation. The Area Under the Curve of the Receiver Operating Characteristic (AUC) is also used to understand how well the model ranks correct and incorrect pairs. AUC tells us mathematically how the true positives rate grows at various false-positive rates.

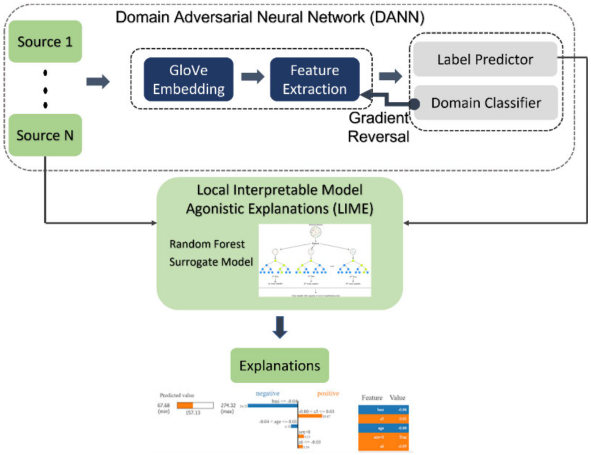

Domain Adversarial Neural Network (DANN)
Source 1
Source N
GloVe Embedding
Feature Extraction
Label Predictor
Domain Classifier
Gradient Reversal
Local Interpretable Model Agonistic Explanations (LIME)
Random Forest Surrogate Model
Explanations

FIGURE 4. An integrated framework consisting of DANN and LIME for Explainable Misinformation Detection.

VI. RESULT AND DISCUSSION

Table 5 shows the different metrics of both the approaches, i.e., Without DANN and With DANN, on the CoAID [71] dataset. To report this result, we have trained the model using "without DANN" and "with DANN." Without DANN, the model includes only the Feature Extractor and Label Predictor. Upon training both models, we then analyze the metrics on how the trained models perform compared to each other. It is observed that the DANN model performs better on the target domain (Twitter) when compared to the model trained on the source (News) and tested directly on the target (Twitter).

Table 5. The results obtained by [68] are compared to those obtained using the DANN architecture mentioned in section IV. Precision, Recall, and F1 are some comparison parameters [71]. DANN improves the Precision by 6% and the F1 score by 40% on target data. Here, DANN outperforms the previous approaches; hence, DANN can effectively learn features from one social media domain to another.

The Training, validation, and test set are made such that each set includes an equal number of samples from each social media platform. While training the model, the metrics (accuracy, loss) are calculated on the entire set rather than looking specifically at each social platform. After achieving the best model, the testing is done to calculate the metrics (Accuracy, A.U.C.) separately for each social platform (Twitter, Instagram, Reddit, Youtube). Refer to table 6 for more details. Our approach in table 6 was to better the previous iteration of models, which were trained on data combined on two domains (CoAID: News and Twitter). The option for training the model using the DANN approach is invalid because there is no target domain after merging the two

TABLE 5. Result of CoAID dataset using our method.

Domain	Without DANN			With DANN		
	Precision	Recall	F1	Precision	Recall	F1
News (Source)	0.8336	0.8264	0.8296	0.7760	0.7805	0.7750
Twitter (Target)	0.9382	0.5212	0.6497	0.9502	0.7285	0.8113

TABLE 6. Result on CoAID dataset using other methods without DANN.

Method	Precision	Recall	F1
H.A.N. [77]	0.6965	0.4659	0.5471
dEFEND [78]	0.8965	0.4847	0.5814
Our model (without DANN using F.E. + L.P.) on the total CoAID dataset	0.9701	0.9145	0.9300

TABLE 7. Result of MISOVAC dataset with and without Dann.

Domain	Without DANN		With DANN	
	Accuracy	A.U.C.	Accuracy	A.U.C.
News (Source)	0.6393	0.6872	0.6885	0.8149
Twitter (Target)	0.6224	0.7413	0.7608	0.8493
Instagram (Target)	0.6945	0.6127	0.6247	0.7019
Reddit (Target)	0.6034	0.5214	0.6027	0.5610
YouTube (Target)	0.6250	0.6250	0.6250	0.6250

domain's data. Hence, we used without DANN approach (FE+LP) for reporting metrics versus the previous state of art methods on this dataset.

Further, we present the results obtained on the MiSoVac data set in table 7. Accuracy/A.U.C. with DANN surpasses most social media platforms. In the MiSoVac dataset, the corpus for Twitter was balanced and substantial, so we observed a significant deviation for both the metrics in DANN and Without DANN. Using DANN, the accuracy of Twitter data increases by 22%, while A.U.C. increases by 15% compared to the normal approach Without DANN. While the data for misinformation on other platforms was limited, only a slight change was observed for the DANN approaches. Only Instagram shows a decrease in accuracy by 10%; however, the A.U.C. increases by 14% when DANN is applied, so the DANN model can distinguish the classes better even if the accuracy is low. An increase of 3% accuracy and 9% in the A.U.C. scores are observed. We observe that DANN can learn the domain invariant features and is good at generalizing data from various social media platforms.

Table 8 compares the H.A.N. [Yang, 2016] architecture on the MiSoVac dataset alongside the "Without DANN" architecture. The "Without DANN" method performs better than the H.A.N. [77] architecture. Among the previous approaches, dEFEND [78] utilized a combination of tweets and their replies, so implementing it on the MiSoVac dataset was impossible.

TABLE 8. Comparison of hann with our method without Dann on the misovac database.

Method	Precision	Recall	F1
H.A.N. [77] (on the entire dataset)	0.75	0.74	0.74
Our model (without DANN using F.E. + L.P.) on the total MiSoVAC dataset	0.77	0.77	0.76

Below is Fig.5 are plots that compare the accuracy and AUC of DANN and Without DANN implementation on the four target domains considered: Reddit, Instagram, Twitter, and YouTube.

Generalization of social media is done by combining multiple datasets into the source and target domains. The samples from the News and Instagram datasets are included in the source domain, and the target domain has samples from Twitter, Reddit, and YouTube datasets. Domain adaptive training using the DANN architecture mentioned in the previous section is carried out. These results are summarized in Table 9. When training the DANN model, the source accuracies (for combined news and Instagram) were 0.8, and the combined AUC was 0.9233. So, while training the DANN model, the model learned to understand complex features from data of various social media domains that increase the source results. As shown in Table 9, we see an improvement in accuracy and AUC for YouTube data for target results. There is a slight increase in the AUC of Twitter data while the accuracy is reduced. For Reddit, as the number of test samples were relatively less, we see little improvement in the results. The same was observed in the results mentioned in Table 8.

The adopted approach is economical as it generalizes to multiple data sources. It saves the time required to build and train individual models and detectors and the cost of annotating vast amounts of data. The overall time taken across the project pipeline is reduced as the step for annotating the newly scraped dataset is reduced. Moreover, individual training models are time-consuming when dealing with separate domains, which can be avoided using this domain adaptation approach. The model trained without the DANN approach trains only on the corpus data of that specific domain. Using the DANN approach, the target data samples are introduced while training; hence the DANN architecture learns the domain invariant features trying to adapt to both the source and target domain. Proving that annotation for the target domain data is not required as the model achieves high accuracy based on the labeled source data.

Explainable Model:

Explainability leads to disentanglement in the domain-specific features and improved generalization to the target domain without hindering performance on the source domain bridging the domain gap [79]. Domain shifts were in the data distribution of the source, and the target domain is

TABLE 9. Testing Results for the combined and individual target domain.

Media (Target Domain)	Without DANN		With DANN	
	Accuracy	A.U.C.	Accuracy	A.U.C.
Twitter	0.6945	0.6127	0.6562	0.7187
Reddit	0.6034	0.5214	0.5000	0.4800
YouTube	0.6250	0.6250	0.8999	0.8000
Combined (Twitter + Reddit + YouTube)	0.8846	0.9689	0.6730	0.6938

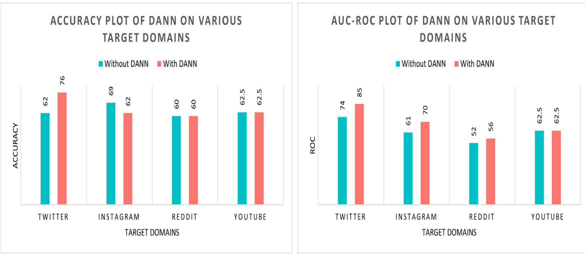


FIGURE 5. Accuracy and A.U.C. plot for various Social media platforms.

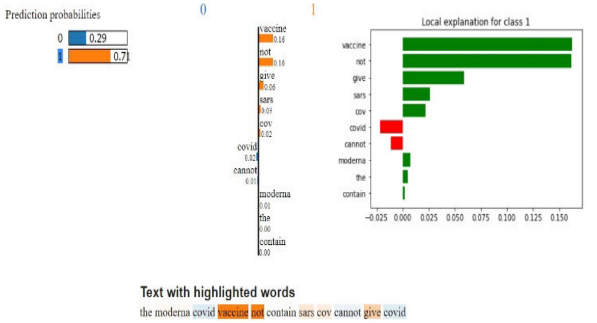


FIGURE 6. Example – 1 LIME explanations for prediction probabilities for both classes (0 and 1) based on the score assigned to each word in the sentence text and its corresponding highlight color.

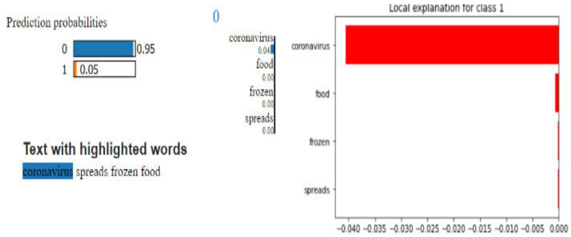


FIGURE 7. Example – 2 LIME explanations for prediction probabilities for both classes (0 and 1) based on the score assigned to each word in the sentence text and its corresponding highlight color. Only the class 0 word (coronavirus) is from the feature space.

different and can be addressed with XAI. Domain adaptation results in learning more discriminative features in the text classification results with the change in evidence in contrast to without domain adaptation, improving generalization on unseen domain learning domain invariant representation.

The LIME visualizations are intuitive and understandable. The results obtained are interpretable for humans. Local explanation reflects the local fidelity, i.e., the classifier's behavior for a particular data instance.

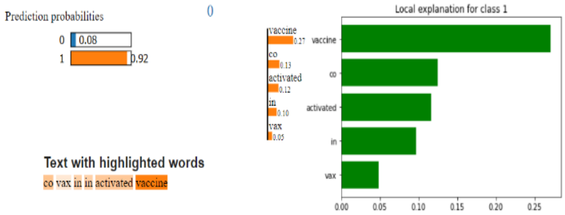


FIGURE 8. Example – 3 LIME explanations for prediction probabilities for both classes (0 and 1) based on the score assigned to each word in the sentence text and its corresponding highlight color. Only the class 0 word (coronavirus) is from the feature space.

Figures 6- 8 demonstrate the output of the LIME for three class instances for the dataset.

The blue represents class 0, and the orange represents class 1. The text is highlighted with the probability of each word being in either class. The bar chart on the left shows float point numbers on the horizontal bars representing the relative importance of these features (green for class 1 and red for class 0). LIME maintains the explanatory ability of significant features regardless of the chosen classifier running independently of the model used. The text explainer finds the top words which primarily drive the model to make the classification decision providing intuitive model behavior. This maps with the original class label providing individual feature relevance and high feature contribution in the final prediction by highlighting the text. For Fig 6, the text has features with a higher probability (∼70%) of being in class 1. In Fig. 7, a higher feature weight is given to the term vaccine with 92% class probability. In Fig. 8, the highest importance is assigned to the term coronavirus with the probability of 95% of class 0. LIME provides excellent results for text classifiers, but random sampling for data instances can be unstable in certain scenarios. It typically acts as an explanation tool for expert and non-expert users with diverse explainability requirements boosting trust for real-world adoption and deployment. The explanations are model-agnostic. The model can distinguish between actual and false sentences for misinformation detection.

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we demonstrated the explainable misinformation detection for social media platforms by employing the DANN and explainable AI LIME-based approach for explaining the target label predictions of the black box DAAN model making it more locally interpretable, trustworthy, and adaptable in real-world applications. Persistent propagation of misinformation in the healthcare domain leads to a direct and significant impact on human social well-being; hence, adopting explainable A.I. in establishing human trust is paramount in healthcare. DANN is employed for misinformation detection across multiple social media platforms. We consider the most relevant case study in current times, COVID-19 misinformation, to implement and test our approach. We use two specific data sets, CoAID, which is available openly and contains samples from news and Twitter sources. In addition, we developed a novel dataset named MiSoVac focusing on COVID-19 vaccine-related misinformation from various social media platforms. We described our data collection procedure, annotation, preprocessing techniques, and the architecture implemented to develop a generic classifier using DANN. Our methodology demonstrates promising results and outperforms other CoAID dataset's target domain approaches. An increase of ∼40% was attained in the F1 score compared to the best model mentioned in the preexisting work [50]. We illustrate the effectiveness of DANN architecture on the MiSoVac dataset and observe that DANN surpasses results obtained by the Without DANN approach by ∼3% in accuracy and ∼9% in AUC on average across all target domains. Domain adaptation and explainability for various social platforms still need to be explored extensively. This is the first of many steps towards developing techniques capable of generalizing on multiple data of a similar domain. Our approach could prove more economical regarding time and processing and generate significantly effective results without training the models for individual platforms powered with the joint prediction and explanation approach for establishing trust and adoption. We also hope the MiSoVac dataset is helpful to fellow researchers to help tackle the COVID-19 misinformation spread. The approach followed in this work is not limited to misinformation detection it can be further explored and extended for more varied tasks in natural language classification were the incoming data amalgamates from multiple domains, such as sentiment analysis, intent detection, and language modeling and detection.

REFERENCES

[1] A. Bobkov, S. Gafurov, V. Krasnoproshin, and H. Vissia, "An approach to web information processing," in *Pattern Recognition and Information Processing* (Communications in Computer and Information Science), vol. 673, V. Krasnoproshin and S. Ablameyko, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-54220-1_13.

[2] *Digital-Trends-Q4-Update @*. www.hootsuite.com. Accessed: Oct. 2022. [Online]. Available: https://www.hootsuite.com/resources/digital-trends-q4-update

[3] B. H. Lim, D. Lu, T. Chen, and M.-Y. Kan, "#mytweet via instagram: Exploring user behaviour across multiple social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 113–120.

[4] H. Daumé, III, "Frustratingly easy domain adaptation," 2009, *arXiv:0907.1815*.

[5] *Towards-Compositional-Understanding-of-the-World-by-Agent-Based-Deep-Learning @ Slideslive.Com*. Accessed: Dec. 14, 2021. [Online]. Available: https://slideslive.com/38922794/towards-compositional-understanding-of-the-world-by-agent-based-deep-learning

[6] P. Przybyła and A. J. Soto, "When classification accuracy is not enough: Explaining news credibility assessment," *Inf. Process. Manage.*, vol. 58, no. 5, Sep. 2021, Art. no. 102653, doi: 10.1016/j.ipm.2021.102653.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[8] M. Nemesh, "Information processing," in *Encyclopedia of Business and Finance*, 2nd ed. Encyclopedia.com, Feb. 2023. [Online]. Available: https://www.encyclopedia.com

[9] C. Wardle and H. Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*, vol. 27. Strasbourg, France: Council of Europe, 2017, pp. 1–107.

[10] A. Carson and K. Farhall, "Understanding collaborative investigative journalism in a 'post-truth' age," *Journalism Stud.*, vol. 19, no. 13, pp. 1899–1911, Oct. 2018, doi: 10.1080/1461670X.2018.1494515.

[11] M. Risdal, "Getting real about fake news," Kaggle, 2016, doi: 10.34740/KAGGLE/DSV/911.

[12] Walter Quattrociocchi. *World Economic Forum*. [Online]. Available: https://www.weforum.org/agenda/2016/01/q-a-walter-quattrociocchi-digital-wildfires/

[13] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1163–1168, doi: 10.18653/v1/n16-1138.

[14] *Signal-Dataset @ Research.Signal-Ai.Com*. Accessed: Oct. 2020. [Online]. Available: https://research.signal-ai.com/newsir16/signal-dataset.html

[15] S. O. Oyeyemi, E. Gabarron, and R. Wynn, "Ebola, Twitter, and misinformation: A dangerous combination?" *Brit. Med. J.*, vol. 349, Oct. 2014, Art. no. g6178, doi: 10.1136/bmj.g6178.

[16] J. P. D. Guidry, K. Carlyle, M. Messner, and Y. Jin, "On pins and needles: How vaccines are portrayed on pinterest," *Vaccine*, vol. 33, no. 39, pp. 5051–5056, Sep. 2015, doi: 10.1016/j.vaccine.2015.08.064.

[17] A. Venkatraman, D. Mukhija, N. Kumar, and S. J. S. Nagpal, "Zika virus misinformation on the Internet," *Travel Med. Infectious Disease*, vol. 14, no. 4, pp. 421–422, Jul. 2016, doi: 10.1016/j.tmaid.2016.05.018.

[18] R. A. Kinchla and R. C. Atkinson, "The effect of false-information feedback upon psychophysical judgments," *Psychonomic Sci.*, vol. 1, nos. 1–12, pp. 317–318, Jan. 1964, doi: 10.3758/bf03342931.

[19] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Anatomy of an online misinformation network," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0196087, doi: 10.1371/journal.pone.0196087.

[20] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, "Persistence-based clustering in Riemannian manifolds," *J. ACM*, vol. 60, no. 6, pp. 1–38, Nov. 2013, doi: 10.1145/2535927.

[21] O. Cordon, O. Cordon, M. J. del Jesus, and F. Herrera, "Analyzing the reasoning mechanisms in fuzzy rule-based classification systems," *Mathware Soft Comput.*, vol. 5, nos. 2–3, pp. 321–332, 1998.

[22] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol.*, vol. 3, no. 11, pp. 1–22, Oct. 2002, doi: 10.1186/gb-2002-3-11-research0059.

[23] T. P. Kelsey and K. K. Saluja, "Fast test generation for sequential circuits," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Nov. 1989, pp. 354–357, doi: 10.1109/iccad.1989.76889.

[24] D. Nauck, F. Klawonn, and R. Kruse, *Foundations of Neuro-Fuzzy Systems*. Hoboken, NJ, USA: Wiley, 1997.

[25] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, pp. 1–9, Oct. 2022, doi: 10.3389/fncom.2022.1019776.

[26] C. Baydogan and B. Alatas, "Sentiment analysis in social networks using social spider optimization algorithm," *Tehnički Vjesnik*, vol. 3651, pp. 1943–1951, Nov. 1943.

[27] C. Baydogan and B. Alatas, "Deep-Cov19-hate?: A textual-based novel approach for automatic detection of hate speech in online social networks throughout COVID-19 with shallow and deep learning models," *Tehnički Vjesnik*, vol. 3651, pp. 149–156, 1848.

[28] A. Arnold, R. Nallapati, and W. W. Cohen, "A comparative study of methods for transductive transfer learning," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Oct. 2007, pp. 77–82.

[29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Jan. 2021, doi: 10.1109/TKDE.2009.191.

[30] T. Tommasi and B. Caputo, "Frustratingly easy NBNN domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 256–263.

[31] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—A survey," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6838–6855.

[32] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—A survey," 2020, *arXiv:2006.00632*.

[33] S. Petrov, P.-C. Chang, M. Ringgaard, H. Alshawi, and G. Research, *Uptraining for Accurate Deterministic Question Parsing*. Cedarville, OH, USA: Association for Computational Linguistics, 2010.

[34] G. Rotman and R. Reichart, "Deep contextualized self-training for low resource dependency parsing," 2019, *arXiv:1911.04286*.

[35] J. Yu, M. Elkaref, and B. Bohnet, "Domain adaptation for dependency parsing via self-training," in *Proc. 14th Int. Conf. Parsing Technol.*, 2015, pp. 1–10, doi: 10.18653/v1/w15-2201.

[36] R. C. Moore and W. Lewis, *Intelligent Selection of Language Model Training Data*. Cedarville, OH, USA: Association for Computational Linguistics, 2010.

[37] R. Aharoni and Y. Goldberg, ''Unsupervised domain clusters in pretrained language models,'' 2020, *arXiv:2004.02105*.
[38] H. Daumé III, ''Frustratingly easy domain adaptation,'' Tech. Rep., 2007.
[39] Y. Ziser and R. Reichart, ''Pivot based language modeling for improved neural domain adaptation,'' in *Proc. Conf. North Amer. Chapter Assoc. for Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 1241–1251, doi: 10.18653/v1/n18-1112.
[40] Y. Ziser and R. Reichart, ''Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 238–249, doi: 10.18653/v1/d18-1022.
[41] Y. Ziser and R. Reichart, ''Neural structural correspondence learning for domain adaptation,'' in *Proc. 21st Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2017, pp. 400–410, doi: 10.18653/v1/k17-1040.
[42] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, ''End-to-end adversarial memory network for cross-domain sentiment classification,'' in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2237–2243, doi: 10.24963/ijcai.2017/311.
[43] E. Ben-David, C. Rabinovitz, and R. Reichart, ''PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models,'' 2020, *arXiv:2006.09075*.
[44] X. Glorot, A. Bordes, and Y. Bengio, ''Domain adaptation for large-scale sentiment classification: A deep learning approach,'' in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 513–520.
[45] M. Chen, K. Q. Weinberger, Z. Xu, and F. Sha, ''Marginalizing stacked linear denoising autoencoders,'' *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 3849–3875, 2015.
[46] S. Clinchant, G. Csurka, and B. Chidlovskii, ''A domain adaptation regularization for denoising autoencoders,'' in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2016, pp. 26–31, doi: 10.18653/v1/p16-2005.
[47] S. Yang, K. Yu, F. Cao, H. Wang, and X. Wu, ''Dual-representation-based autoencoder for domain adaptation,'' *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7464–7477, Aug. 2022, doi: 10.1109/TCYB.2020.3040763.
[48] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, ''Domain separation networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Aug. 2016, pp. 343–351.
[49] Y. Ganin et al., ''Domain-adversarial training of neural networks,'' *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[50] J. Shen, Y. Qu, W. Zhang, and Y. Yu, ''Wasserstein distance guided representation learning for domain adaptation,'' 2017, *arXiv:1707.01217*.
[51] J. Yuan, Y. Zhao, and B. Qin, ''Learning to share by masking the non-shared for multi-domain sentiment classification,'' *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 9, pp. 2711–2724, 2022.
[52] Y. Li, T. Baldwin, and T. Cohn, ''What's in a domain? Learning domain-robust text representations using adversarial training,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., (Short Papers)*, vol. 2, 2018, pp. 474–479, doi: 10.18653/v1/n18-2076.
[53] D. Shah, T. Lei, A. Moschitti, S. Romeo, and P. Nakov, ''Adversarial domain adaptation for duplicate question detection,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1–8.
[54] M. Ryu, G. Lee, and K. Lee, ''Knowledge distillation for bert unsupervised domain adaptation,'' *Knowl. Inf. Syst.*, vol. 64, no. 11, pp. 3113–3128, 2022.
[55] G. Hinton, O. Vinyals, and J. Dean, ''Distilling the knowledge in a neural network,'' 2015, *arXiv:1503.02531*.
[56] K. Lim, J. Y. Lee, J. Carbonell, and T. Poibeau, ''Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, Apr. 2020, pp. 8344–8351, doi: 10.1609/aaai.v34i05.6351.
[57] F. Alam, S. Joty, and M. Imran, ''Domain adaptation with adversarial training and graph embeddings,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 1077–1087, doi: 10.18653/v1/p18-1099.
[58] X. Cui and D. Bollegala, ''Self-adaptation for unsupervised domain adaptation,'' in *Proc. Int. Conf. Recent Adv. Natural Lang. Process., (RANLP)*, Sep. 2019, pp. 1–10, doi: 10.26615/978-954-452-056-4_025.
[59] A. Adadi and M. Berrada, ''Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),'' *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[60] P. Linardatos and V. Papastefanopoulos, ''Explainable A.I.?: A review of machine learning interpretability methods,'' *Entropy*, vol. 23, no. 1, p. 18, 2021.
[61] G. Ras, N. Xie, M. van Gerven, and D. Doran, ''Explainable deep learning: A field guide for the uninitiated,'' 2020, *arXiv:2004.14545*.
[62] S. Gholizadeh, ''Model explainability in deep learning based natural language processing,'' 2021, *arXiv:2106.07410*.
[63] D. Gunning and D. W. Aha, ''DARPA's explainable artificial intelligence program,'' *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019.
[64] G. Vilone and L. Longo, ''Explainable artificial intelligence: A systematic review,'' 2020, *arXiv:2006.00093*.
[65] O. Biran and C. Cotton, ''Explanation and justification in machine learning: A survey,'' in *Proc. IJCAI Workshop Explainable A.I. (X.A.I.)*, vol. 8, no. 1, 2017, pp. 8–13.
[66] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, ''Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,'' *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[67] C. Molnar, *Interpretable Machine Learning*. Morrisville, NC, USA: Lulu.Com, 2020.
[68] G. Joshi, R. Walambe, and K. Kotecha, ''A review on explainability in multimodal deep neural nets,'' *IEEE Access*, vol. 9, pp. 59800–59821, 2021, doi: 10.1109/ACCESS.2021.3070212.A.
[69] S. Mohseni, N. Zarei, and E. D. Ragan, ''A survey of evaluation methods and measures for interpretable machine learning,'' 2018, *arXiv:1811.11839*.
[70] M. Ribeiro, S. Singh, and C. Guestrin, '''Why should I trust you?': Explaining the predictions of any classifier,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Demonstrations*, 2016, pp. 97–101, doi: 10.18653/v1/n16-3020.
[71] L. Cui and D. Lee, ''CoAID: COVID-19 healthcare misinformation dataset,'' 2020, *arXiv:2006.00885*.
[72] A. Srivastava, M. Hasan, B. Yagnik, R. Walambe, and K. Kotecha, ''Role of artificial intelligence in detection of hateful speech for Hinglish data on social media,'' in *Proc. Appl. Artif. Intell. Mach. Learn.: Select (ICAAAIML)*. Singapore: Springer, 2021, pp. 83–95.
[73] E. Ma, ''NLP augmentation,'' NLPAug, 2019. [Online]. Available: https://github.com/makcedward/nlpaug
[74] J. Pennington, R. Socher, and C. Manning, ''Glove: Global vectors for word representation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/d14-1162.
[75] S. Hochreiter and J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[76] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[77] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, ''Hierarchical attention networks for document classification,'' in *Proc. NAACL*, 2016, pp. 1480–1489, doi: 10.18653/V1/N16-1174.
[78] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, ''dEFEND: Explainable fake news detection,'' in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 395–405, doi: 10.1145/3292500.3330935.
[79] A. Zunino, ''Explainable deep classification models for domain generalization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2021, pp. 3233–3242.

GARGI JOSHI received the B.Tech. degree in computer science and engineering from Dr. Babasaheb Ambedkar Marathwada University (BAMU), Aurangabad, in 2012, and the M.E. degree from the University of Pune, in 2014. She is currently pursuing the Ph.D. degree in computer engineering in AI and deep learning specific to multimodal AI explainable AI domain with Symbiosis International (Deemed University), Pune. She is also a Junior Research Fellow with Symbiosis International (Deemed University). Her research interests include AI, machine learning, deep learning, and the advents of multimodal AI and XAI.

ANANYA SRIVASTAVA is currently pursuing the B.Tech. degree with the Department of Electronics and Telecommunication, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune.

BHARGAV YAGNIK is currently pursuing the B.Tech. degree with the Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune.

MOHAMMED HASAN is currently pursuing the B.Tech. degree with the Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune.

ZAINUDDIN SAIYED is currently pursuing the B.Tech. degree with the Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune.

LUBNA A. GABRALLA received the B.Sc. and M.Sc. degrees in computer science from the University of Khartoum and the Ph.D. degree in computer science from the Sudan University of Science and Technology, Khartoum, Sudan. She is currently an Associate Professor with the Department of Computer Science and Information Technology, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her current research interests include soft computing, machine learning, and deep learning. She became a Senior Fellow of HEA (SFHEA), in 2021.

AJITH ABRAHAM (Senior Member, IEEE) received the Master of Science degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a not-for-profit scientific network for innovation and research excellence connecting industry and academia (the network with HQ in Seattle, USA, has more than 1,500 scientific members from over 105 countries). As an investigator/a co-investigator, he has won research grants worth over more than U.S. $100 million. He also holds two university professorial appointments. He works as a Professor in artificial intelligence with Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair of Artificial Intelligence with UCSI University, Malaysia. He works in a multidisciplinary environment. He has authored/coauthored more than 1,400 research publications, out of which are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as per Google Scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), from 2008 to 2021, and served as a Distinguished Lecturer for IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and serves/served on the editorial board of over 15 international journals indexed by Thomson ISI.

RAHEE WALAMBE received the M.Phil. and Ph.D. degrees from Lancaster University, U.K., in 2008. From 2008 to 2017, she was a research consultant with various organizations in the control and robotics domain. She is currently associated with the Symbiosis Centre for Applied Artificial Intelligence, SIU. Her research interests include applied deep learning and AI in robotics and healthcare. She is a recipient of multiple research grants.

KETAN KOTECHA was an Administrator with Parul University and Nirma University and has several achievements in these roles to his credit. He is currently a Team Member of the nationwide initiative on AI and deep learning skilling and research named Leadingindia.ai initiative sponsored by the Royal Academy of Engineering, U.K., under the Newton Bhabha Fund. He heads the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). He is also considered a foremost expert in AI and aligned technologies. In addition, he has pioneered education technology with his vast and varied experience in administrative roles. He has expertise and experience in cutting-edge research and projects in AI and deep learning for more than 25 years. He has widely published in several excellent peer-reviewed journals, including education policies, teaching-learning practices, and AI.