Received: 8 March 2022 | Accepted: 19 January 2023

DOI: 10.1111/poms.13959

ORIGINAL ARTICLE

Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach

Ka Chung Ng[1] | **Ping Fan Ke**[2] | **Mike K. P. So**[3] | **Kar Yan Tam**[3]

[1]Department of Management and Marketing, Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

[2]School of Computing and Information Systems, Singapore Management University, Singapore, Singapore

[3]Department of Information Systems, Business Statistics and Operations Management, School of Business and Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Correspondence
Ping Fan Ke, School of Computing and Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902.
Email: pfke@smu.edu.sg

Handling Editor: Vijay Mookerjee

Abstract

Online platforms are experimenting with interventions such as content screening to moderate the effects of fake, biased, and incensing content. Yet, online platforms face an operational challenge in implementing machine learning algorithms for managing online content due to the labeling problem, where labeled data used for model training are limited and costly to obtain. To address this issue, we propose a domain adaptive transfer learning via adversarial training approach to augment fake content detection with collective human intelligence. We first start with a source domain dataset containing deceptive and trustworthy general news constructed from a large collection of labeled news sources based on human judgments and opinions. We then extract discriminating linguistic features commonly found in source domain news using advanced deep learning models. We transfer these features associated with the source domain to augment fake content detection in three target domains: political news, financial news, and online reviews. We show that domain invariant linguistic features learned from a source domain with abundant labeled examples can effectively improve fake content detection in a target domain with very few or highly unbalanced labeled data. We further show that these linguistic features offer the most value when the level of transferability between source and target domains is relatively high. Our study sheds light on the platform operation in managing online content and resources when applying machine learning for fake content detection. We also outline a modular architecture that can be adopted in developing content screening tools in a wide spectrum of fields.

KEYWORDS
adversarial domain adaptation, augmented AI, deception detection, fake news, transfer learning

1 | INTRODUCTION

Increasingly, online platforms like social media and review websites are applying machine learning techniques to screen user-generated content for quality control (Lee et al., 2018; X. Zhang et al., 2022). This is in response to surging societal concerns about the spread of fake content that fuel extreme emotions (Allcott & Gentzkow, 2017; Ng et al., 2021; Vosoughi et al., 2018). In 2016, Facebook began to tag articles that were identified as fake by third-party fact-checkers.[1] Similarly, in early 2020, Twitter started to flag what it judged to be inappropriate content related to COVID-19.[2] Without proper quality control, the value of the platforms will be reduced (Choi et al., 2018; Cui et al., 2018; Wei et al., 2021; Yan & Pedraza-Martinez, 2019). Hence, content screening will be one of the most crucial processes of the backend operations for platforms, especially those specializing in content provision.

In quality management, the value loss in a value chain is measured as the deviation between the expected and the actual product characteristics (Taguchi, 1985), with the expense of quality control processes such as costs of sampling and inspection (Kanyamibwa & Ord, 2000). In a traditional supply chain with physical goods, such value loss could be reduced by contractual agreement with the upstream business partners, like deterred payment (Rui & Lai, 2015) and price negotiation based on defective rate (Leng et al., 2016). However, online platforms typically have a much larger upstream (i.e., users who generate content) with anonymous identity, making contractual agreement practically infeasible.

Accepted by Vijay Mookerjee, after three revisions.

As a result, interventions to curb the propagation of fake content will be necessary for platforms. To efficiently routinize such inventions, real-time assessment built on computational models will be required. The capability of machine learning in fake content detection has been reported in the literature, in which fake content has been found to contain linguistic cues that reveal its dissociation from genuine content (Clarke et al., 2021; D. Zhang et al., 2016; X. Zhang et al., 2022; L. Zhou & Zhang, 2008), and these cues could be captured by machine learning and natural language processing (NLP) techniques (Bloomfield, 2012; Clarke et al., 2021; Sharma et al., 2019; Q. Wang et al., 2018).

Despite the effectiveness of machine learning in fake content detection, online platforms always suffer from a lack of accurately labeled data to train machine learning models for content screening as these data are limited and costly to obtain (Ganin et al., 2016; Van Vlasselaer et al., 2017; Zhu et al., 2020). As summarized in Supporting Information Appendix A, our literature review reveals two challenges in developing machine learning models for fake content detection: (1) constructing a large sample of labeled examples for discriminant analysis is difficult because many domains have few confirmed cases of fake content and (2) collecting labeled fake content is laborious and costly. These challenges raise important operational issues for online platforms in utilizing machine learning for managing and exploiting online content (Wei et al., 2021). To address this issue, this study adopts an augmented artificial intelligence (AI) perspective, defined as a type of human–AI hybrid where humans and AI augment one and another (A. Rai et al., 2019), to advance fake content detection. Specifically, we augment fake content detection by identifying discriminating and domain invariant linguistic features based on a large collection of human judgments and perceptions of truth and deception, representing an approach that incorporates collective human intelligence to enhance AI (Yau et al., 2021).

In a nutshell, the idea is to (1) identify a news domain (source domain) where there is an abundance of human judgments and opinions in fake news detection, (2) distill the common linguistic features of news that are effective in discriminating against fake and non-fake news based on the collective human inputs in the source domain, (3) transfer these linguistic features to another domain (target domain) where verified fake cases and human inputs are scare, and (4) augment these transferred features with traditional AI techniques to alleviate the problems associated with limited/unbalanced verified cases for model development in the target domain. In this regard, our work connects to the nascent literature on human–AI interaction (Fügener et al., 2021; Ge et al., 2021; A. Rai et al., 2019; Raisch & Krakowski, 2021) by contributing a unique form of human-in-the-loop case that leverages collective human intelligence to augment fake content detection as part of a regular platform operation.

To rigorously validate the proposed approach, we identify three target domains: political news, financial news, and online reviews. All three domains have insufficient accurately labeled samples but to differing degrees. Fake political news is defined as "news articles that are intentionally and verifiably false and could mislead readers" (Allcott & Gentzkow, 2017, p. 213). It has received much attention since the 2016 U.S. presidential election, and a body of fake political news data has been manually assembled. Since fake or manipulated political news can undermine social media's credibility and users' experience, flagging this content is important to platform operations (Lee et al., 2018). On the other hand, fake financial news articles are written with malicious intentions to manipulate the financial market (Clarke et al., 2021; X. Zhang et al., 2022). Unlike political news, wrongly labeled financial news could have significant legal consequences. Besides, financial news generally contains domain knowledge and insider information, causing ground-truth labeling to be hard to verify (X. Zhang et al., 2022). These explain why financial news verified by the respective regulator as fake are very few (Clarke et al., 2021; X. Zhang et al., 2022). As firms increasingly incorporate various online information into their operations (Wei et al., 2021), identifying fake financial news provides valuable insights into the operational risk in financial services (Xu et al., 2017). Lastly, fake online reviews are defined as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed" (D. Zhang et al., 2016, p. 457). Fake online reviews can significantly influence platform operations by shaping consumers' purchasing decisions and affecting merchants' revenue (Wu et al., 2020). They also serve as an appropriate context in the current study as their ground truth is difficult to establish, and the usual practice of manually labeling fake reviews is inefficient (D. Zhang et al., 2016).

To represent human intelligence, we consider linguistic features associated with deception and truth extracted from a large general news dataset constructed based on the consensus of human labelers. We regard these linguistic features as opinion based because they reveal the difference between deceptive and trustworthy news according to human heuristic judgment. Specifically, we collect three sets of deceptive news (fake, biased, and clickbait news) from a comprehensive list of labeled deceptive news sites (Zimdars, 2016) and combine these with trustworthy news collected from a list of reliable news sites maintained by *Wikipedia*. In total, we obtain more than 2.2 million deceptive news articles and more than 1.9 million trustworthy news articles to serve as source domain data for extracting opinion-based linguistic features. In a way, labeling these 4 M+ articles corresponds to the collective wisdom of human labelers in judging whether a piece of news is fake or not from a linguistic standpoint.

To augment fake content detection with human intelligence, we propose an advanced model that combines deep learning, transfer learning, domain adaptation, and adversarial training within a single framework. Under the framework, we first apply deep learning to extract discriminating linguistic features from the source domain. We then use transfer learning with domain adaptation via adversarial training to transfer domain invariant linguistic features to the three target

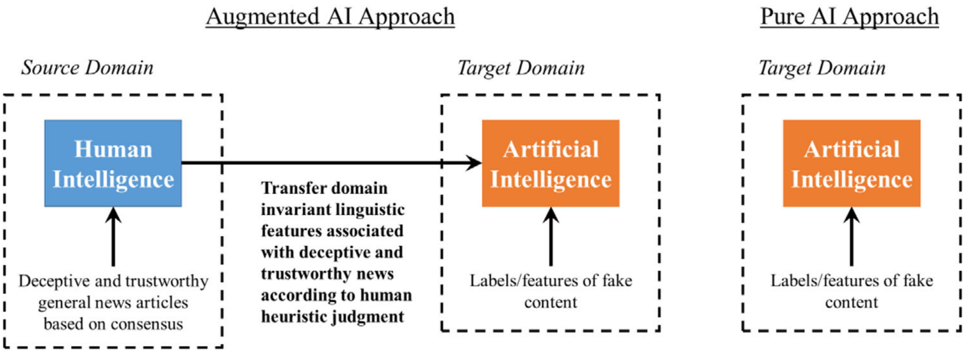
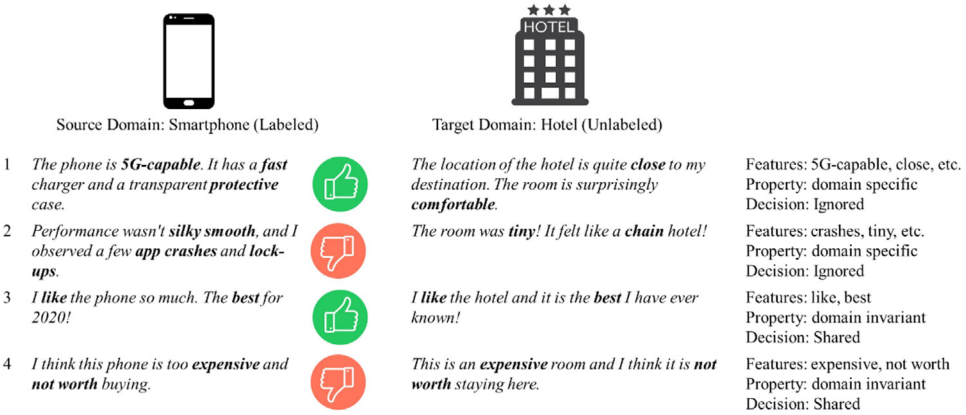

Augmented AI Approach
Pure AI Approach
Source Domain
Target Domain
Target Domain
Human Intelligence
Deceptive and trustworthy general news articles based on consensus
Transfer domain invariant linguistic features associated with deceptive and trustworthy news according to human heuristic judgment
Artificial Intelligence
Labels/features of fake content
Artificial Intelligence
Labels/features of fake content

FIGURE 1 Augmented AI approach versus pure AI approach. [Color figure can be viewed at wileyonlinelibrary.com]



	Source Domain: Smartphone (Labeled)		Target Domain: Hotel (Unlabeled)	
1	*The phone is* ***5G-capable****. It has a* ***fast*** *charger and a transparent* ***protective*** *case.*	👍	*The location of the hotel is quite* ***close*** *to my destination. The room is surprisingly* ***comfortable****.*	Features: 5G-capable, close, etc. Property: domain specific Decision: Ignored
2	*Performance wasn't* ***silky smooth****, and I observed a few* ***app crashes*** *and* ***lock-ups****.*	👎	*The room was* ***tiny****! It felt like a* ***chain*** *hotel!*	Features: crashes, tiny, etc. Property: domain specific Decision: Ignored
3	*I* ***like*** *the phone so much. The* ***best*** *for 2020!*	👍	*I* ***like*** *the hotel and it is the* ***best*** *I have ever known!*	Features: like, best Property: domain invariant Decision: Shared
4	*I think this phone is too* ***expensive*** *and* ***not worth*** *buying.*	👎	*This is an* ***expensive*** *room and I think it is* ***not worth*** *staying here.*	Features: expensive, not worth Property: domain invariant Decision: Shared

FIGURE 2 Illustration of a bag-of-words idea of domain adaptation. [Color figure can be viewed at wileyonlinelibrary.com]

domains (political news, financial news, and online reviews). Lastly, we fine-tune the domain adaptive transfer learning with labeled target domain data to allow for AI adjustment in reducing human biases and errors. We refer to this model framework as the "augmented AI" approach. For comparison, we consider three machine learning models (multi-layer perceptron, random forest, and multinomial naïve Bayes) directly trained on target domain data for fake content detection, which we regard as the "pure AI" approach. Figure 1 gives an overview of the two approaches.

The domain adaptive transfer learning via adversarial training is the state-of-the-art model (Chen et al., 2022; Y. Shi et al., 2022) that aims to extract discriminative and domain invariant features from one domain (source domain) and then transfer these features to another domain (target domain) to solve the same or different tasks (Pan & Yang, 2010; Zhuang et al., 2021). Transfer learning has been used in machine learning applications in technical and engineering domains where insufficient or no labeled data are available (Zhuang et al., 2021). The resulting trained models have generally performed well, requiring fewer data and less computational time than traditional models (Shin et al., 2020).

We illustrate the concept of domain adaptation in its simplest form using a sentiment classification task that comprises two review datasets: labeled reviews of smartphones and unlabeled reviews of hotels (see Figure 2). A typical machine learning model will discover several features that differentiate between positive and negative reviews of smartphones, such as "5G-capable" and "fast," as shown by examples 1 and 2 in Figure 2. However, these features are specific to smartphones (i.e., domain specific features) and are thus not useful to differentiate between positive and negative reviews of hotels. Conversely, features such as "best," "expensive," and "not worth," as shown by examples 3 and 4, are applicable to both smartphones and hotels and can be used to differentiate between positive and negative reviews of smartphones and hotels. The goal of domain adaptation is to extract features from a source domain containing sufficient labeled examples and then transfer only discriminative and domain invariant features to a target domain. This transfer process can be achieved within a single learning design based on adversarial training that trains models in a competing way (Ganin et al., 2016). This approach requires null or only a very small percentage of labeled examples to be initially present in the target domain. Figure 2 illustrates a simple bag-of-words idea of domain adaptation, while more complex feature forms (e.g., latent features) are difficult to visualize but remain relevant and applicable.

In general, we expect the AI-based fake content detection with augmentation to outperform the pure AI approach. Our empirical results confirm this expectation. To further explore the boundary condition of the augmented AI model, we examine how the performance of domain adaptive transfer learning varies according to the level of domain transferability. Specifically, domain adaptive transfer learning should be less effective in situations where the transferability between the source and target domains is low, as these domains will share few common features. As transferability increases, performance increases correspondingly until it reaches a point where incremental improvement starts to plateau. The reason is that when the source and target domains are highly similar, many features learned from the source domain are directly applicable to the target domain, and thus the benefit of transfer learning becomes marginal. Accordingly, we construct a simple and generalizable score to quantify domain transferability and show that the augmented AI approach's performance varies according to domain transferability.

In terms of operation support, the domain transferability score provides an implication for online platforms to allocate resources efficiently when applying machine learning for fake content detection. For instance, consider two product categories, with one containing limited labeled fake reviews as the target domain and another containing a sufficiently large amount of labeled fake reviews that can serve as the source domain. Using the transferability score as a guiding indicator, if domain transferability is above medium, the platform can consider applying domain adaptive transfer learning directly to detect fake content in the target domain without spending effort and time on manual labeling. However, if the domain transferability is very low, the platform needs to collect extra labels for training machine learning models. In this way, the transferability score guides online platforms through identifying domains that require extra resources (i.e., labels) when using machine learning to detect fake content.

This study makes several contributions to emerging research in operations management (OM) and information system (IS) interface, including AI, deep learning, social media, and digital platforms (S. Kumar et al., 2018). First, we propose an augmented AI approach that operationalizes collective human intelligence in assisting online platforms in identifying fake content as a routine practice. Our approach can also help online platforms resolve the inefficiency of labeling problem when applying machine learning models. Second, the idea of domain transferability implies a data-driven solution to help online platforms effectively allocate resources for augmenting fake content detection in different domains or categories. Lastly, by recognizing an increasing number of research studies that employ machine learning and big data analytics in solving OM-related problems (Choi et al., 2018; Cui et al., 2018; S. Kumar et al., 2018; Lee et al., 2018), our work contributes an early effort in applying advanced deep learning to improve online content management. To the best of our understanding, domain adaptive transfer learning via adversarial training has not been used in fake content detection for supporting platform operations (see Supporting Information Appendix A).

2 | LITERATURE REVIEW

2.1 | Fake content detection based on linguistic features and machine learning

There is a body of research in deception theory and computational linguistics demonstrating that deceptive content contains distinct linguistic features that can be used for its detection (Ho et al., 2016; Rashkin et al., 2017; Q. Wang et al., 2018; Zahedi et al., 2015). L. Zhou et al. (2004) showed that linguistic constructs are useful for detecting deception in text-based asynchronous computer-mediated communication. Larrimore et al. (2011) found that loan descriptions with extended narratives and concrete descriptions increased lenders' perceptions of borrowers' trustworthiness. Bloomfield (2012) leveraged linguistic features to identify deceptive messages relayed during quarterly earnings conference calls. Toma and D'Angelo (2015) showed that online medical advice was perceived as more trustworthy if it contained more words, especially long words, and fewer "I"-pronouns and anxiety-related words. Ho et al. (2016) identified that word counts and words that were associated with cognitive and affective processes were important factors for detecting deception in online communication. Rubin et al. (2016) highlighted that fake articles typically contained more humorous, ironic, and absurd words. Siering et al. (2016) studied fraudulent behavior in crowdfunding platforms and found that the content-based and linguistic cues of suspended projects differed from those of non-suspended projects. Rashkin et al. (2017) determined that compared with trustworthy news, fake news contained more first- and second-person pronouns; more subjective, superlative, and modal adverbs; fewer assertive and "hear" category words (Tausczik & Pennebaker, 2010); and fewer hedging words. Furthermore, Yang et al. (2017) suggested that satirical political news was more emotional and unprofessional than trustworthy news and Clarke et al. (2021) reported that there were substantial differences between the linguistic features of fake news and those of legitimate financial news.

Previous studies have also attempted to use NLP and machine learning techniques to capture the linguistic features of fake content. For example, Abbasi et al. (2010) developed a design science framework based on a support vector machine to identify textual cues embedded in a web page, such as word phrases and grammar, and thus identify fake websites. Q. Wang et al. (2018) studied deception in the mobile app market and developed a machine learning model that used app descriptions and reviews to identify copycat apps. Several studies on the analysis of online reviews have demonstrated the utility of combining textual cues with machine learning models to detect fake or inauthentic reviews (N. Kumar et al., 2019; Ott et al., 2011; D. Zhang et al., 2016). Many studies in the political sphere have analyzed the language patterns associated with deception and extracted linguistic features from fake political news (Rashkin et al., 2017; Rubin et al., 2016; W. Y. Wang, 2017; Yang et al., 2017). These patterns and linguistic features have been used to develop advanced deep learning models to detect and assess the truthfulness of news (Sharma et al., 2019; X. Zhou & Zafarani, 2020). In the finance and accounting context, Clarke et al. (2021) used a psycholinguistic and word categories lexicon (Tausczik & Pennebaker, 2010) to develop machine learning models to detect deceptive financial news articles. Bloomfield (2012) used the same lexicon and logistic regression to detect deceptive conference calls.

The above studies show that combining linguistic features with machine learning for fake content detection is a burgeoning and promising field of research. However, previous studies on deceptive linguistic cues generally relied on small samples of deceptive data and on the direct application of standalone machine learning models (see Supporting Information Appendix A) that are susceptible to the inefficiency of labeling problem. To address this research gap, the current study aims to advance fake content detection in online platforms by augmenting domain invariant linguistic features representing collective human intelligence with the help of domain adaptive transfer learning via adversarial training.

2.2 | Transfer learning and its application

Transfer learning has received increasing research attention in recent years and has been proven useful in various applications, such as pattern recognition and sentiment analysis (Zhuang et al., 2021). It involves using knowledge acquired in one task to solve a related task in the same or similar domains. According to a survey by Zhuang et al. (2021), there are several ways to categorize transfer learning based on different criteria. Given that our paper aims to address the labeling issue in fake content detection in online platforms, in the following, we adopt the label-setting perspective of Pan and Yang (2010) and Zhuang et al. (2021) to categorize transfer learning into three types: inductive transfer learning, transductive transfer learning, and unsupervised transfer learning.

Inductive transfer learning is used when the source and target domains are the same (e.g., finance), and the source and target tasks are different but related (e.g., opinion mining vs. sentiment analysis). Accordingly, inductive transfer learning requires labeled data in the target domain but either labeled or unlabeled data in the source domain. For example, Zheng et al. (2020) observed that the distribution of credit card transactions changed with users' transaction behavior over time, thus developed a boosting-based transfer learning model to improve credit scoring. Peng (2020) proposed an inductive transfer learning model to improve public firms' earnings forecasts within dynamic data environments. Kratzwald and Feuerriegel (2019) pioneered a transfer learning approach to transfer useful knowledge from other NLP tasks to improve the performance of question–answering systems.

Transductive transfer learning is used when the source and target tasks are the same (e.g., deception detection), while the source and target domains are different (e.g., political vs. finance). Unlike inductive transfer learning, transductive learning requires substantial labeled data in the source domain but few or no labeled data in the target domain. When a source and a target domain are very similar and highly related, a pre-trained model based on the source domain data can be directly applied to solve the target domain task. For instance, Kraus and Feuerriegel (2017) used a deep learning model pre-trained on Form 8-K filings to analyze regulated ad hoc announcements for stock price prediction. In contrast, when a source and a target domain are different, domain adaptation is required to ensure that only domain invariant features are transferred to the target domain (Ganin et al., 2016). Zhu et al. (2020) illustrated that domain adaptive transfer learning improved motion sensor-based human identification based on features extracted from rich sets of wearable motion sensor data. Our paper is connected to this type as we leverage state-of-the-art transductive transfer learning (i.e., domain adaptive transfer learning via adversarial training) to resolve the labeling issue for platform operations.

Finally, unsupervised transfer learning is used when domains and tasks are different. It thus aims to solve unsupervised machine learning problems where the source and target domains contain no labeled data. For example, Shen et al. (2020) used unsupervised transfer learning in loan rejection inference analysis to estimate loan applicants' possible repayments for better credit scoring.

A growing body of literature has demonstrated the potential of transfer learning. To realize its potential, it is imperative to understand the notion of source-to-target transferability and to develop techniques to quantify transferability to avoid negative transfer (Pan & Yang, 2010). Early works to quantify transferability were mainly theoretical studies (Bao et al., 2019; Ben-David et al., 2007; Mansour et al., 2009). Later works have proposed various ways to construct an empirical measure of transferability (Achille et al., 2019; Bao et al., 2019; Nguyen et al., 2020; Tran et al., 2019; Zamir et al., 2019). However, the proposed measures were either very complex, not easy to interpret, or less generalizable due to the strong assumptions implied. Besides, these works mainly focused on computer vision and quantifying the transferability between source and target tasks (e.g., sentiment detection vs. deception detection) rather than the domains (e.g., fake general news vs. fake financial news).

This paper introduces a transferability score to quantify domain transferability using a similarity measure that is very simple to apply and calculate. This score is also informative and convenient as it falls between 0 and 1, which provides an easy and fast assessment of source–target domain transferability in transfer learning. Besides, this score is generalizable as it can be applied to any machine learning model that can generate a feature map/vector, which is very common in deep learning. Lastly, the score depends on "domain" only but not "class," as it is generated from models that do not train on labeled target domain data. From a resource allocation perspective, online platforms can leverage our score to assess domain transferability first before starting to collect labels.

3 | DOMAIN ADAPTIVE TRANSFER LEARNING VIA ADVERSARIAL TRAINING

In this study, we leverage collective human intelligence using domain invariant linguistic features extracted from a source domain consisting of three types of deceptive general news labeled as fake, biased, or clickbait. These news articles are classified as such by human labelers. Afterward, domain adaptive transfer learning via adversarial training is applied to transfer only relevant linguistic features to three target domains—political news, financial news, or online reviews. Our model works in a semisupervised way, with source and target domain data trained together. Specifically, we train a deep learning model to extract useful linguistic features to differentiate deceptive general news from trustworthy news (supervised learning) and then apply transfer learning within a domain adaptation framework to assess the deceptiveness of content in the three target domains (unsupervised learning). This approach assumes that the extracted linguistic features are discriminative in content and invariant across the source and target domains. Finally, we fine-tune the pre-trained domain adaptive transfer learning model with labeled target domain data to represent our augmented AI approach.

3.1 | Theoretical background

First, we outline why and how domain adaptive transfer learning works. Prior studies have devised theoretical measures to assess how well a domain adaptive model performs in a target domain (Ben-David et al., 2007; 2010). Formally, we consider a binary classification task in which X denotes the input space and $Y = \{0, 1\}$ denotes the set of possible labels. We let S be a labeled source sample of size n_S drawn independently and identically distributed (i.i.d.) from the source domain D_S and let T be an unlabeled target sample of size n_T drawn i.i.d. from the target domain D_T^X over the input space X. We let the hypothesis class $\mathcal{H}$ be a set of binary classifiers $\theta : X \to \{0, 1\}$. According to Ben-David et al. (2010) and Ganin et al. (2016), the generalization bound on the target error for every $\theta \in \mathcal{H}$ is

$$\varepsilon_T(\theta) \leq \varepsilon_{\hat{S}}(\theta) + \hat{d}_{\mathcal{H}}\left(\hat{S}, \hat{T}\right) + O\left(\sqrt{\text{Complexity}/n_S}\right) + \vartheta, \tag{1}$$

where $\varepsilon_T(\theta)$ is the target error, $\varepsilon_{\hat{S}}(\theta)$ is the empirical source error, $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$ represents an empirical divergence term that captures the distance between the source and target domain distributions, $O(\sqrt{\text{Complexity}/n_S})$ is a constant complexity term weighted by source sample size that is expected to be small, and $\vartheta \geq \inf[\varepsilon_S(\theta^*) + \varepsilon_T(\theta^*)]$. Equation (1) shows that the target error bound depends on a good hypothesis θ^* that minimizes the combined source and target error ϑ. We assume that there exists such an ideal hypothesis θ^* and our goal is to train a classifier that exhibits performance that approximates this hypothesis, given that we have no label information about the target domain. Equation (1) also shows that the target error bound depends on $\varepsilon_{\hat{S}}(\theta)$ and the divergence between the source domain and target domain distributions $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$. In particular, $\varepsilon_{\hat{S}}(\theta)$ increases as $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$ decreases, because minimizing the divergence between the source domain and target domain distributions discards source domain specific features that are important to minimizing the source error. Hence, the objective of domain adaptive transfer learning is to minimize the trade-off between $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$, which is equivalent to learning features that are discriminative in the source domain (which reduces $\varepsilon_{\hat{S}}(\theta)$) and invariant across domains (which reduces $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$). If domain adaptation is not implemented, the value of $\varepsilon_{\hat{S}}(\theta)$ is decreased, but the divergence between domains increases. Therefore, compared to a model with domain adaptation, a model without domain adaptation will achieve poorer performance in the target task if the increase in the divergence term $\hat{d}_{\mathcal{H}}(\hat{S}, \hat{T})$ outweighs the reduction in $\varepsilon_{\hat{S}}(\theta)$.

3.1.1 | Fine-tuning with labeled target domain data

As explained, the best model configuration for a situation where no information on the target label is available should achieve a minimum balance between the source error and domain divergence, as expressed in Equation (1). However, suppose a small number of labeled examples (e.g., verified deceptive examples) are present in the target domain. In that case, it may be possible to increase model performance by fine-tuning the model parameters and supplementing target domain specific features for prediction (Daumé, 2007; Ganin et al., 2016). This involves replacing the original empirical source error in Equation (1) with a weighted average of the empirical source error and the empirical target error based on the small number of labeled target examples (Ben-David et al., 2010; Daumé, 2007; Ganin et al., 2016; Mansour et al., 2009; P. Rai et al., 2010). Consequently, the theoretical target error bound defined by Equation (1) can be further reduced.

3.2 | Overview of the model framework

A domain adaptive transfer learning via adversarial training framework is depicted in Figure 3. It is developed with reference to the domain adversarial training for

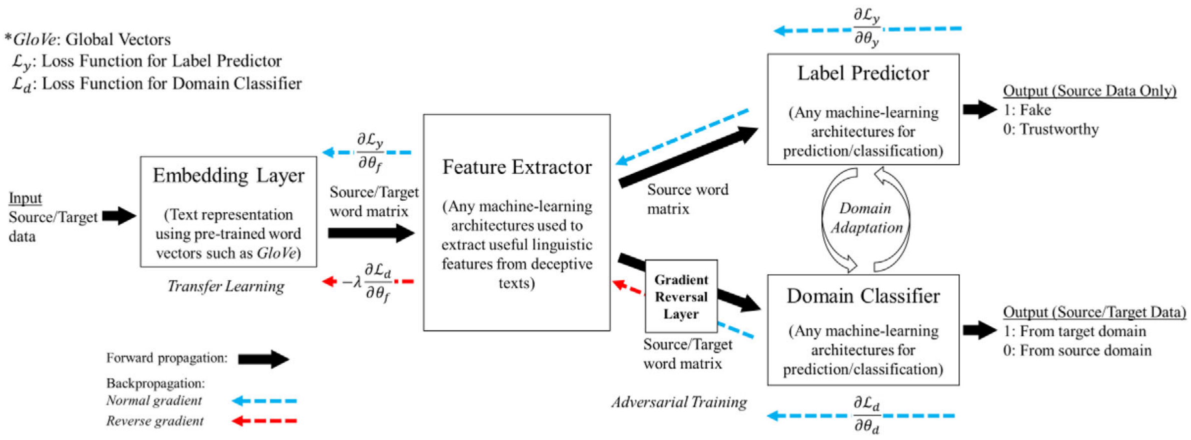

*GloVe: Global Vectors
$\mathcal{L}_y$: Loss Function for Label Predictor
$\mathcal{L}_d$: Loss Function for Domain Classifier
Input
Source/Target data
Embedding Layer
(Text representation using pre-trained word vectors such as GloVe)
Transfer Learning
$\frac{\partial \mathcal{L}_y}{\partial \theta_f}$
Source/Target word matrix
$-\lambda \frac{\partial \mathcal{L}_d}{\partial \theta_f}$
Feature Extractor
(Any machine-learning architectures used to extract useful linguistic features from deceptive texts)
Source word matrix
Gradient Reversal Layer
Source/Target word matrix
Adversarial Training
$\frac{\partial \mathcal{L}_y}{\partial \theta_y}$
Label Predictor
(Any machine-learning architectures for prediction/classification)
Domain Adaptation
Domain Classifier
(Any machine-learning architectures for prediction/classification)
$\frac{\partial \mathcal{L}_d}{\partial \theta_d}$
Output (Source Data Only)
1: Fake
0: Trustworthy
Output (Source/Target Data)
1: From target domain
0: From source domain
Forward propagation:
Backpropagation:
Normal gradient
Reverse gradient

FIGURE 3 A model framework for fake content detection. [Color figure can be viewed at wileyonlinelibrary.com]

transfer learning outlined by Ganin et al. (2016), which is the state-of-the-art method in domain adaptation (Chen et al., 2022; Y. Shi et al., 2022). It consists of four modules: an embedding layer, a feature extractor, a label predictor, and a domain classifier (the adversarial part). Learning occurs as training data are channeled through these modules, from left to right. The embedding layer module transforms text content (e.g., deceptive and non-deceptive text) into meaningful numerical representations. Various methods can be used in this module, such as traditional methods based on bag-of-words, term frequency-inverse document frequency, or more advanced techniques based on word embedding. Next, the feature extractor (represented by $\mathcal{G}_f(\cdot;\theta_f)$ with parameters θ_f) extracts deceptive linguistic features from training data, which can be handled by any machine learning architecture capable of extracting common features of fake text content. In adversarial domain adaptation, source and target domain data are used together for training. After a training example (i.e., x_i) passes through the feature extractor, it takes one of the two paths: If it originates from the source domain, it passes through the label predictor (represented by $\mathcal{G}_y(\cdot;\theta_y)$ with parameters θ_y) and the domain classifier (represented by $\mathcal{G}_d(\cdot;\theta_d)$ with parameters θ_d); if it originates from the target domain, it passes through the domain classifier only. The label predictor, with a loss function defined as $\mathcal{L}_y^i(\theta_f,\theta_y) = \mathcal{L}_y(\mathcal{G}_y(\mathcal{G}_f(x_i;\theta_f);\theta_y), y_i)$, aims to detect whether source domain data are trustworthy or fake (i.e., a label y_i), whereas the domain classifier, with a loss function defined as $\mathcal{L}_d^i(\theta_f,\theta_d) = \mathcal{L}_d(\mathcal{G}_d(\mathcal{G}_f(x_i;\theta_f);\theta_d), d_i)$, aims to classify input data as source domain or target domain data (i.e., a label d_i). However, it is adversely designed to "perform worse" in this task by introducing the gradient reversal layer between the feature extractor and domain classifier. Specifically, the gradient reversal layer contains no parameters and hence does not require any parameter update. It has no impact during the forward propagation; however, during the backpropagation, the gradient learned from the domain classifier (i.e., $\frac{\partial \mathcal{L}_d}{\partial \theta_d}$) will be multiplied by a negative constant (i.e., $-\lambda$) when passing through the gradient reversal layer back to the feature extractor (i.e., $-\lambda\frac{\partial \mathcal{L}_d}{\partial \theta_f}$). In this regard, we ensure that the model is trained in an adversarial way to maximize the domain classification loss so that, in the end, it is indistinguishable between source and target domain data. Mathematically stated, the combined loss function under this domain adversarial training is defined as

$$C\left(\theta_f,\theta_y,\theta_d\right) = \frac{1}{n_S}\sum_{i=1}^{n_S}\mathcal{L}_y^i\left(\theta_f,\theta_y\right) - \lambda\left(\frac{1}{n_S}\sum_{i=1}^{n_S}\mathcal{L}_d^i\left(\theta_f,\theta_d\right) + \frac{1}{n_T}\sum_{i=n_S+1}^{n_S+n_T}\mathcal{L}_d^i\left(\theta_f,\theta_d\right)\right),$$

from which $(\hat{\theta}_f,\hat{\theta}_y) = \text{argmin}_{\theta_f,\theta_y}\, C(\theta_f,\theta_y,\hat{\theta}_d)$ to be achieved by the feature extractor and the label predictor and $\hat{\theta}_d = \text{argmax}_{\theta_d}\, C(\hat{\theta}_f,\hat{\theta}_y,\theta_d)$ to be achieved by the domain classifier.

In sum, the label predictor and the domain classifier function to simultaneously train a model that is able to identify fake data (i.e., is discriminative) but unable to differentiate between source and target domains (i.e., is domain invariant), which ensures that the model only transfers invariant features. In other words, we can perceive this training approach as introducing a domain regularizer that prevents the model from distinguishing the origin of the input data, which is achieved by the inclusion of the domain classifier and the gradient reversal layer. Equation (1) shows that the divergence term $\hat{d}_{\mathcal{H}}(\hat{S},\hat{T})$ is crucial as it governs the target error bound $\varepsilon_T(\theta)$. It is thus important to determine this term to understand how good domain adaptation performs in theory. We can approximate $\hat{d}_{\mathcal{H}}(\hat{S},\hat{T})$ by training a classifier that differentiates between source and target domain data to capture the divergence between source and target domains. As suggested by Ganin et al. (2016), this is equivalent to using the generalization error term ϵ obtained from the domain classifier to approximate $\hat{d}_{\mathcal{H}}(\hat{S},\hat{T})$ as follows:

$$\hat{d}_{\mathcal{H}}\left(\hat{S},\hat{T}\right) \cong \hat{d}_{\mathcal{A}} = 2\left|1-2\epsilon\right|, \quad (2)$$

where $\hat{d}_{\mathcal{A}}$ is called the *proxy $\mathcal{A}$-distance* (Ganin et al., 2016). A well-performing domain adaptive model has a domain classifier error ϵ approximately equal to 0.5, which indicates that it performs unbiased domain classification. Therefore, for a given pair of source and target domains, we can use $\hat{d}_{\mathcal{A}}$ together with $\varepsilon_{\hat{S}}(\theta)$ to assess the feasibility of domain adaptation.

We develop a domain adaptive transfer learning model that instantiates the framework (Figure 3) and uses the model (and its variants) for analysis and validation in the remainder of this paper. More details of the model are provided in Supporting Information Appendix B.

4 | MODEL DEVELOPMENT

This section provides more information on data collection, model construction, and model training.

4.1 | Source domain data

To construct the source domain dataset, we collect labeled deceptive general news articles from a comprehensive list of deceptive websites (Zimdars, 2016), which are curated by volunteers from the OpenSources project. This list is derived from various lists from the Internet and fact-checking sites (e.g., Wikipedia[3] and Snopes.com[4]) and compiled using six heuristic rules, including writing style analysis, which reflects opinion-based linguistic features of deceptive news. This list has been referenced in various university libraries (e.g., NDNU library[5] and NJS library[6]) and research studies (Allcott et al., 2019; Grinberg et al., 2019; Guess et al., 2018). It contains 1001 news websites, each tagged with various labels (e.g., fake, biased) that indicate the types of deception. We focus on news sites with one of three identified labels ("fake," "biased," or "clickbait"), which results in 223 sites for news collection. Our focus on these three labels is in line with the definition of fake content stated earlier, and these labels capture the intentions to achieve various goals, such as distorting facts, misleading the audience, and attracting attention (see Table 1). We first collected content from these 223 labeled sites to create three datasets of deceptive news corresponding to fake, biased, and clickbait to provide insights into their differences in the prediction task in the three target domains. We then combine these three datasets to form a single dataset of deceptive news of the source domain to conduct further analyses. By doing so, we aim to identify as many opinion-based linguistic features as possible in the source domain and transfer them to the target domains to identify fake content. A complete list of each type of deceptive news site is given in Supporting Information Appendix C.

To collect trustworthy news articles, we refer to a Wikipedia list[7] of 80 reliable news sites that are frequently referred to by readers with high consensus. Table 1 summarizes the definitions and descriptive statistics for the news article data collected for the source domain. We collect news articles from these sites over a period from 2014 to 2018. In brief, the labeled general news dataset, serving as the source domain data, is constructed from lists of deceptive and trustworthy websites agreed upon by diverse human opinions. We thus extract and transfer linguistic features from these news articles to represent human opinion on truth and deception.

4.2 | Target domain data

We consider three target domains for fake content detection to conduct a comprehensive analysis. The first domain is political news, and the corresponding dataset, which contains fake and authentic political news articles, is obtained from data repositories like *FakeNewsNet* (Shu et al., 2017; 2020). Both fake and authentic political news articles are identified by PolitiFact.com, which is a reputable source for fact-checking political news. We collect 856 fake political news articles and 8767 authentic articles from the repositories, and these comprise our first target domain dataset. We observe that fake and authentic political news articles have similar word statistics, such as word count, average words per sentence, and function words.

The second domain is financial news, for which we obtain 383 verified fake financial articles from Clarke et al. (2021) and X. Zhang et al. (2022). The Securities and Exchange Commission (SEC) has verified that these financial news articles are biased news written for monetary compensation to promote stocks.[8] These fake financial news articles were published on various financial websites from August 1, 2011, to December 31, 2013. We also use the Factiva API to obtain legitimate financial news articles published in *The Wall Street Journal*, and these comprise a dataset of 68,409 news articles from the same period as above. We then construct the second target domain dataset by combining the 383 verified fake items with the 68,409 legitimate items. We observe that legitimate financial news articles contain fewer words overall and fewer words per sentence than verified fake financial news articles.

The third domain is online reviews, for which we collect fake and authentic online reviews from Yelp that have been used in other studies (Mukherjee et al., 2013; Rayana & Akoglu, 2015; D. Zhang et al., 2016). Yelp uses its own review filtering algorithm to identify fake or suspicious reviews, which are added to a filtered list that is publicly available. It has been shown that Yelp's fake review detection algorithm is sufficiently accurate and reliable, such that the reviews it identifies as fake are close to the ground truth (Mukherjee et al., 2013; D. Zhang et al., 2016). We obtain

TABLE 1 Definitions and descriptive statistics for source domain data

	Trustworthy	Fake	Biased	Clickbait
Definitions (Zimdars, 2016)	News published by media that circulate news and information in a manner consistent with traditional and ethical journalism practices.	News published by media that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports.	News published by media that present a particular point of view and may rely on propaganda, decontextualized information, and opinions presented as facts.	News published by media that provide generally credible content but use exaggerated, misleading, or questionable headlines, social media descriptions, and images. These news outlets may also use sensational language to generate interest, clickthroughs, and shares, but their content is typically verifiable.
Number of articles	1,913,222	894,746	1,138,998	231,949
Number of sites	80	107	95	21
Number of authors	102,926	9094	39,147	7495
Average word count	510 (566)	451 (661)	440 (733)	396 (388)

Note: Standard deviations are shown in parentheses.

a random sample of 4268 fake reviews and 34,818 authentic reviews from the filtered list of Yelp and use these reviews to construct our third target domain dataset. We observe that authentic reviews contain more words overall and more words per sentence than fake reviews, while fake and authentic reviews contain similar numbers of complex words (words with more than six letters), function words, and punctuation marks. All three domains exhibit the labeling problem to various degrees.

As shown in previous studies in Supporting Information Appendix A, machine learning models have typically been trained on relatively small datasets in fake content detection, as few verified fake examples are available. Although some studies have used balanced datasets, these datasets did not correspond to actual population distributions. We validate our approach using highly imbalanced datasets in target domains to mimic real scenarios for which verified fake data are scarce.

4.3 | Model preparation and training

Before preparing the models, we first compile four versions of source domain datasets. The first three source domain datasets are constructed by combining deceptive news articles of each type with randomly sampled trustworthy news articles of equal number. For example, we combine 894,746 fake news articles with 894,746 randomly sampled trustworthy news articles. The remaining source domain news dataset is constructed by combining deceptive news articles (i.e., 231,949 randomly sampled fake news articles + 231,949 randomly sampled biased news articles + 231,949 clickbait news articles) with an equal number of randomly sampled trustworthy news articles (i.e., 695,847 randomly sampled trustworthy news articles).

To comprehensively compare different models under the general transfer learning framework in Figure 3, we consider four specific types of transfer learning models to be described in detail below. We provide complete pseudocode for training each type of transfer learning in Supporting Information Appendix D.

4.3.1 | Baseline: Transfer learning without domain adaptation

The first type is simple transfer learning without domain adaptation. Specifically, we train this type of model using the architecture depicted in Supporting Information Appendix B by discarding the domain-classifier module and using only the source domain dataset. We then apply the trained models to directly score the deceptiveness of content in the three target domains. No target domain data are used for model training. To simplify the interpretation of results, we label each model as follows. We first use the following symbols to indicate which version of the source domain dataset is used for model training: "*Fake_*" for fake and trustworthy news dataset, "*Bias_*" for biased and trustworthy news dataset, "*Ckbt_*" for clickbait and trustworthy news dataset, and "*Pool_*" for pooled all deceptive and trustworthy news dataset. We then use the symbol "*TLnoDA*" to indicate that these models are transfer learning without domain adaptation. In sum, we end up with four models denoted by *Fake_TLnoDA*, *Bias_TLnoDA*, *Ckbt_TLnoDA*, and *Pool_TLnoDA*.

4.3.2 | Baseline: Transfer learning with domain adaptation

The second type is transfer learning with domain adaptation. We follow the procedures described in Section 3 to train the model on source domain data and non-fake data in the target domains. The source domain data ensure that the model has sufficient predictive power to discriminate deceptive from trustworthy news articles, while the non-fake data in the target domains guide the training process to extract only domain invariant features. We merge the source domain dataset with 8767 authentic political news articles, 68,409 legitimate financial news articles, or 34,818 authentic online reviews that represent the three target domains. To emphasize again, no data labeled as fake in the target domains are used for model training. Similarly, we denote models in this type by changing the symbol to "*TLDA*," for example, *Fake_**TLDA***, *Bias_**TLDA***, *Ckbt_**TLDA***, and *Pool_**TLDA***.

In brief, we regard these two types of transfer learning with/without domain adaptation as baselines for comparison since they only rely on opinion-based linguistic features without leveraging the labeling information of the target domains. These two types of models simulate the situation where the target domain's label on fake versus non-fake is unavailable. In addition, we consider a random guess classifier (*RG*) that predicts the two classes with equal probability and is thus the simplest classification baseline.

4.3.3 | Augmented AI: Transfer learning with domain adaptation and fine-tuning on a small sample of the target domain

The third and fourth types of transfer learning models represent the augmented AI approach proposed in this study. The third type is transfer learning with domain adaptation, fine-tuned with a few labeled data in the target domain. We use this type to illustrate that the performance of a domain adaptive model comprised of opinion-based linguistic features can be increased by supplementation with a very small number of target fake examples. Thus, after preparing the pre-trained domain adaptive model (the second type of model described above), we randomly sample a very small amount of labeled data in an approximate ratio of 1:10 from the target domain, resulting in random samples of 172 labeled political news articles, 80 labeled financial news articles, and 860 labeled online reviews. We consider a ratio of 1:10 as it is consistent with the 10-fold cross-validation applied to the pure AI models described later. These small samples of fake content are used to fine-tune the label predictor of the pre-trained models, with the other model parameters, held constant. Based on the second type of model, we denote models belonging to the third type by extending the symbol with "*_FTsmall*" to indicate that they are fine-tuned with a small amount of labeled data in the target domain. We thus have four models named *Fake_TLDA_**FTsmall***, *Bias_TLDA_**FTsmall***, *Ckbt_TLDA_**FTsmall***, and *Pool_TLDA_**FTsmall***.

4.3.4 | Augmented AI: Transfer learning with domain adaptation and fine-tuning on full sample of the target domain

The fourth type is an extension of the third type and uses all labeled target domain data. Specifically, we fine-tune the trained domain adaptive models using all available fake content and randomly sampled non-fake samples of equal size. This type of model is the ultimate augmented AI model to enable comparison with the direct application of traditional machine learning to the target domain dataset. We denote models based on this type by *Fake_TLDA_FT**full***, *Bias_TLDA_FT**full***, *Ckbt_TLDA_FT**full***, and *Pool_TLDA_FT**full***, with the last part named "*full*" to indicate that full labeled target domain data is used for fine-tuning. Supporting Information Appendix E summarizes the details of all types of transfer learning models and their corresponding short names.

For each transfer learning model and each training epoch, we randomly split the source domain news into two sets of 75% and 25% each. The 75% set is used for training, and the 25% set is used to test the performance of the model and to calculate the empirical source error. We combine the training set generated from the source domain news articles with an equal number of random target domain data items to form the final training set for the domain adaptive models. Except for models that incorporate fine-tuning, only non-fake target domain data are used. We run each model for at least 50 epochs to ensure that appropriate model convergence is achieved. In all cases, we use Python with the PyTorch and TorchText libraries and run the model on a server using Intel Xeon E5 CPUs and four Nvidia Tesla P100 GPU cards.

4.3.5 | Pure AI and evaluation metrics

As detailed in Supporting Information Appendix F, we consider two direct machine learning approaches (pure AI) for comparison. The first approach is a machine learning model trained directly on the full target dataset using all the available labeled data. This approach corresponds to the best performance a direct machine learning model can achieve when it uses all the available information. The second approach is a machine learning model trained using a small subset (10%) of verified fake content in the target dataset. Thus, we envision a situation in which the first approach corresponds to the actual population distribution of fake content in the target domain, which is unknown in practice, whereas the second approach corresponds to the sample distribution of the target domain, which represents only a small subset of observable and verified fake content that is scarce and hard to obtain in practice. To minimize algorithmic bias, three popular machine learning techniques (a multi-layer perceptron classifier, a random forest classifier, and a multinomial naïve Bayes model) are implemented, and their average performance is used as the performance measure of each approach.

All the models (augmented AI models, pure AI models, and the baselines) are compared in terms of four metrics: balanced accuracy, precision, recall, and F1 score, with a focus on balanced accuracy. We also conduct paired t-tests to assess whether there are statistically significant differences between model performances (Abbasi et al., 2015). More details on the construction of pure AI models and the evaluation metrics are provided in Supporting Information Appendix F.

TABLE 2 Domain adaptation performance

	Target error bound[a]			Domain transferability[b]		
Source domain dataset	$\varepsilon_{\hat{S}}(\theta)$	$\hat{d}_{\mathcal{A}}$	$\varepsilon_{\hat{S}}(\theta) + \hat{d}_{\mathcal{A}}$	$cos(\bar{V}_{S,noDA}, \bar{V}_{T,noDA})$	$cos(\bar{V}_{S,DA}, \bar{V}_{T,DA})$	*TranScore*
Political news						
Fake	0.0433	0.0012	0.0445	0.8782	0.9817	0.8946
Biased	0.0889	0.1374	0.2263	0.8922	0.9947	0.8970
Clickbait	0.0436	0.1035	0.1471	0.8763	0.9925	0.8829
Political news[c]	0.1650	0.0000	0.1650	1.0000	1.0000	1.0000
Financial news						
Fake	0.0859	0.2424	0.3283	0.8659	0.9987	0.8670
Biased	0.0517	0.1331	0.1848	0.8518	0.9976	0.8538
Clickbait	0.0845	0.0520	0.1365	0.8713	0.9978	0.8732
Financial news[c]	0.0052	0.0000	0.0052	1.0000	1.0000	1.0000
Online reviews						
Fake	0.0464	0.1465	0.1929	0.7166	0.9950	0.7202
Biased	0.0695	0.1434	0.2129	0.7546	0.9928	0.7601
Clickbait	0.0931	0.1137	0.2068	0.7339	0.9934	0.7388
Online reviews[c]	0.1745	0.0000	0.1745	1.0000	1.0000	1.0000

[a] $\varepsilon_{\hat{S}}(\theta)$ is the empirical source error that is computed from the label predictor error based on the last epoch. $\hat{d}_{\mathcal{A}}$ is the *proxy $\mathcal{A}$-distance* that is computed from Equation (2) using the generalization error obtained from the minimum value of the domain classifier errors based on 1000 iterations, where each iteration contains equal numbers of randomly sampled source domain and target domain data.

[b] $\cos(\bar{V}_{S,noDA}, \bar{V}_{T,noDA})$ ($\cos(\bar{V}_{S,DA}, \bar{V}_{T,DA})$) is the cosine similarity between the centroids of feature maps of source and target domain data generated from the model without (with) domain adaptation. *TranScore* is the transferability score computed from Equation (3) based on these two cosine similarity measures.

[c] We train models using the same target domain as the source domain for reference to examine the ideal best performance of domain adaptation when there is no divergence between domains.

5 | MODEL VALIDATION

Based on the model setup described before, we compare the prediction performance of the augmented AI approach, represented by the domain adaptive transfer learning with fine-tunning, and the pure AI approach in three target domains (political news, financial news, and online reviews) that contain highly unbalanced examples with different levels of domain transferability.

5.1 | Domain adaptation performance

We first assess the effectiveness of domain adaptation, as reported in Table 2. For reference, we also train models using the same target domain as the source domain so that there will be no divergence between domains. The value of the empirical source error $\varepsilon_{\hat{S}}(\theta)$ is obtained from the label predictor in the last epoch. For transfer learning models without domain adaptation, the values of $\varepsilon_{\hat{S}}(\theta)$ are all less than 2%, meaning that these models achieve greater than 98% predictive accuracy in differentiating deceptive news from trustworthy news. When domain adaptation is implemented, the values of $\varepsilon_{\hat{S}}(\theta)$ increase. According to Equation (1), this is expected because some domain specific linguistic features that only discriminate source domain data are filtered out. Nevertheless, these models demonstrate good performance by achieving greater than 90% predictive accuracy in the source domain. The empirical source error of the models does not differ substantially across the three target domains.

Next, we examine the *proxy $\mathcal{A}$-distance* $\hat{d}_{\mathcal{A}}$ computed according to Equation (2) for each target domain. We use the following procedure to obtain the generalization error term to calculate $\hat{d}_{\mathcal{A}}$. First, we construct a random sample containing 10,000 source data points and 10,000 target data points. Second, we apply the domain adaptive transfer learning model to this sample to obtain the domain classifier error. Third, we repeat the previous two steps 1000 times and retain only the minimum value as the generalization error term (Ganin et al., 2016). We sum $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{A}}$ to obtain an approximate upper bound of the target error, which indicates the efficacy of a domain adaptive model. A large value suggests that the model fails to achieve domain adaptation, as it fails to retain the discriminant power in detecting fake content (i.e., a large $\varepsilon_{\hat{S}}(\theta)$) or to reduce divergence between the source domain and the target domain (i.e., a large $\hat{d}_{\mathcal{A}}$). Thus, this sum provides an indication of whether the assumption of domain adaptation holds.

For the political news domain, the average value of $\hat{d}_{\mathcal{A}}$ is 0.0807. The best source domain news is achieved by fake general news, as it delivers the smallest sum of $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{A}}$ (0.0445). This theoretical target error bound is much smaller than that obtained from the model trained on the same data for both source and target domains (0.1650), meaning that the transfer learning model with domain adaptation can conceptually outperform the direct learning model. For the financial news domain, the average value of $\hat{d}_{\mathcal{A}}$ is 0.1425. The best source domain news is achieved by clickbait general news, as it minimizes the trade-off between $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{A}}$, achieving a sum of 0.1365. For the online review domain, the average value of $\hat{d}_{\mathcal{A}}$ is 0.1345, and the best source domain news is achieved by fake general news, as it delivers a sum of 0.1929. Overall, none of our source domain news has a sum of $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{A}}$ greater than 0.5.

5.1.1 | Domain transferability

Although a small sum of $\varepsilon_{\hat{S}}(\theta)$ and $\hat{d}_{\mathcal{A}}$ suggests that a specific domain adaptive model may be effective, it provides little information on the extent of domain transferability between source and target domains. In this regard, we propose a measure that assesses domain adaptation performance by quantifying the level of transferability. As shown in Figure 3 and Supporting Information Appendix B, the feature extractor fits the model by compressing the incoming input via convolution and pooling it into a final feature map that should be a good representation of the original input. Thus, a feature map generated from a domain adaptive model should represent features that are domain invariant and transferable between domains, while the feature map generated from a non-domain adaptive version of the same model will represent features that are domain specific. Based on this concept, we propose a transferability score to quantify domain transferability and outline the steps as follows. First, we pass source and target domain data through the domain adaptive model to obtain a feature map for each input data. We then compute the average of feature maps of source and target domain data separately to obtain two centroids. If two domains are very similar, the two centroids will be sufficiently close to each other as well. We can thus apply a similarity function to these source and target domain centroids to capture their proximity. We repeat these procedures for the model without domain adaptation and finally calculate the score as

$$TranScore = \cos\left(\bar{V}_{S,noDA}, \bar{V}_{T,noDA}\right) / \cos\left(\bar{V}_{S,DA}, \bar{V}_{T,DA}\right), \tag{3}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity function,[9] $\bar{V}_{S,noDA}$ ($\bar{V}_{T,noDA}$) is the source (target) domain centroid generated from the model without domain adaptation, and, similarly, $\bar{V}_{S,DA}$ ($\bar{V}_{T,DA}$) is those generated from the domain adaptive model. The cosine similarity measure is informative, as we apply the rectified linear unit function to ensure that only positive values are retained in the feature map. Note that, theoretically, the similarity between the source and target domains with adaptation (denominator) is always larger than that between the source and the target domains without adaptation (numerator) as the model will retain common features and ignore discriminating features after adaptation, so the score ideally falls between 0 and 1.

The transferability scores of all three domains are shown in Table 2. We also provide a visual analysis of domain transferability in Supporting Information Appendix G for robustness check. First, from the table, we observe that the cosine similarity between source and target domain centroids are all very near to 1 after domain adaptation, suggesting that domain adaptive models are successful in adapting the feature space of source and target domains. Next, we observe that the average transferability score is 0.89 for the political news domain, which indicates that the source and political news domains are similar and contain many shareable features. The average transferability score for the financial news domain is 0.86, which is just slightly lower than that of the political domain. Finally, the average transferability score for the online review domain is 0.74 that is the lowest among the three target domains. Based on this gradation of the transferability scores, we denote the domain transferability in the order of political news, financial news, and online reviews, respectively.

5.2 | Relative performance based on three target domains

We now assess the overall performance of augmenting fake content detection by domain invariant linguistic features in three target domains. It is expected that (1) all learning models (both augmented AI and pure AI models) will outperform the random guess model, (2) the domain adaptive transfer learning model will outperform the non-domain adaptive transfer learning model, and (3) the domain adaptive transfer learning model fine-tuned with a set of fake content (augmented AI models) will outperform the direct machine learning models that use the same set of fake content (pure AI models). The relative performance of the four transfer learning models depends on how well common opinion-based linguistic features are transferred from the source domain to the target domain. Thus, we hypothesize that the relative performance of the different models for a particular target domain will follow an order as depicted in Table 3. To test our hypothesis, we consider the version of the source domain dataset that combines all types of opinion-based deceptive news. Therefore, when we compare the relative performance of various models, we mainly focus on these seven models: two augmented AI models (*Pool_TLDA_FTsmall* and *Pool_TLDA_FTfull*), two pure AI models, and three baselines (*Pool_TLnoDA*, *Pool_TLDA*, and *RG*).

5.2.1 | Results for fake political news

We first examine the results of the political news validation (a high transferability domain), which are summarized in Table 4. We confirm that the non-domain adaptive transfer learning model (*Pool_TLnoDA*) is the worst-performing transfer learning model, with low precision and only approximately 50% balanced accuracy. Nevertheless, it still outperforms a random guess classifier.

TABLE 3 Summary of three different learning perspectives

	Model	**Expected relative performance (1 = best)**	**Opinion-based linguistic features (source domain)**	**Labels/features of fake content (three target domains)**
Augmented AI	Transfer learning with domain adaptation and fine-tuning on full sample of the target domain	1	✓	All
	Transfer learning with domain adaptation and fine-tuning on small sample (10%) of the target domain	3	✓	A small subset (10%)
Pure AI	Direct learning on full sample of the target domain	2	–	All
	Direct learning on small sample (10%) of the target domain	4	–	A small subset (10%)
Baseline	Transfer learning with domain adaptation	5	✓	–
	Transfer learning without domain adaptation	6	✓	–
	Random guess	7	–	–

Notes: Domain transferability: political news ~ financial news > online reviews.
Abbreviation: AI, artificial intelligence.

As expected, the domain adaptive transfer learning model (*Pool_TLDA*) performs better than the non-domain adaptive model (*Pool_TLnoDA*), as the former achieves a higher value of balanced accuracy and has better precision and high recall. In addition, the domain adaptive model (*Pool_TLDA*) that does not include examples of fake news articles in the target domain performs similarly to the direct learning models trained on a small sample.

We then consider the transfer learning model with domain adaptation that is fine-tuned using a small subset of the available fake political news articles (*Pool_TLDA_FTsmall*) and find that *Pool_TLDA_FTsmall* has approximately 2.93% better balanced accuracy than the corresponding transfer learning model with domain adaptation that is not fine-tuned (*Pool_TLDA*). Furthermore, the direct learning models trained on a small sample have an average balanced accuracy and F1 score of 52.49% and 15.40%, respectively, and achieve low precision and moderate recall. Interestingly, we observe that the performance of the domain adaptive model (*Pool_TLDA_FTsmall*), which is fine-tuned with a small subset of fake content, is significantly better than the direct learning on a small sample and approaches the direct learning models that use all fake content in its target domain.

Finally, we consider the case in which all labeled target domain data are used to fine-tune the domain adaptive model. We observe that the resulting model (*Pool_TLDA_FTfull*) significantly outperforms the pure AI models, which provides further evidence of the effectiveness of augmented AI models. The results of the first target domain confirm our expectations regarding the relative performance of these transfer learning models (Table 3).

5.2.2 | Results for fake financial news

Table 5 summarizes the results from our analysis of financial news articles (a domain with transferability slightly lower than the political news domain). Similar to the analysis of the political news domain, we observe that the non-domain adaptive transfer learning model (*Pool_TLnoDA*) significantly outperforms a random guess classifier (by >6.41%) in identifying fake financial news articles with respect to balanced accuracy. The *Pool_TLDA* model has significantly better performance than the *Pool_TLnoDA* model, as shown by its balanced accuracy and F1 scores.

In addition, the domain adaptive model (*Pool_TLDA*) outperforms the direct learning models trained on a small sample in terms of balanced accuracy, without any fake financial news articles present in its target domain. Similarly, we find that fine-tuning substantially improves its performance, as the *Pool_TLDA_FTsmall* model has approximately 17.94% higher balanced accuracy than the *Pool_TLDA* model and a higher F1 score. The fine-tuned domain adaptive model (*Pool_TLDA_FTsmall*) achieves better results than the *Pool_TLDA* model by more accurately identifying fake financial news articles. Furthermore, we observe that the *Pool_TLDA_FTsmall* model substantially outperforms the direct learning models trained on a small sample (by approximately 41.40% on average) and has a balanced accuracy close to the direct learning models using all available fake financial news articles.

The performance of the domain adaptive model fine-tuned using all labeled data (*Pool_TLDA_FTfull*) is slightly lower than the performance of direct learning models based on full datasets. This may be attributable to a small set of labeled target domain data (comprising a random sample of 40 fake financial news articles) being sufficient for the domain adaptive model such that supplementation with additional labeled target domain data is of no additional benefit. The results obtained from the financial news domain are largely consistent with those obtained from the political news domain, further confirming the relative performance order depicted in Table 3.

TABLE 4 Validation results for fake political news

	Balanced accuracy	Precision	Recall	F1	Difference in balanced accuracy	Paired *t*-test *p*-value
Direct learning on full sample						
Multi-layer perceptron	0.5973[a] (0.0176)	0.4466[a] (0.0446)	0.2219[a] (0.0387)	0.2941[a] (0.0369)	−8.06% (vs. *Pool_TLDA_FTfull*)	<0.001*
Random forest	0.5468[a] (0.0171)	0.8073[a] (0.1385)	0.0969[a] (0.0340)	0.1719[a] (0.0560)	−13.11% (vs. *Pool_TLDA_FTfull*)	<0.001*
Naïve Bayes	0.5248[a] (0.0124)	0.8900[a] (0.1556)	0.0502[a] (0.0246)	0.0941[a] (0.0444)	−15.31% (vs. *Pool_TLDA_FTfull*)	<0.001*
Direct learning on small sample						
Multi-layer perceptron	0.5139[a] (0.0100)	0.1132[a] (0.0235)	0.1530[a] (0.0432)	0.1247[a] (0.0153)	−3.59% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Random forest	0.5266[a] (0.0245)	0.0998[a] (0.0128)	0.5249[a] (0.1095)	0.1661[a] (0.0162)	−2.32% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Naïve Bayes	0.5343[a] (0.0234)	0.1001[a] (0.0084)	0.6109[a] (0.0965)	0.1712[a] (0.0117)	−1.55% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on full sample[b]						
Fake_TLDA_FTfull	0.6571	0.1444	0.7453	0.2420	11.02% (vs. *Fake_TLDA_FTsmall*)	<0.001*
Bias_TLDA_FTfull	0.6775	0.1716	0.6717	0.2734	12.23% (vs. *Bias_TLDA_FTsmall*)	<0.001*
Ckbt_TLDA_FTfull	0.6999	0.1893	0.6869	0.2968	15.92% (vs. *Ckbt_TLDA_FTsmall*)	<0.001*
Pool_TLDA_FTfull	0.6779	0.2003	0.5829	0.2982	12.81% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on small sample[c]						
Fake_TLDA_FTsmall	0.5469	0.1050	0.5596	0.1768	3.27% (vs. *Fake_TLDA*)	<0.001*
Bias_TLDA_FTsmall	0.5552	0.1051	0.6554	0.1811	5.05% (vs. *Bias_TLDA*)	<0.001*
Ckbt_TLDA_FTsmall	0.5407	0.1187	0.2956	0.1694	3.02% (vs. *Ckbt_TLDA*)	<0.001*
Pool_TLDA_FTsmall	0.5498	0.1120	0.4404	0.1786	2.93% (vs. *Pool_TLDA*)	<0.001*
Transfer learning with domain adaptation						
Fake_TLDA	0.5142	0.0915	0.9311	0.1667	2.38% (vs. *Fake_TLnoDA*)	<0.001*
Bias_TLDA	0.5047	0.0897	0.9848	0.1645	2.40% (vs. *Bias_TLnoDA*)	<0.001*
Ckbt_TLDA	0.5105	0.0907	0.9871	0.1662	1.95% (vs. *Ckbt_TLnoDA*)	<0.001*
Pool_TLDA	0.5205	0.0935	0.7710	0.1667	1.78% (vs. *Pool_TLnoDA*)	<0.001*
Transfer learning without domain adaptation						
Fake_TLnoDA	0.4904	0.0822	0.2150	0.1190	−0.96% (vs. *RG*)	<0.001*
Bias_TLnoDA	0.4807	0.0855	0.8657	0.1556	−1.93% (vs. *RG*)	<0.001*
Ckbt_TLnoDA	0.4910	0.0871	0.7827	0.1568	−0.90% (vs. *RG*)	<0.001*
Pool_TLnoDA	0.5027	0.0894	0.9965	0.1641	0.27% (vs. *RG*)	<0.001*
Random guess						
RG	0.5000	0.0890	1.0000	0.1634	–	–

Notes: The results are based on 9623 political news articles, 856 of which are fake political news.

Abbreviation: RG, random guess.

[a] Average value is reported with the standard deviation shown in parentheses.

[b] Fine-tuning is based on 856 fake news articles and 856 randomly sampled real political news articles.

[c] Fine-tuning is based on a random sample of 86 fake news articles and 86 real political news articles.

**p*-values significant at $\alpha = 0.05$.

5.2.3 | Results for fake online reviews

The fake online reviews dataset represents a relatively lower transferability scenario. The corresponding results are detailed in Table 6. We observe that the non-adaptive transfer learning (*Pool_TLnoDA*) achieves reasonable performance in this context, compared with that of a random guess classifier, as the *Pool_TLnoDA* model affords a balanced accuracy of 51.47% and an F1 score of 20.11%. The *Pool_TLDA* model performs similarly to the *Pool_TLnoDA* model. Again, we find that fine-tuning improves transfer learning performance. The domain adaptive model (*Pool_TLDA_FTsmall*), which is fine-tuned with a small sample of the target dataset, significantly outperforms the direct learning models that are

TABLE 5 Validation results for fake financial news

	Balanced accuracy	Precision	Recall	F1	Difference in balanced accuracy	Paired t-test p-value
Direct learning on full sample						
Multi-layer Perceptron	0.9922[a] (0.0104)	0.9974[a] (0.0077)	0.9844[a] (0.0208)	0.9907[a] (0.0104)	6.83% (vs. *Pool_TLDA_FTfull*)	<0.001*
Random Forest	0.8510[a] (0.0315)	1.0000[a] (0.0000)	0.7021[a] (0.0629)	0.8234[a] (0.0437)	−7.29% (vs. *Pool_TLDA_FTfull*)	<0.001*
Naïve Bayes	0.8859[a] (0.0185)	0.8034[a] (0.0431)	0.7728[a] (0.0369)	0.7873[a] (0.0333)	−3.80% (vs. *Pool_TLDA_FTfull*)	<0.001*
Direct learning on small sample						
Multi-layer perceptron	0.5334[a] (0.0432)	0.0505[a] (0.0538)	0.0836[a] (0.1091)	0.0238[a] (0.0195)	−40.32% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Random forest	0.5278[a] (0.0213)	0.1891[a] (0.2005)	0.0574[a] (0.0433)	0.0793[a] (0.0618)	−40.88% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Naïve Bayes	0.5067[a] (0.0689)	0.0056[a] (0.0008)	0.9499[a] (0.1383)	0.0112[a] (0.0017)	−42.99% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on full sample[b]						
Fake_TLDA_FTfull	0.8821	0.3341	0.7728	0.4665	−6.10% (vs. *Fake_TLDA_FTsmall*)	<0.001*
Bias_TLDA_FTfull	0.7492	0.3416	0.5039	0.4072	−9.84% (vs. *Bias_TLDA_FTsmall*)	<0.001*
Ckbt_TLDA_FTfull	0.8534	0.1754	0.7258	0.2825	−1.56% (vs. *Ckbt_TLDA_FTsmall*)	<0.001*
Pool_TLDA_FTfull	0.9239	0.3596	0.8564	0.5066	−1.27% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on small sample[c]						
Fake_TLDA_FTsmall	0.9431	0.0498	0.9922	0.0948	15.24% (vs. *Fake_TLDA*)	<0.001*
Bias_TLDA_FTsmall	0.8476	0.0207	0.9452	0.0406	24.00% (vs. *Bias_TLDA*)	<0.001*
Ckbt_TLDA_FTsmall	0.8690	0.0222	0.9791	0.0435	20.95% (vs. *Ckbt_TLDA*)	<0.001*
Pool_TLDA_FTsmall	0.9366	0.0472	0.9843	0.0901	17.94% (vs. *Pool_TLDA*)	<0.001*
Transfer learning with domain adaptation						
Fake_TLDA	0.7907	0.0131	1.0000	0.0261	10.06% (vs. *Fake_TLnoDA*)	<0.001*
Bias_TLDA	0.6076	0.0071	0.9791	0.0141	15.09% (vs. *Bias_TLnoDA*)	<0.001*
Ckbt_TLDA	0.6595	0.0085	0.9295	0.0168	25.19% (vs. *Ckbt_TLnoDA*)	<0.001*
Pool_TLDA	0.7572	0.0118	0.9661	0.0234	19.31% (vs. *Pool_TLnoDA*)	<0.001*
Transfer learning without domain adaptation						
Fake_TLnoDA	0.6901	0.0091	0.9713	0.0181	19.01% (vs. *RG*)	<0.001*
Bias_TLnoDA	0.4567	0.0046	0.3943	0.0090	−4.33% (vs. *RG*)	<0.001*
Ckbt_TLnoDA	0.4076	0.0041	0.5170	0.0082	−9.24% (vs. *RG*)	<0.001*
Pool_TLnoDA	0.5641	0.0064	1.0000	0.0127	6.41% (vs. *RG*)	<0.001*
Random guess						
RG	0.5000	0.0056	1.0000	0.0111	–	–

Notes: The results are based on 68,792 financial news articles, 383 of which are fake financial news articles.
Abbreviation: RG, random guess.
[a] Average value is reported with the standard deviation shown in parentheses.
[b] Fine-tuning is based on 383 fake financial news articles and 383 randomly sampled legitimate financial news articles.
[c] Fine-tuning is based on a random sample of 40 fake news articles and 40 legitimate financial news articles.
*p-Values significant at $\alpha = 0.05$.

trained on the same small sample of fake reviews, as the former achieve similar accuracy to the latter but a higher recall. However, a comparison of the performance of the domain adaptive model fine-tuned with the full target dataset (*Pool_TLDA_FTfull*) against the direct machine learning models based on full samples shows that the latter performs better. As expected, the effectiveness of transfer learning declines as transferability decreases; this is also an indication of negative transfer (Pan & Yang, 2010), as applying transfer learning reduces the predictive accuracy of models, such that transfer learning models perform worse than models (multi-layer perceptron and random forest) trained directly on the full data sample. Figure 4 summarizes the main comparisons shown in Tables 4–6.

To conclude, the validation results largely match our general expectations regarding the relative performance of the

TABLE 6 Validation results for fake online reviews

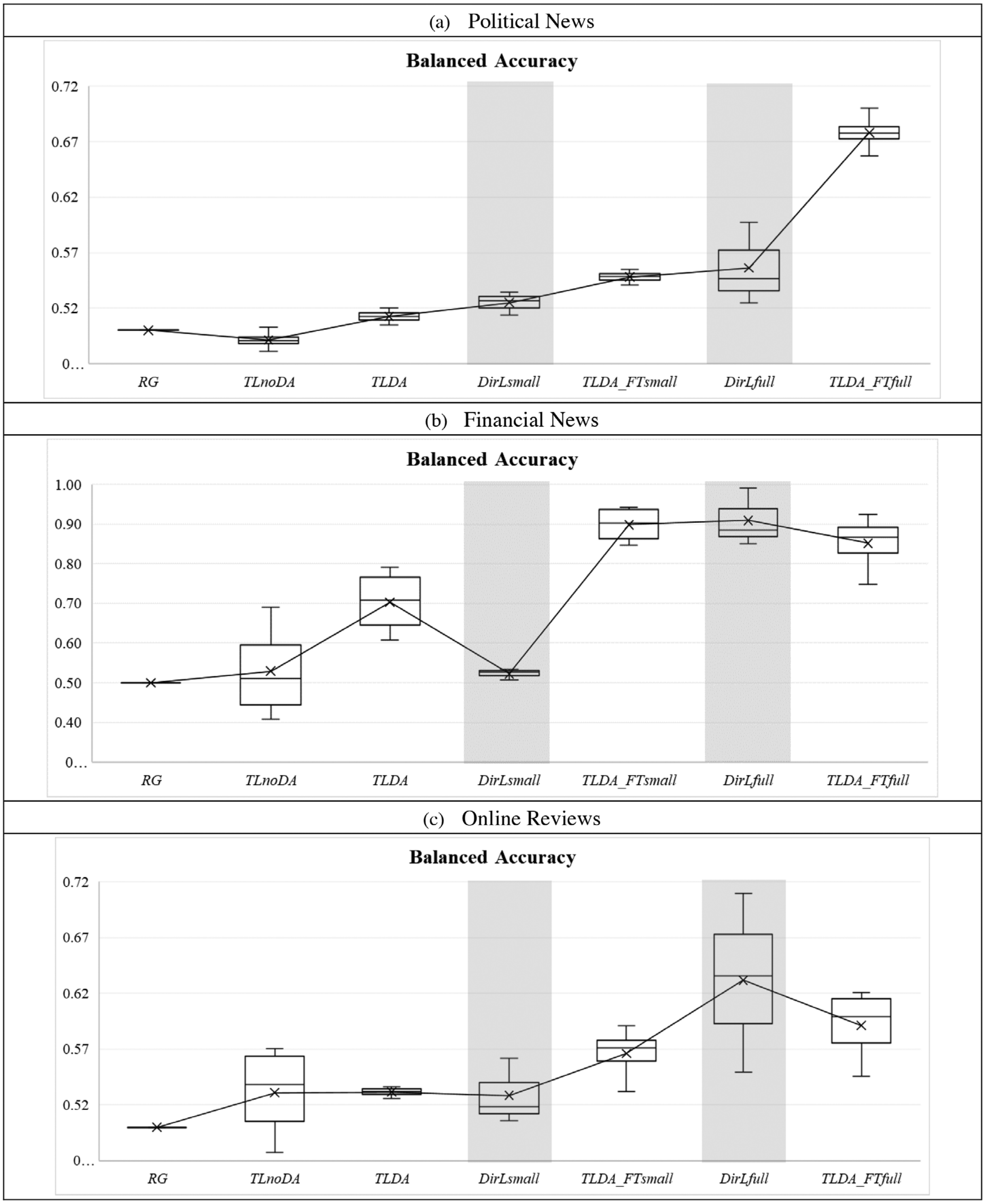
	Balanced accuracy	Precision	Recall	F1	Difference in Balanced Accuracy	Paired *t*-test *p*-value
Direct learning on full sample						
Multi-layer perceptron	0.7098[a] (0.0109)	0.7179[a] (0.0334)	0.4412[a] (0.0250)	0.5454[a] (0.0158)	9.67% (vs. *Pool_TLDA_FTfull*)	<0.001*
Random forest	0.6360[a] (0.0092)	0.8843[a] (0.0237)	0.2765[a] (0.0186)	0.4209[a] (0.0223)	2.29% (vs. *Pool_TLDA_FTfull*)	<0.001*
Naïve Bayes	0.5496[a] (0.0084)	0.8636[a] (0.0387)	0.1012[a] (0.0174)	0.1806[a] (0.0272)	−6.35% (vs. *Pool_TLDA_FTfull*)	<0.001*
Direct learning on small sample						
Multi-layer perceptron	0.5057[a] (0.0040)	0.1603[a] (0.0379)	0.0442[a] (0.0199)	0.0648[a] (0.0207)	−6.25% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Random forest	0.5618[a] (0.0121)	0.1322[a] (0.0083)	0.6784[a] (0.1399)	0.2185[a] (0.0079)	−0.64% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Naïve Bayes	0.5181[a] (0.0116)	0.1261[a] (0.0147)	0.3323[a] (0.1476)	0.1729[a] (0.0164)	−5.01% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on full sample[b]						
Fake_TLDA_FTfull	0.5459	0.1326	0.4625	0.2062	1.39% (vs. *Fake_TLDA_FTsmall*)	<0.001*
Bias_TLDA_FTfull	0.5851	0.1992	0.3355	0.2500	−0.56% (vs. *Bias_TLDA_FTsmall*)	<0.001*
Ckbt_TLDA_FTfull	0.6206	0.1824	0.5354	0.2721	4.69% (vs. *Ckbt_TLDA_FTsmall*)	<0.001*
Pool_TLDA_FTfull	0.6131	0.1862	0.4873	0.2694	4.49% (vs. *Pool_TLDA_FTsmall*)	<0.001*
Transfer learning with domain adaptation and fine-tuning on small sample[c]						
Fake_TLDA_FTsmall	0.5320	0.1169	0.8679	0.2060	−0.43% (vs. *Fake_TLDA*)	<0.001*
Bias_TLDA_FTsmall	0.5907	0.1391	0.7519	0.2348	5.71% (vs. *Bias_TLDA*)	<0.001*
Ckbt_TLDA_FTsmall	0.5737	0.1309	0.7917	0.2246	4.30% (vs. *Ckbt_TLDA*)	<0.001*
Pool_TLDA_FTsmall	0.5682	0.1290	0.7919	0.2219	4.26% (vs. *Pool_TLDA*)	<0.001*
Transfer learning with domain adaptation						
Fake_TLDA	0.5363	0.1177	0.8910	0.2080	5.88% (vs. *Fake_TLnoDA*)	<0.001*
Bias_TLDA	0.5336	0.1166	0.9438	0.2075	−2.78% (vs. *Bias_TLnoDA*)	<0.001*
Ckbt_TLDA	0.5307	0.1160	0.9313	0.2063	−3.95% (vs. *Ckbt_TLnoDA*)	<0.001*
Pool_TLDA	0.5256	0.1197	0.5223	0.1947	1.09% (vs. *Pool_TLnoDA*)	<0.001*
Transfer learning without domain adaptation						
Fake_TLnoDA	0.4775	0.0993	0.4007	0.1591	−2.25% (vs. *RG*)	<0.001*
Bias_TLnoDA	0.5614	0.1377	0.5284	0.2185	6.14% (vs. *RG*)	<0.001*
Ckbt_TLnoDA	0.5702	0.1359	0.6375	0.2240	7.02% (vs. *RG*)	<0.001*
Pool_TLnoDA	0.5147	0.1122	0.9651	0.2011	1.47% (vs. *RG*)	<0.001*
Random guess						
RG	0.5000	0.1092	1.0000	0.1969	–	–

Notes: The results are based on 39,086 online reviews, 4,268 of which are fake online reviews.
Abbreviation: RG, random guess.
[a] Average value is reported with the standard deviation shown in parentheses.
[b] Fine-tuning is based on 4268 fake and 4268 randomly sampled authentic online reviews.
[3] Fine-tuning is based on a random sample of 430 fake online reviews and 430 authentic online reviews.
**p*-Values significant at $\alpha = 0.05$.

various transfer learning models, as depicted in Table 3. A comparison of the pure AI approach using traditional machine learning models shows that the augmented AI approach based on domain adaptive transfer learning effectively alleviates the problem of insufficient labeled data for fake content detection, provided that there is an appropriate level of transferability between domains, as indicated by the transferability score.

6 | DISCUSSION AND CONCLUSION

Online platforms are experimenting with fact-checking and content screening interventions as a regular operation to moderate the adverse effects of deceptive and incensing content (Moravec et al., 2019; Papanastasiou, 2020). However, platforms have limited capacity for manual fact-checking, as it is labor- and time-intensive (Sharma et al., 2019). To


(a) Political News

(b) Financial News

(c) Online Reviews

FIGURE 4 Summary of validation results. (a) Political news. (b) Financial news. (c) Online reviews. *RG*: random guess; *TLnoDA*: transfer learning without domain adaptation; *TLDA*: transfer learning with domain adaptation; *DirLsmall*: direct learning trained on a small sample; *TLDA_FTsmall*: transfer learning with domain adaptation and fine-tuning on a small sample; *DirLfull*: direct learning trained on the full sample; *TLDA_FTfull*: transfer learning with domain adaptation and fine-tuning on the full sample. The shaded areas represent the two pure AI approaches

effectively curb the spread of fake content, intelligent screening systems must be developed that can maximize the accuracy of assessing deceptive content, as failure to do so may lead to devastating consequences that undermine the value of online content (Cui et al., 2018; S. Kumar et al., 2018). Failing to stop fake content from spreading on online platforms will reinforce filter bubbles and generate extreme emotions and highly polarized opinions (Allcott & Gentzkow, 2017; Ng et al., 2021; Vosoughi et al., 2018), or being exploited by threat actors to manipulate the business environment

(Lee et al., 2018). However, legitimate content could be wrongly labeled with the presence of false alarms. This may create an impression of censorship and exaggerated filtering, potentially leading to backlash from the user community (Freeze et al., 2020).

This study adopts an augmented AI with a human intelligence perspective and proposes the domain adaptive transfer learning via an adversarial training framework to address the inefficiency in labeling as well as maximize fake content detection performance. We find that transfer learning with domain adaptation and fine-tuning can effectively extract and transfer opinion-based linguistic features to augment AI-based fake content detection, as domain adaptation ensures only relevant features are transferred and fine-tuning reduces human biases and errors in features learned from crowd-based opinions. We also derive a measure to operationalize the notion of domain transferability. We show that domain adaptive transfer learning offers the most value when the level of transferability is high. In our validation analyses, both augmented AI and pure AI models are tested against the general expectations depicted in Table 3; these expectations are largely confirmed, as shown in Figure 4.

6.1 | Theoretical implications

Our study contributes to research on applying machine learning to detect fake content in online platforms and, to some extent, the literature on human–AI interaction (A. Rai et al., 2019; Yau et al., 2021). The literature has discussed a wide range of scenarios of how we can keep humans and AI in a loop to achieve maximum performance (Fügener et al., 2021; Ge et al., 2021; Lou & Wu, 2021; Raisch & Krakowski, 2021). Our research empirically compares augmented and pure AI approaches using an important and highly relevant context of fake content detection. Our findings largely confirm that fake content detection based on machine learning models can achieve better performance when augmenting domain invariant linguistic features extracted from deceptive and trustworthy news identified based on consensus. We thus contribute a unique use case of having collective human intelligence (opinion-based linguistic features) to supplement the AI model (fake content detection) when there is limited and even no labeled data for training the AI model. This implication is important to online platforms in applying machine learning to operationalize fake content detection on a regular basis.

Second, we explain the efficacy of transfer learning with respect to the transferability between a source domain and a target domain. To this end, a transferability score is developed to quantify the transferability between domains. A low transferability (close to 0) means that few features are shared by the source and target domains, and thus the utility of domain adaptive transfer learning is highly limited. Similarly, a very high transferability (nearly 1) means that the utility of domain adaptive transfer learning is also limited. In this situation, a transfer learning model developed with examples from the source domain already incorporates many shared features of the source and target domains and can thus be applied directly to the target domain. For example, if the comparison depicted in Figure 2 is between smartphones and digital notepads (instead of hotels), which are similar in many ways, models developed using smartphone examples could be applied directly to digital notepads. Thus, the performance gain delivered by domain adaptation would be limited. This finding addresses a research gap regarding the condition that underscores transfer learning performance.

Third, in a more general sense, features extracted from domain adaptation training can be regarded as universal features, as they hold across different domains and contexts (Hao, 2019). These features are particularly useful for increasing the generalizability of machine learning algorithms, as the trained models can be deployed in multiple applications. The invariant property of these features also facilitates a certain amount of model interpretation, as it avoids spurious features that undermine the performance of machine learning models and thus enables better evaluation of the behavior of AI systems (Meske et al., 2022). As an illustration, we outline an approach to interpret and visualize domain invariant linguistic features and present it in Supporting Information Appendix H. Invariant features can be identified by performing domain adaptation via adversarial training using multi-context data, which can be collected at different times, from different locations, or on different subjects. In this regard, important universal information is retained, and spurious correlations are filtered out, leading to more robust and trustworthy machine learning models.

6.2 | Practical implications

A direct practical implication of the result of this study is to illustrate a cost-effective way to implement an AI-based fake content detection model for online platforms such as social media and review websites. On the one hand, domain experts' judgments are required to determine whether domain specific content is fake or not, which is costly to collect for building an AI model. On the other hand, platform users' feedbacks on the content (e.g., like, downvote, flag as inappropriate) are relatively easier to collect. Still, the quality cannot be guaranteed and may result in a poor and biased AI model when these data are used directly for model training. To address the dilemma, platforms could adopt the augmented AI model to leverage inputs from the crowd (in this study, opinion-based linguistic features) to supplement the costly domain specific task. In our empirical analysis, results suggest that when the transferability score is not low, the model performance of a pure AI model based on direct learning with full sample of the target domain is actually close to that of an augmented AI model based on transfer learning with domain adaptation and fine-tuning using a small fraction of the target sample. Given that this augmented AI approach only uses 10% data of the target domain, platforms can save up to 90% cost on expert judgment while maintaining a similar model performance when this approach is adopted.

The findings of this study also have practical implications beyond the detection of fake content on online platforms. Our study sheds light on big data research focused on extracting information signals from unstructured text data coming from different sources (Choi et al., 2018; Einav & Levin, 2014; Z. M. Shi et al., 2020). Big data analytics is at the core of OM since many OM-related problems need to deal with data of high volume, high variety, and high velocity (Choi et al., 2018; Jha et al., 2020). Our proposed domain adaptive transfer learning via adversarial training approach can thus advance big data analytic techniques in OM when addressing uncertainty with learning (e.g., limited or missing label problem). Our model also provides insights into emerging topics like fintech at the interface between OM and IS (S. Kumar et al., 2018). For instance, financial technology applications increasingly rely on NLP to deliver novel financial services to customers (Jagtiani & John, 2018; Jagtiani & Lemieux, 2019). Thomson Reuters News Analytics aggregates various news sentiment analysis techniques to support trading and investment decisions, whereas Sentifi delivers actionable insights by analyzing large-scale financial news data from 13 million media outlets. In these contexts, domain adaptive transfer learning facilitates efficient and scalable inferencing of the information content of news. For example, analyzing the market reaction to domain invariant linguistic features is a novel way to utilize transfer learning, illustrating the potential utility of domain adaptive transfer learning in deciphering financial news for market signaling applications.

6.3 | Limitations and future research

While the current study opens a new avenue for online fake content and business analytics research, it also has several limitations that warrant future research. First, in situations whereby human judgments must be used to determine the veracity of information, our model can be used as an initial screening mechanism to aid fact-checkers. Content with very high/low scores (as determined by the domain adaptive transfer learning model) could be labeled automatically, while content with scores in the intermediate range could be forwarded to checkers for final judgment. The resulting human decisions could then serve as training samples for regular updating of the model. Second, the ability to deceive continuously improves, as deceivers constantly learn from authentic and legitimate content (Moravec et al., 2019; D. Zhang et al., 2016). Therefore, the source domain data used to extract human opinion-based features must be updated regularly. Third, our domain adaptive transfer learning model is likely sensitive to the domain itself as moving content categorized from one domain to another will affect the domain transferability and the amount of domain specific features. While our work leverages datasets from various sources and assumes that text contents are categorized according to the predefined domains (general news, political news, financial news, and online review), a promising direction for future research is to define a domain for a piece of content using machine learning and NLP techniques, such as clustering analysis and topic modeling. We can thus examine the relationship between transfer learning and domain sensitivity. Lastly, this paper considers a purely data-driven approach to identifying domain specific and domain invariant features. In this regard, we consider the linguistic features to be latent and unobservable and do not make any presumption about the forms of features (e.g., lexical, syntactic, semantic, thematic). Future research can directly investigate what kind of linguistic features are regarded as domain invariant in different domains for advancing theoretical understanding. We illustrate this idea in Supporting Information Appendix H to stimulate future studies that seek to build on our work.

ACKNOWLEDGMENTS

The authors are grateful to the department editor, the senior editor, and two anonymous reviewers for their insightful and constructive comments that greatly improve the content and the exposition of the paper. The authors also thank Dr. Hailiang Chen from the University of Hong Kong for sharing the fake financial news dataset. This project was supported by the Theme-based Research Grant on Fintech (T31-608/18N) of the Research Grant Council of Hong Kong.

ORCID

Ka Chung Ng https://orcid.org/0000-0001-7875-8194
Ping Fan Ke https://orcid.org/0000-0002-4205-7801
Mike K. P. So https://orcid.org/0000-0003-0781-8166
Kar Yan Tam https://orcid.org/0000-0003-3242-0184

ENDNOTES

[1] https://www.facebook.com/journalismproject/programs/third-party-fact-checking

[2] https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[3] https://en.wikipedia.org/wiki/List_of_fake_news_websites

[4] https://www.snopes.com/news/2016/01/14/fake-news-sites/

[5] https://library.ndnu.edu/fakenews/identifying

[6] https://libguides.njstatelib.org/facts/fake_news

[7] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources#Stale_discussions.

[8] A copy of the complete list of fake financial news items is available at https://ftalphaville-cdn.ft.com/wp-content/uploads/2017/04/10231526/Stock-promoters.pdf (last accessed May 24, 2018). In this study, we focus on a reduced sample of 383 fake financial news articles, which is provided by Clarke et al. (2021).

[9] Cosine similarity between vectors is defined as $\cos(v_1, v_2) = \frac{(v_1 \cdot v_2)}{\|v_1\| \, \|v_2\|} = \frac{\sum_{i=1}^{n} v_{1i} v_{2i}}{\sqrt{\sum_{i=1}^{n} {v_{1i}}^2}\sqrt{\sum_{i=1}^{n} {v_{2i}}^2}}$, where v_{1i}, v_{2i} are components of vector v_1, v_2 respectively.

REFERENCES

Abbasi, A., Zahedi, F. M., Zeng, D., Chen, Y., Chen, H., & Nunamaker, J. F. (2015). Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems*, *31*(4), 109–157. https://doi.org/10.1080/07421222.2014.1001260

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker, J. F. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, *34*(3), 435–461. https://doi.org/10.2307/25750686

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C., Soatto, S., & Perona, P. (2019). Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 6429–6438). IEEE.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research and Politics*, *6*(2), 1–8. https://doi.org/10.1177/2053168019848554

Bao, Y., Li, Y., Huang, S. L., Zhang, L., Zheng, L., Zamir, A., & Guibas, L. (2019). An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing* (pp. 2309–2313). IEEE.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. In B. Scholkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*. (pp. 137–144). MIT Press.

Bloomfield, R. (2012). Discussion of detecting deceptive discussions in conference calls. *Journal of Accounting Research*, *50*(2), 541–552. https://doi.org/10.1111/j.1475-679X.2012.00448.x

Chen, K., Zhuang, D., & Chang, J. M. (2022). Discriminative adversarial domain generalization with meta-learning based cross-domain validation. *Neurocomputing*, *467*, 418–426. https://doi.org/10.1016/j.neucom.2021.09.046

Choi, T.-M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management*, *27*(10), 1868–1883. https://doi.org/10.1111/poms.12838

Clarke, J., Chen, H., Du, D., & Hu, Y. J. (2021). Fake news, investor attention, and market reaction. *Information Systems Research*, *32*(1), 35–52. https://doi.org/10.1287/isre.2019.0910

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749–1769. https://doi.org/10.1111/poms.12707

Daumé, H. (2007). Frustratingly easy domain adaptation. In *ACL 2007 - Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, 256–263.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1–24. https://doi.org/10.1086/674019

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2020). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*, 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, *45*(3), 1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(1), 1–35.

Ge, R., Zheng, Z., Tian, X., & Liao, L. (2021). Human-robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research*, *32*(3), 774–785. https://doi.org/10.1287/isre.2021.1009

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign* (Working paper). Princeton University.

Hao, K. (2019). *Deep learning could reveal why the world works the way it does*. MIT Technology Review, Massachusetts Institute of Technology.

Ho, S. M., Hancock, J. T., Booth, C., & Liu, X. (2016). Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication. *Journal of Management Information Systems*, *33*(2), 393–420. https://doi.org/10.1080/07421222.2016.1205924

Jagtiani, J., & John, K. (2018). Fintech: The impact on consumers and regulatory responses. *Journal of Economics and Business*, *100*, 1–6. https://doi.org/10.1016/j.jeconbus.2018.11.002

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the Lendingclub consumer platform. *Financial Management*, *48*(4), 1009–1029. https://doi.org/10.1111/fima.12295

Jha, A. K., Agi, M. A. N., & Ngai, E. W. T. (2020). A note on big data analytics capability development in supply chain. *Decision Support Systems*, *138*, 1–9. https://doi.org/10.1016/j.dss.2020.113382

Kanyamibwa, F., & Ord, J. K. (2000). Economic process control under uncertainty. *Production and Operations Management*, *9*(2), 184–202. https://doi.org/10.1111/j.1937-5956.2000.tb00333.x

Kratzwald, B., & Feuerriegel, S. (2019). Putting question-answering systems into practice: Transfer learning for efficient domain customization. *ACM Transactions on Management Information Systems*, *9*(4), 15.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38–48. https://doi.org/10.1016/j.dss.2017.10.001

Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2019). Detecting anomalous online reviewers: An unsupervised approach using mixture models. *Journal of Management Information Systems*, *36*(4), 1313–1346. https://doi.org/10.1080/07421222.2019.1661089

Kumar, S., Mookerjee, V., & Shubham, A. (2018). Research in operations management and information systems interface. *Production and Operations Management*, *27*(11), 1893–1905. https://doi.org/10.1111/poms.12961

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, *39*(1), 19–37. https://doi.org/10.1080/00909882.2010.536844

Lee, S.-Y., Qiu, L., & Whinston, A. (2018). Sentiment manipulation in online platforms: An analysis of movie tweets. *Production and Operations Management*, *27*(3), 393–416. https://doi.org/10.1111/poms.12805

Leng, M., Li, Z., & Liang, L. (2016). Implications for the role of retailers in quality assurance. *Production and Operations Management*, *25*(5), 779–790. https://doi.org/10.1111/poms.12501

Lou, B., & Wu, L. (2021). AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms. *MIS Quarterly*, *45*(3), 1451–1482. https://doi.org/10.25300/MISQ/2021/16565

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. Paper presented at 22nd Conference on Learning Theory, Montreal, QC, Canada.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). *Fake review detection: Classification and analysis of real and pseudo reviews (Report No. UIC-CS-03-2013)*. Department of Computer Science, University of Illinois at Chicago.

Ng, K. C., Tang, J., & Lee, D. (2021). The effect of platform intervention policies on fake news dissemination and survival: An empirical examination. *Journal of Management Information Systems*, *38*(4), 898–930. https://doi.org/10.1080/07421222.2021.1990612

Nguyen, C. V., Hassner, T., Seeger, M., & Archambeau, C. (2020). LEEP: A new measure to evaluate transferability of learned representations.

In *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 7250–7261). Curran Associates.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *ACL-HLT 2011 - Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies* (pp. 309–319). Association for Computational Linguistics.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papanastasiou, Y. (2020). Fake news propagation and detection: A sequential model. *Management Science*, *66*(5), 1826–1846. https://doi.org/10.1287/mnsc.2019.3295

Peng, J. (2020). Learning in dynamic business environments: An application in earnings forecast for public firms. In *Proceedings of the 40th international conference on information systems* (AIS) (p. 6). Springer, Cham.

Rai, P., Saha, A., Daumé, H., & Venkatasubramanian, S. (2010). Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing (ALNLP '10)* (pp. 27–32). Association for Computational Linguistics. https://dl.acm.org/doi/abs/10.5555/1860625.1860629

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, *43*(1), 3–9.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *EMNLP 2017 - Conference on empirical methods in natural language processing* (pp. 2931–2937). Association for Computational Linguistics.

Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In L. Cao, & C. Zhang (Eds.), *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 985–994). Association for Computational Linguistics.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In T. Fornaciari, E. Fitzpatrick, & J. Bachenko (Eds.), *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17). Association for Computational Linguistics.

Rui, H., & Lai, G. (2015). Sourcing with deferred payment and inspection under supplier product adulteration risk. *Production and Operations Management*, *24*(6), 934–946. https://doi.org/10.1111/poms.12313

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, *10*(3), 1–41. https://doi.org/10.1145/3305260

Shen, F., Zhao, X., & Kou, G. (2020). Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. *Decision Support Systems*, *137*, 1–10. https://doi.org/10.1016/j.dss.2020.113366

Shi, Z. M., Lee, G. M., & Whinston, A. B. (2020). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, *40*(4), 1035–1056. https://doi.org/10.25300/MISQ/2016/40.4.11

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., & Synnaeve, G. (2022). Gradient matching for domain generalization. *Proceedings of the 10th international conference on learning representations* (pp. 1–28). Morgan Kaufmann.

Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K. C. (2020). Enhancing social media analysis with visual data analytics: A deep learning approach. *MIS Quarterly*, *44*(4), 1459–1492. https://doi.org/10.25300/MISQ/2020/14870

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, *8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Siering, M., Koch, J. A., & Deokar, A. V. (2016). Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems*, *33*(2), 421–455. https://doi.org/10.1080/07421222.2016.1205930

Taguchi, G. (1985). Quality engineering in Japan. *Communications in Statistics – Theory and Methods*, *14*(11), 2785–2801. https://doi.org/10.1080/03610928508829076

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Toma, C. L., & D'Angelo, J. D. (2015). Tell-tale words: Linguistic cues used to infer the expertise of online medical advice. *Journal of Language and Social Psychology*, *34*(1), 25–45. https://doi.org/10.1177/0261927X14554484

Tran, A., Nguyen, C., & Hassner, T. (2019). Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1395–1405). IEEE.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management Science*, *63*(9), 3090–3110. https://doi.org/10.1287/mnsc.2016.2489

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, Q., Li, B., & Singh, P. V. (2018). Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research*, *29*(2), 273–291. https://doi.org/10.1287/isre.2017.0735

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In R. Barzilay, & M.Y. Kan (Eds.), *Proceedings of The ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics (Long Papers)* (pp. 422–426). Association for Computational Linguistics.

Wei, Z., Xiao, M., & Rong, R. (2021). Network size and content generation on social media platforms. *Production and Operations Management*, *30*(5), 1406–1426. https://doi.org/10.1111/poms.13328

Wu, Y., Ngai, E. W. T., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, *132*, 1–15. https://doi.org/10.1016/j.dss.2020.113280

Xu, Y., Pinedo, M., & Xue, M. (2017). Operational risk in financial services: A review and new research opportunities. *Production and Operations Management*, *26*(3), 426–445. https://doi.org/10.1111/poms.12652

Yan, L., & Pedraza-Martinez, A. J. (2019). Social media for disaster management: Operational value of the social conversation. *Production and Operations Management*, *28*(10), 2514–2532. https://doi.org/10.1111/poms.13064

Yang, F., Mukherjee, A., & Gragut, E. (2017). Satirical news detection and analysis using attention mechanism and linguistic features. In L. Specia, M. Post, & M. Paul (Eds.), *Proceedings of The EMNLP 2017 - Conference on empirical methods in natural language processing* (pp. 1979–1989). Association for Computational Linguistics.

Yau, K. L. A., Lee, H. J., Chong, Y. W., Ling, M. H., Syed, A. R., Wu, C., & Goh, H. G. (2021). Augmented intelligence: Surveys of literature and expert opinion to understand relations between human intelligence and artificial intelligence. *IEEE Access*, *9*, 136744–136761. https://doi.org/10.1109/ACCESS.2021.3115494

Zahedi, F. M., Abbasi, A., & Chen, Y. (2015). Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *Journal of the Association for Information Systems*, *16*(6), 448–484. https://doi.org/10.17705/1jais.00399

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2019). Taskonomy: Disentangling task transfer learning. *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence* (pp. 6241–6245). International Joint Conferences on Artificial Intelligence.

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, *33*(2), 456–481. https://doi.org/10.1080/07421222.2016.1205907

Zhang, X., Du, Q., & Zhang, Z. (2022). A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, *31*(8), 3160–3179. https://doi.org/10.1111/poms.13743

Zheng, L., Liu, G., Yan, C., Jiang, C., Zhou, M., & Li, M. (2020). Improved TrAdaBoost and its application to transaction fraud detection. *IEEE Transactions on Computational Social Systems*, *7*(5), 1304–1316. https://doi.org/10.1109/TCSS.2020.3017013

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

Zhou, L., & Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM*, *51*(9), 119–122. https://doi.org/10.1145/1378727.1389972

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 109.

Zhu, H., Samtani, S., Chen, H., & Nunamaker, J. F. (2020). Human identification for activities of daily living: A deep transfer learning approach. *Journal of Management Information Systems*, *37*(2), 457–483. https://doi.org/10.1080/07421222.2020.1759961

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zimdars, M. (2016). False, misleading, clickbait-y, and/or satirical "news" sources. https://d279m997dpfwgl.cloudfront.net/wp/2016/11/Resource-False-Misleading-Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9D-Sources-1.pdf

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ng, K. C., Ke, P. F., So, M. K. P., & Tam, K. Y. (2023). Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach. *Production and Operations Management*, *32,* 2101–2122. https://doi.org/10.1111/poms.13959