Evaluating Fake News Detection Models from Explainable Machine Learning Perspectives

Raed Alharbi, Minh N. Vu, and My T. Thai
Computer and Information Science and Engineering Department
University of Florida
Gainesville, FL, USA
{r.alharbi, minhvu, mythai}@ufl.edu

***Abstract*—Many research efforts recently have aimed at understanding the phenomenon of fake news, including recognizing their common features and patterns, leading to several fake news detection models based on machine learning. Yet, the real distinct strength of those models remains uncertain: some perform well only with particular data, but others are more general. Most of the models classified the fake news as a black-box without giving any explanations to users. In this work, therefore, we conduct an exploratory investigation that evaluates and interprets fake new detection models, including looking into which important features that contribute to the models' prediction from the explainable machine learning perspective. This give us some insights on how the detection models work and their trustworthiness.**

***Index Terms*—Fake news; explainable machine learning; social network**

I. Introduction

More than a decade after their development, social media systems have dramatically changed the way people communicate and interact online, bringing forth an entire unused application stream and reshaping current data environments [1]. In specific, social media platforms have fundamentally changed the direction on how news are distributed, dispersed, and devoured in our community. However, these modifications have begun genuine data war in recent years, boosting drives for misinformation and fake news, diminishing the validity of news venues in these ecosystems [2], and possibly influencing readers' on the risky matters for our community [3], [4].

To battle this, a huge effort on automatically detecting fake news based on machine learning (ML) methods has been investigated intensively [5]–[10]. Generally, a key method of these detection models is to train an ML based model such that it can classify/predict a news as fake or real with high accuracy, which in turns helps fact-checkers to distinguish articles that are deserved to be further examined [11], [12]. Despite the overwhelming significance of the current researches in this area, the true accuracy level of these models (not tested with the similar distribution data that were used for training) is not confirmed and why the model made such a classification is remained as a black-box to users and fact-checkers. Therefore, the usefulness and trustworthy of these fake news detection models are questionable.

Findings. In this paper, we evaluate three representative fake news detection models, namely DEFEND [5], HDSF [7], and TCNN-URG [6]. To understand the classification reasons of these three detection models, we use model-agnostic explainers, a recent research direction in explainable machine learning, which can explain how the classification was made. Our investigation reveals a significant impact of understanding models' predictions with the explanation for fake news detection, summarized as follows.

- Our investigation on the global and local features of efficient fake news detection models reveals unnecessary and incomprehensible features that diminish a human confidence in the detection models.
- Our human subjective experiment shows that the explanation on fake news detection models may weaken the public trust in these models regardless their performance accuracy.
- We found that understanding the trade-off, in many respects, between trusting an explanation or accuracy of a model is crucial if action is planned on the basis of the model's prediction.

Organization. The remainder of the paper is structured as follows. Section II presents the background, involving related works and an overall picture of the common fake news detection models. Experiment setup is introduced in section III then the analysis and discussion of our findings and their implications are presented. Finally, section IV concludes our paper with some discussion on the future direction.

II. Background and Evaluated Models

A. Fake News Detection

In general, the spread of fake news on social media is linked to a number of factors, such as information content and user behaviour. Thus, we categorize current fake new detection models using their key input sources as: models concentrating only on news content and models that utilize both the news content and comments [2]. Those input sources are derived as *textual and visual* bases.

Approaches that focus only on news content include: 1) Knowledge-based methods: the use of external sources, such as experts to examine the statements made in the news content [13]. 2) Style-based methods: extracting distorted and misleading information in writing style, such as craftiness [2]. 3) Discourse-level-based methods: constructing a hierarchical

978-1-7281-0137-8/19/$31.00

978-1-7281-7122-7/21/$31.00 ©2021 IEEE

discourse-level structure for news stories in a data-driven and automated manner [7]. In addition to news content, the social context of news articles, such as social engagement, provides rich knowledge to help detect fake news. The features in social context primarily express users' social comments and news content [13].

For a fair comparison, we choose to analyze in depth three detection models from both categories: DEFEND [5], HDSF [7], and TCNN-URG [6]. A brief summary of these models is provided as follows.

- **DEFEND:** an interpretable fake news detection tool that can utilize both posts and user feedback to collect interpretable top-k check-worthy sentences along with user feedback for identifying fake news.
- **HDSF:** learns a discourse-level structure for fake/real news content in an automated and data-driven manner.
- **TCNN-URG:** TCNN extracts semantic data from news content by representing it at the word and sentence level, and URG learns a generative model of user comments to news content from historical user comments which it can use to produce comments to news articles.

DEFEND is the state-of-the-art detection model that has some self-explanation on its prediction, whereas HDSF utilizes the content's structure to detect fake news. TCNN-URG combines both news contents and user comments features to identify the fake news. Therefore, these three models can well represent the best fake news detection in the literature.

B. Explainable Machine Learning

Regardless ubiquitous adoption, fake news detection models still remain as black-boxes to users. Nonetheless, when evaluating confidence, knowing the reasons behind predictions is very important. Such interpretation provides insights to understand the model, which can be used to gain trust and build a more reliable model [14], [15].

ML explainers can be divided into two groups: post-hoc explainability and intrinsic explainability [5]. Post-hoc achieves interpretability to evaluate the model using different statistical techniques for understanding the model's behaviors [16]–[18]. In contrast, intrinsic explainability is accomplished by limiting the complexity of the machine-learning model [5], [19].

As mentioned earlier, to analyze the trustworthiness of the above three detection models, we delve into the behaviors of these models. In particular, we use ML explainers to understand the reasons why these three models made such predictions. We select three well-known explainers, Captum [20] is the state-of-art interpretable model that build on PyTorch, SHAP [15] leverages game theory methods to reverse engineer any predictive algorithm 's output globally, and LIME which concentrates on explaining individual predictions of models [14]. These explainers are summarized as follows.

- **LIME:** explains in an interpretable and faithful manner the predictions of any classifier, by approximating an interpretable model locally around the prediction.
- **Captum:** consists mainly of primary-based and layer-based attribution. Primary-based attribution investigates one single input feature at time of a model, whereas layer-based attribution concentrates in each neuron in a provided layer. In this paper, we focus on primary-based attribution, which use integrated gradients to identifies global model behavior such as features weights.
- **SHAP:** a theoretical framework that generalizes LIME to clarify any machine learning model 's results.

Platform	PolitiFact
# Users	81,745
# Comments	105,034
# Candidate news	1084
# Fake news	443
# True news	641

TABLE I: The statistics of PolotiFact dataset

III. Model Analysis

This section provides our evaluation on DEFEND, TCNN-URG, and HSFD by using Captum, SHAP, and LIME, which return the top most important features to the models' predictions. By observing these explanations along with the weight features, we can evaluate the trustworthiness of the models. In particular, we aim to answer the following questions:

- **Q1:** How robust are the real/fake features that contribute to the models' classification?
- **Q2:** Does the interpreted outcome reflect the predicted result in term of the most contributed features?
- **Q3:** Which fake news detection model is more desirable to trust? The one with the high accuracy percentage? Or the well explained model?

This section starts with our experimental setup, then a brief re-examination on the performance of DEFEND, TCNN-URG, and HSFD. Next, we discuss in depth our results on using the explainers to evaluate these three models, together with the human subjective testing on the trustworthiness of the detection models.

A. Dataset

We employed one of the holistic fake news detection benchmark dataset called FakeNewsNet [2]. The dataset is obtained from a platform called PolitiFact. It contains 81K users and news articles (105K comments) with their labels and social context data. News articles include the media meta characteristics such as document body and user media contributions to the articles. Note that we maintain news articles with 2 comments or more. Table I displays detailed statistics of the dataset.

B. Evaluation metrics

Because of the unbalanced dataset, we use the following metrics to re-examine the performance of the three detection models in terms of accuracy, precision, recall, and F1 score. It has been proven that these metrics work well on unbalanced dataset [5]–[7].

Accuracy. The ratio of correctly predicted observations to the total observations.

Precision. The ratio of correctly predicted positive observations to the total predicted positive observations.

Recall. The ratio of correctly predicted positive observations to the all observations in actual class.

F1 score. The weighted average of Precision and Recall.

C. Fake News Detection Performance

Before explaining the reasons behind predictions of each detection model, we first evaluate their performance on our selected dataset using the above metrics. We arbitrarily select 77% of the news for training and the remaining 23% for testing. The results are shown in Table II.

Dataset	Metric	DEFEND	HDSF	TCNN-URG
Politifact	Accuracy	**0.82**	0.74	0.80
	Precision	0.79	0.69	**0.80**
	Recall	**0.81**	0.70	0.79
	F1	**0.80**	0.72	0.78

TABLE II: Performance comparison for fake news detection

As shown in Table II, HRSF performs the worst in comparison with the other two methods. This indicates that HRSF, based on only new content, was not able to catch the syntactic and semantic efficiently through hierarchical attention neural networks in posts to differentiate fake and real news.

On the other hand, for the methods that combine the use of posts and user comments, DEFEND and TCNN-URG, we see that DEFEND outperforms TCNN-URG. This implies that features extracted from posts and corresponding user comments in DEFEND have additional details such as extracting features that have most impact on the model prediction, thereby increasing its detection efficiency.

Overall, methods based on both posts and user comments such as DEFEND and TCNN-URG are more preferable. In addition, we can see that DEFEND consistently performs better than TCNN-URG and HSDF in terms of all evaluation metrics on the PolitiFact dataset. For instance, DEFEND accomplishes an average relative improvement of 2.0% comparing with TCNN-URG in terms of accuracy and F1 score.

Content Words	Weights	Comments words	Weights
content	.9345	trump	.77
old	.74428	fakenews	.722
gallup	.564	Knox	.603
it's	.584	wood	.60
remembered	.6023	tweet	.43887
end	.57	you're	.354
I'm	.547	msnbc	.227

TABLE III: The global interpretation of DEFEND model using Captum explainer for top 7 contributed words.

Content Words	Weights
of	.90
2016	.85
ohio	.713
incrdiabily	.697
thank	.695
training	.694
lowering	.734

TABLE IV: The global interpretation of HDSF model using Captum explainer for top 7 contributed words.

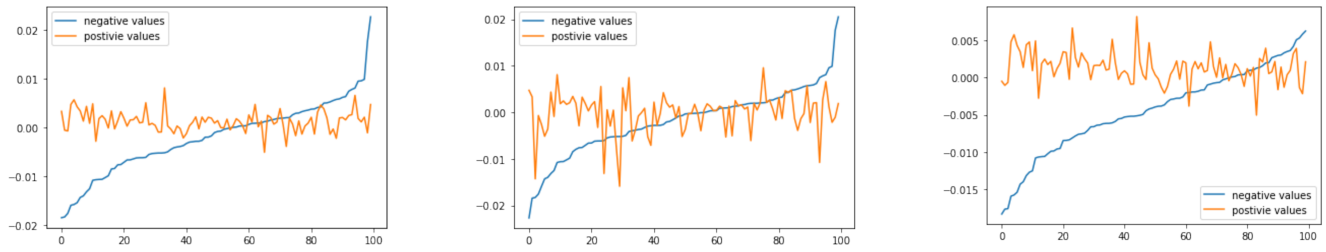
D. Understanding Global Behaviors using SHAP and Captum

Our first goal is to understand and explain the global behaviors of DEFEND, HRSF, and TCNN-URG by using SHAP and Captum. The SHAP values illustrate how much each predictor contributes to the target label, whether positively or negatively. In contrast, the Captum algorithm produces contribution scores which, in descending order, list the most important features [20]. In this set of experiments, we select the top 100 features and average their SHAP values to analyze both positive and negative predictions that affect the three detection models decision.

We first observe the distinct contribution gaps between negative and positive prediction as the bigger the gap, the more trustworthy a model is, as shown in Figure 1. The x-axis shows the features in the increasing order of their important score in negative prediction and the y-axis shows the important score. As can be seen in Figures 1a and 1b for DEFEND and HDSF, we observe a distinct set of features contributing significantly to negative predictions (features 90 to 100). In this case, we can interpret that the models make a negative prediction by observing the occurrence of these features and make positive prediction by not seeing those features (not by observing some other features). On the other hand, from Figure 1c for TCNN-URG, since most features as low score in negative prediction, we can infer that the model relies on the occurrences of some features to decide the positive predictions.

We next analyze the top extracted meaningful words using Captum algorithm, presented in Tables III and IV. The result depicts the top seven used words by the models that assist in detecting fake news, where the weights evaluate the contribution of each input feature to the output of the model. The higher the weight, the more important the feature is. For DEFEND model, the highest captured Captum weights in the contents in Table III represent by the word content, whereas the rest of words' weights represent approximately an average with 0.64 per word. Although importance, however, we observe that some of those words in Table III do not carry much meaning, such as "it's", "gallup", and "I'm". This may not give insights to understand the reason behind the model prediction to classify certain news as fake. For instance, in the sentence "I'm the most humble man in the world," the feature "I'm" will not have much meaningful contribution to the prediction.

Similar behavior can be concluded for HDSF as shown in Table IV. HDSF successfully captures meaningful words that support the classifier to make the correct prediction to some extent. However, some meaningless words are shown such as "of". On the other hand, the TCNN-URG model uses document embedding to effectively capture rich semantic information and text characteristics from not only short, but also long posts. However, the non-invertible embedding used in the model makes it more difficult to interpret using Captum.

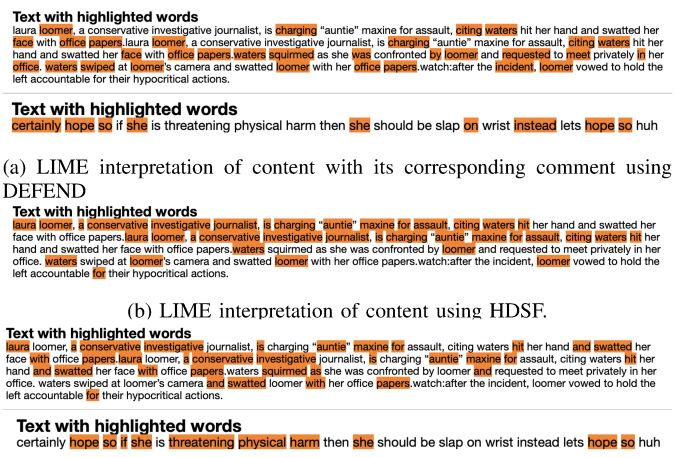


(a) The evaluation of DEFEND for the top 100 features

(b) The evaluation of HDSF for the top 100 features

(c) The evaluation of TCNN-URG for the top 100 features

Fig. 1: The analysis of global behaviors for DEFEND, HDSF and TCNN-URG models using SHAP explainer.



(a) LIME interpretation of content with its corresponding comment using DEFEND

(b) LIME interpretation of content using HDSF.

(c) LIME interpretation of content with its corresponding comment using TCNN-URG

Fig. 2: The local interpretation for DEFEND, HDSF and TCNN-URG models using LIME explainer for one real news prediction.

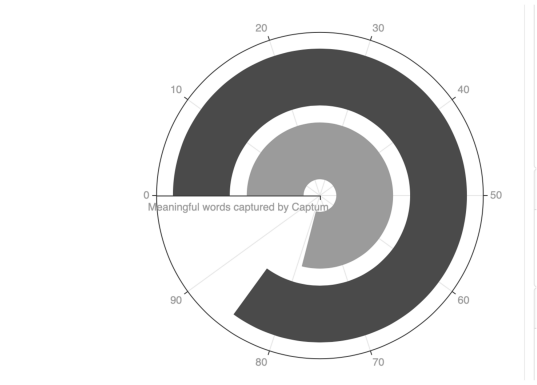
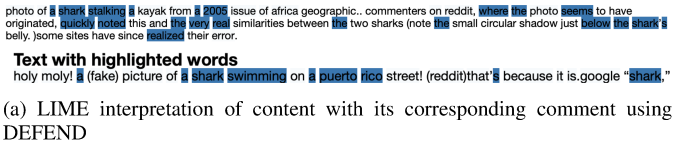


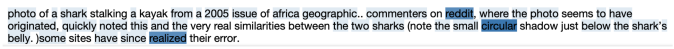
Fig. 3: Evaluation of top 100 interpretable words for DEFEND (black color) and HDSF (gray color).

To better understand the impact of key features, we extract the top 100 captured words for correct predictions and then **eliminate meaningless** words. As can be seen in Figure 3 DEFEND outperforms HDSF in term of extracted meaningful words by approximately 7.5%.

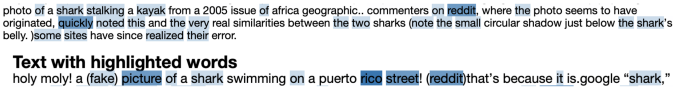
Overall, TCNN-URG has more distinct contribution gaps



(a) LIME interpretation of content with its corresponding comment using DEFEND



(b) LIME interpretation of content using HDSF.



(c) LIME interpretation of content with its corresponding comment using TCNN-URG

Fig. 4: The local interpretation for DEFEND, HDSF and TCNN-URG models using LIME explainer for one correct fake news prediction.

other than DEFEND and HDSF, whereas DEFEND, in terms of interpretability step, beats TCNN-URG and slightly HDSF. HDSF also effectively describes features better than TCNN-URG, which, due to non-invertible embedding used in the model, does not interpret any features.

E. Understanding Local Behaviors using LIME

To better visualize and understand the local behaviors of the three evaluated detection models, we randomly select one real news and one fake news from the PolotiFact dataset. The detection models successfully classify these two news as real news and fake news, accordingly. We next use LIME to highlight the top 15 features that contribute the most to the models' predictions. The results are depicted in Figures 2 (for the real news sample) and 4 (for the fake news sample).

As can be seen in Figures 2a, 2b, and 2c, DEFEND, HDSF, and TCNN-URG in general have captured the key related features. However, among these top 15 features, there are still some less meaningful words. More specifically, DEFEND, using its co-attention layer, aims to capture the important features among the content and comment [5]. In our real news sample, DEFEND has relied its decision on the important feature words such as "loomer" and "requested". Unfortunately, it also highlights "was", "by", and "in." Likewise, the word "hope

so" in the associated comment of the news is also captured as one of the most important features. Note that the provided explanation in Figure 2a shows that highlighted words do not capture the main related features between the content and the comment.

Similarly, Figure 2b shows that HDSF is able to capture the hierarchical discourse-level structure to some extents by highlighting the following parts "Laura, loomer er, a consecutive investigative journalist, is charging "auntie" for assault, citing water hit", but it does not explain well the whole structure by misidentifying the other part of the content which related the highlighted sentence such as "she was confronted by loomer". As a consequence, the structure-related properties of HDSF framework break down the major structural discrepancies between false and real news [7].

As shown in 2c, TCNN-URG has a slightly better performance in the explainable perspective. The top part of Figure 2c reflect the TCNN content level while the bottom part is the URG level. As noted, the words "consecutive", "investigative", and "swatted" are the most supportive. Additionally, the URG level shows that the comment is related to the article, and it supports the claim that "lura has to punished." Taking a closer look, we notice that the words "treating" and "harm" are captured and assist the model to predict the right prediction. Unfortunately, it also captures pointless words "so".

On the other hand, Figures 4a, 4b and 4c illustrate the top captured features by LIME that help correctly identify the fake news sample (highlighted in blue). Similar to what we have observed earlier, DEFEND partially is able to detect key important features such the word "shark", but it also captures less meaningful words. HDSF behaves the worst as can be seen in Figure 4b. There are several light blue highlighted words, indicating that the words are important, but their weight are negligible. This explanation may mislead users. TCNN-URG, again, performs the best in term of explanation. It extracts the main key features between the content and the comment as shown in Figure 4c.

F. Understanding False Positive Behaviors

Up to now, we see that DEFEND outperforms the other two models in terms of accuracy, F1 score and recall. Also, Captum values show that the more understandable and meaningful words were captured by DEFEND than that of HDSF and TCNN-URG. However, TCNN-URG is better at capturing the important common features between the content and user comments using LIME interpretation. SHAP also shows that TCNN-URG has more unique contribution gap among positive and negative samples. Which model is more desirable?

We examine the three detection models from another angle: the false positive prediction. We randomly select one fake news sample which the models misclassified as a real news. We use LIME to detail which features, highlighted in blue, constitute to the fake news detection, and which ones, highlighted in orange, play an important role in classifying the news as real. The results are shown in Figures 5, 6, and 7.

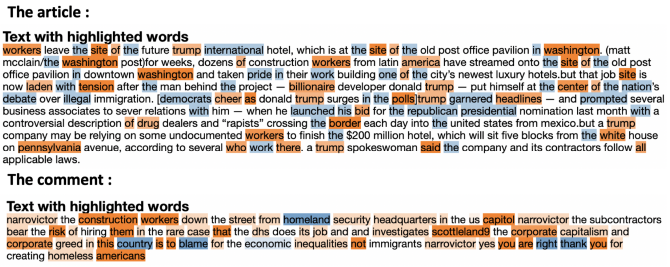


Fig. 5: Sample false positive prediction for DEFEND

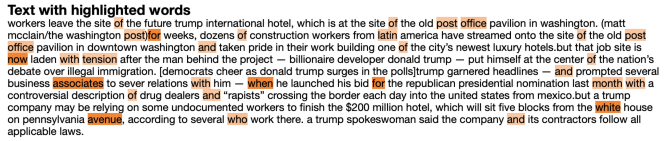


Fig. 6: Sample false positive prediction for HDSF

For content-based method, HDSF does not give any sign (blue highlight) for the news to be fake, as can be seen in Figure 6 and therefore does not make the right prediction. This result indicates that HDSF totally misclassified the news.

Although DEFEND and TCNN-URG also misclassified the news as well, both of these models show some fake news signals (blue highlight), as shown in Figures 5 and 7, respectively. The blue highlighted words provide some clues about why the paragraph (Figure 5) may be fake, which may give some insights for the fact checkers. Furthermore, TCNN-URG explained the article slightly better than DEFEND. The explicit relationship between the content and comment was predicted as fake in TCNN-URG. For instance, in the article sentence "leave **the site for** future Trump **international hotel**", and the comment sentence "**construction** workers **down** the **street**" in Figure 7, TCNN-URG was able to highlight in blue (indicator for fake news) the main claim words and its responding words. In other words, TCNN-URG successfully captured the relationship between the content and the comments to some extent. In contrast, DEFEND detected that relationship as indicator for real news (orange highlight).

G. Human-Subjective Experiment

In our human-subjective test, we randomly chose 30 sample fake posts whose predictions made *correctly* by DEFEND, HDSF, and TCNN-URG using the Polilifact dataset. Then we use LIME to explain their prediction locally. Thirty six end-users were asked to give a score on a scale of 1 (Untrustworthy) to 5 (Most trustworthy) for the explanation based on the their thoughts on the importance of the highlighted text for each model prediction.

From the distribution of the scores in Figure 8, the explanations of TCNN—URG model are evaluated higher than those of other methods. In addition, the average score of the three models DEFEND, HDSF and TCNN-URG is approximately 3.05, 2.5 and 3.5, respectively. The single most striking observation to emerge from the data comparison is that TCNN-URG

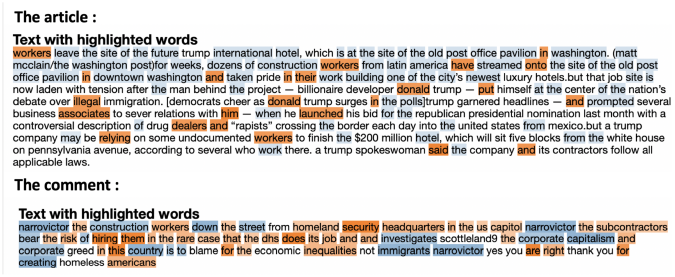


Fig. 7: Sample false positive prediction for TCNN-URG

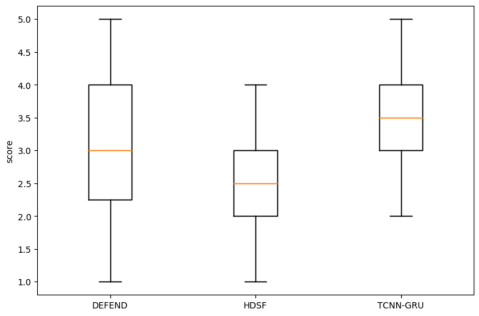


Fig. 8: Score of explanation

outperform DEFEND and HDSF, whereas DEFEND beats the other two models in term of the accuracy performance as illustrated in section III-C. Therefore, the result has further strengthened our observation on that the explanation does a major role in gaining trust than that of acuracy.

IV. Conclusions and Future Work

Recently, the issue of fake news has gained much attention. However, most of current efforts concentrate on identifying fake news but not for explaining its decision. Thus, this paper analyzed fake news detection models using Captum, LIME, and SHAP expainers. We notice that the models with high accuracy may not be the best choice to gain people trust but to explain the models well. Also, different explainers interpret different features, which may decrease our trust in fake news detection models. This requires us to design more interpretable explainers that can do a deeper analysis for fake news detection models and provide a more understandable explanation to human to gain trust.

Acknowledgment

This work was supported in part by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939725.

References

[1] K. Gallagher, "The social media demographics report: Differences in age, gender, and income at the top platforms," *Business Insider*, 2017.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[3] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett *et al.*, "Fake news vs satire: A dataset and analysis," in *Proceedings of the 10th ACM Conference on Web Science*, 2018, pp. 17–21.

[4] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[5] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 395–405.

[6] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence." in *IJCAI*, vol. 18, 2018, pp. 3834–3840.

[7] H. Karimi and J. Tang, "Learning hierarchical discourse-level structure for fake news detection," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3432–3442.

[8] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[9] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 312–320.

[10] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, 2018, pp. 637–645.

[11] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 324–332.

[12] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.

[13] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 943–951.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[16] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[17] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 772–10 781.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[19] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[20] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.