Deception Detection with Feature-Augmentation by soft Domain Transfer

Sadat Shahriar*, Arjun Mukherjee, and Omprakash Gnawali

University of Houston, Houston, Tx 77004, USA
sshahria@cougarnet.uh.edu, arjun@cs.uh.edu, odgnawal@central.uh.edu

Abstract. In this era of information explosion, deceivers use different domains or mediums of information to exploit the users, such as News, Emails, and Tweets. Although numerous research has been done to detect deception in all these domains, information shortage in a new event necessitates these domains to associate with each other to battle deception. To form this association, we propose a feature augmentation method by harnessing the intermediate layer representation of neural models. Our approaches provide an improvement over the self-domain baseline models by up to 6.60%. We find Tweets to be the most helpful information provider for Fake News and Phishing Email detection, whereas News helps most in Tweet Rumour detection. Our analysis provides a useful insight for domain knowledge transfer which can help build a stronger deception detection system than the existing literature.

Keywords: Deception · BERT · LSTM · Phishing · Fake News · Rumour

1 Introduction

Deception in the text implies a deliberate attempt of a sender to misconstrue an affair or create a false impression [2]. Deception in the text can occur in multiple domains like News, Tweets, Emails, and research has been done to detect deception in domain-specific settings [13, 16, 19]. Although deceivers use their con in each domain with a unique style, all kinds of deception have the same agenda of deceiving people. Hence, detecting deception in one domain can be leveraged with detection in the other domain. In this paper, we perform a soft domain transfer by investigating how to harness the power of deception detection in domain A to detect deception in domain B. We also investigate the effectiveness of domain transfer when the source domain is non-deceptive.

Researchers looked at deception from a holistic point of view in the hope of capturing the nuances in the style of deception [12]. However, it is not clear if such a clear pattern exists since deceptions in different domains are very different. Additionally, further investigation is needed to quantify the "help" received from one domain to the other. To this end, there is a significant research gap in achieving the domain knowledge transfer. We define *Deceptive Domain* as

* Corresponding Author

different sources of the information through which deception occurs, and we use fake news, phishing emails, and rumours as deception in different domains. As non-deceptive domains, we use Newsgroup topics, sentiment detection, and Wikipedia ontology detection. Therefore, we formulate our first research question as (**RQ1**): Can knowledge transfer from different domains help improve deception detection? Additionally, our second research question is posed as (**RQ2**) Between the deceptive and non-deceptive domains, which set of domains are most helpful in detecting deception? To answer these questions, we train six different BERT and LSTM models [3,6] for three deceptive and three non-deceptive domains. We collect the intermediate-layer information of the target domain and harness the power of the external domain by combining the intermediate-layer and train a Fully-Connected Deep Neural Network (FC-DNN) to detect deception. In this way of feature augmentation, we leverage the knowledge of other domains in the FC-DNN model by injecting that knowledge into the input information.

The significance of this study is manifold. First, in many domains, deceptive data is significantly scarce. For example, individuals and corporations are reluctant to share the phishing emails they receive to evade embarrassment [1]. Second, with the influx of social media, the information is flown through different domains when a new event emerges. For example, the emergence of COVID-19 created a significant misinformation upsurge in news, tweets, and Facebook posts; thus, learning deception by relying on one domain only results in missing other domain information. Finally, this study can guide researchers to lay out a selective knowledge transfer scheme from different domains and find the generalized pattern of deception. The novelty of our work is, to the best of our knowledge, this is the first work that explores the effectiveness of domain transfer in deception detection, and opens up new avenues of further promising research.

2 Related Works

Hernández-Castañeda et al. proposed a cross-domain deception detection using SVM, where the datasets were of opinions on different topics [5]. They aimed to find a general set of features in different experimental train-test settings. Similar research was done by [14] and [9]. In the Fake News detection task, Pèrez-Rosas et al. performed cross-domain experiment on two different datasets and showed the challenges of generalizability [11]. Gautam and Jerripothula used Spinbot, Grammarly and GloVe-based method to for cross-domain fake news detection [4]. However, these research were done to make the deception detection system topic-agnostic rather than mediums-of-deception agnostic. Hence, harnessing the deception-detection capability in the cross-domain setting remains an unexplored area.

3 Dataset

For the Email domain, we use a phishing email dataset from the Anti-Phishing Pilot at ACM IWSPA 2018 [17]. The training set has 5092 legitimate and 629

phishing emails, and the test data size is 4300, with 3825 legitimate and 475 phishing emails. We label phishing emails as deceptive and non-phishing as non-deceptive. For the News domain, we use LIAR dataset [18], which comes with six labels of the news, namely, True, Mostly-True, Half-True, Mostly-False, False, Pants-on-Fire False. We consider the first two as non-deceptive text, and the last four as deceptive following the work in [15]. For the Tweet domain, the PHEME dataset is used, which had 2402 rumour texts, and 4023 non-rumour texts [20]. We label rumour tweets as deceptive and non-rumour tweets as non-deceptive.

For non-deceptive tasks, three datasets are used. The IMDB movie review dataset comes with 50,000 reviews, labeled as positive or negative [10]. The 20 newsgroups dataset consists of around 18000 samples with labels on newsgroups posts about 20 topics [7]. The Wikipedia topic classification dataset consists of 342,782 articles with 9 topic classes [8]. We randomly sample 10,000 texts for each non-deceptive domain and use 80-20 ratio for the train-test.

4 Methodology

We use two neural models– the Bidirectional Encoder Representations from Transformers (BERT) model and Long Short-Term Memory (LSTM) model as baseline methods [3, 6]. The BERT model is built with transformer layers consisting of encoders and decoders with self-attention capability. We fine-tune our baseline self-domain BERT model, extract the model's last [CLS] layer, and use an FC layer and a softmax for the downstream classification task. LSTMs are an efficient variation of Recurrent Neural Network (RNN) with added long-term dependency solution. We use the sequence of words as the input of a two-layer LSTM model and use an FC layer on top to classify the text.

For the feature augmentation process, we perform the Intermediate Layer Concatenation (**ILC**), which is explained in Figure 1. For the BERT model, we extract the self-domain trained [CLS] layers of different domains and concatenate them, representing our target domain's augmented feature set. Similarly, for the LSTM model, we extract the output from the final LSTM layer representation of different domains and concatenate them. The augmented feature set is then fed to a 2-layer Fully-Connected (FC) model to detect deception. Finally, all the network hyperparameters are set using validation sets generated by sampling 20% data from the training set.

5 Results and Discussion

The Table 1 shows the performance of feature augmentation with different deception domains using the BERT-based ILC models. For the Email domain, we observe that the News domain improves the F1 score of phishing detection by 2.31% and the Tweet domain improves the performance by 4.89%. While both Tweet and News domains are combined, we observe a performance boost in F1-score by 6.60%. For News and Tweet domain, we also observe an improved performance with deceptive feature augmentation. Emails help detect fake news

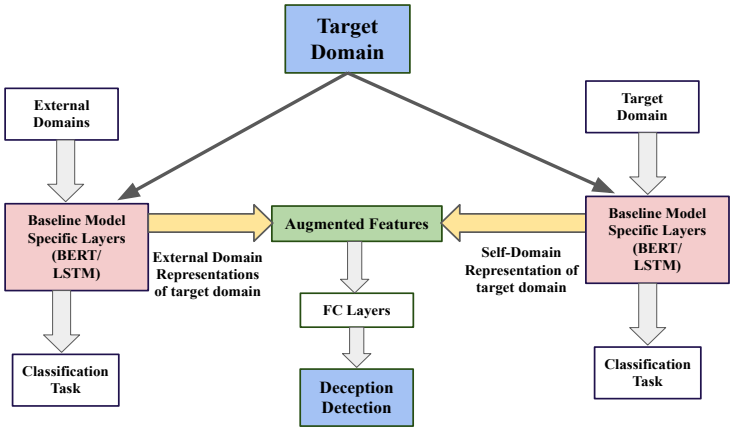

Fig. 1. Feature augmentation by soft domain transfer to improve deception detection. We augment the deceptive features by concatenating the intermediate layer representation of baseline models of both target and external domains and the augmented features are fed to a FC network to detect deception.

by 0.75%, and Tweets help by 1.69%. Tweet rumour detection gets performance improvement of 1.36% from News and 1.14% from Email domain. However, compared to the News and Tweet, performance improvement is higher in the Email domain. Being a pretrained model, BERT is more likely to perform well with public texts like News and Tweet, and thus the baseline model achieves a better understanding of deception in these two domains. Hence, the augmentation from other deceptive domains improves phishing email detection more than deception detection in other domains.

Table 1. Cross-Domain deception detection based on BERT models. **E**, **T**, and **N** stands for Email, Tweet and News respectively. For example, "ILC-TN" stands for ILC model where Tweet and News domains are combined.

	Baseline – BERT		**ILC – EN**		**ILC – TN**		**ILC – ET**		**ILC – ETN**	
Domains	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**
Email	80.99	95.41	83.31	96.03	–	–	85.88	96.79	**87.59**	**97.39**
News	76.88	63.55	77.63	63.93	78.57	**67.80**	–	–	**78.80**	67.56
Tweet	80.34	84.79	–	–	81.70	86.23	81.48	85.99	**82.07**	**86.77**

We further investigate the effectiveness of cross-domain feature augmentation by projecting the data to a 2-D subspace using Singular Value Decomposition (SVD) method. Figure 2 clearly shows an improved feature separation while Tweet and News domains are added with Email, increasing the distance between

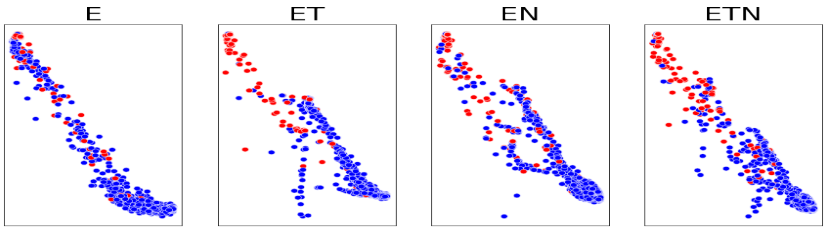

Fig. 2. SVD-reduced representation (BERT model) of Email domain with their self-domain features vs intermediate layer concatenated features with different deception domains. Blue points represent non-deception and red points represent deception

the deceptive and non-deceptive samples' center of cloud by 50.23%, when all three domains are concatenated.

Using the LSTM-based feature augmentation technique, we compromise overall performance, but unlike BERT, we do not use a pretrained model. Therefore, we observe a consistent performance improvement in all three deception domains (Table 2). In the Email domain, like the BERT-based ILC model, the Tweet domain helps the most, and overall improvement is up to 3.97%, with a combined augmentation. Tweets are the most helpful domain both for Email and News. However, the best performance is obtained while all three domains are combined, giving a performance raise of 4.82% in the News domain and 3.39% in the Tweet domain. As standalone domains, News helps the Tweet domain most, providing a boost of 1.76% in the F1-score.

Table 2. Cross-Domain deception detection based on LSTM models.

	Baseline – LSTM		**ILC – EN**		**ILC – TN**		**ILC – ET**		**ILC – ETN**	
Domains	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**
Email	71.23	92.23	72.04	92.45	–	–	74.96	94.68	**75.20**	**94.87**
News	72.71	62.03	74.69	62.43	75.45	63.18	–	–	**77.53**	**63.59**
Tweet	73.11	77.16	–	–	74.87	79.29	74.18	79.00	**76.50**	**82.13**

Table 3. BERT-based Deception detection by feature augmentation from non-deceptive domains.

	Sentiment		**Newsgroup**		**Wikipedia**		**Combined**	
Domains	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**
Email	81.24	96.09	80.95	96.14	81.04	96.13	81.26	96.09
News	77.43	63.10	77.58	63.19	77.89	63.51	78.05	63.79
Tweet	81.41	86.07	81.32	85.91	81.24	85.80	81.89	86.30

Next, we investigate deception detection performance while augmented with non-deceptive domains using BERT models. From Table 3, we observe that Sentiment and Wikipedia slightly improve the performance of phishing email detection, and with combined domains, it improves by 0.27% in F1-score. For the News domain, Wikipedia helps the most, and overall we get a 1.16% improvement in F1 score with all the domains combined. The Sentiment is the most helpful domain for detecting rumour in Tweets, improving the performance by 1.07% in the F1-score, and with the combined domains, the improvement is 1.55%. We also find a similar performance with LSTM-based ILC models, with the best performance in combined domains, improving the Email, News, and Tweet domain deception detection by 2.60%, 5.04%, and 3.10% respectively (Table 4).

From the above discussion, we find that the feature augmentation from different domains helps improve the deception detection task. However, the performance boost is greater when the external domain is deceptive than a non-deceptive one, and thus, a soft domain transfer takes place.

Table 4. LSTM-based Deception detection by feature augmentation from non-deceptive domains.

	Sentiment		**Newsgroup**		**Wikipedia**		**Combined**	
Domains	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**
Email	73.18	93.41	71.49	92.18	72.17	92.77	73.83	94.11
News	72.80	61.97	72.96	62.11	75.50	63.37	76.48	63.41
Tweet	73.93	78.32	73.00	77.91	75.17	81.56	76.21	82.02

6 Conclusion and Future Work

Despite the research on deception detection in many existing domains, there is a research gap on how to harness cross-domain deception detection by transferring the knowledge gained from one domain to the other. In this paper, we bridge the gap using an intermediate-layer concatenation approach from the neural model. There are several future research directions for this work. First, our analysis is limited to three domains only. Several other domains, e.g., reviews, Facebook posts, and Whatsapp message forwards, can also be explored for cross-domain deception detection. Furthermore, we use only one dataset in each domain. Additional research with more datasets in these domains will help solidify our hypothesis.

Acknowledgements The research was supported in part by grants NSF 1838147, ARO W911NF-20-1-0254. The views and conclusions contained in this document are those of the authors and not of the sponsors. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Aassal, A.E., Moraes, L., Baki, S., Das, A., Verma, R.: Anti-phishing pilot at acm iwspa 2018 evaluating performance with new metrics for unbalanced datasets. pp. 2–10. http://ceur-ws.org/Vol-2124/anti-phishing-pilot
2. Burgoon, J.K., Buller, D.B.: Interpersonal deception: Iii. effects of deceit on perceived communication and nonverbal behavior dynamics. Journal of Nonverbal Behavior **18**(2), 155–184 (Jun 1994). https://doi.org/10.1007/BF02170076, https://doi.org/10.1007/BF02170076
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2019)
4. Gautam, A., Jerripothula, K.R.: Sgg: Spinbot, grammarly and glove based fake news detection. 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM) pp. 174–182 (2020)
5. Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., Flores, J.J.G.: Cross-domain deception detection using support vector networks. Soft Computing **21**(3), 585–595 (Feb 2017). https://doi.org/10.1007/s00500-016-2409-2, https://doi.org/10.1007/s00500-016-2409-2
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
7. Lang, K.: Newsweeder: Learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning. pp. 331–339 (1995)
8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**, 167–195 (2015)
9. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1566–1576. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-1147, https://aclanthology.org/P14-1147
10. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-1015
11. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: COLING (2018)
12. Shahriar, S., Mukherjee, A., Gnawali, O.: A domain-independent holistic approach to deception detection. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1308–1317 (2021)
13. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
14. Sánchez-Junquera, J., Villaseñor-Pineda, L., y Gómez, M.M., Rosso, P., Stamatatos, E.: Masking domain-specific information for cross-domain deception detection. Pattern Recognition Letters **135**, 122–130 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2020.04.020, https://www.sciencedirect.com/science/article/pii/S0167865520301422

15. Upadhayay, B., Behzadan, V.: Sentimental liar: Extended corpus and deep learning models for fake claim classification (2020)
16. Varshney, G., Misra, M., Atrey, P.K.: A survey and classification of web phishing detection schemes. Security and Communication Networks **9**(18), 6266–6284 (2016)
17. Verma, R.M., Zeng, V., Faridi, H.: Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. p. 2605–2607. CCS '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3319535.3363267, https://doi.org/10.1145/3319535.3363267
18. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-2067, https://www.aclweb.org/anthology/P17-2067
19. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR) **51**(2), 1–36 (2018)
20. Zubiaga, A., Hoi, G.W.S., Liakata, M., Procter, R.: PHEME dataset of rumours and non-rumours (10 2016). https://doi.org/10.6084/m9.figshare.4010619.v1