|  |
| --- |
| **Analysis of the HEDGE Model on Text Classification and its Potential for Applications in Fake News Explainability** |
|  |
| **Mazal Bethany** |
| Department of Information Systems and Cyber Security |
| University of Texas at San Antonio |
|  |
|  |

Abstract

Deep neural network classifiers are often seen as black box models. Methods that can improve explainability in AI are sought after and should be examined for their suitability in different domains. To make it known why our models are making certain classification decisions is important for building trust in deep learning models. In this work, we analyze the generations of the HEDGE model for hierarchical explanations on BERT text classifiers on the IMDB dataset for sentiment analysis and the LIAR dataset for fake news discrimination. HEDGE on the IMDB dataset had been demonstrated to be effective in the paper that presented HEDGE, but this work extends the analysis to find common errors between the IMDB and LIAR datasets and examine the suitability of HEDGE in fake news explainability. The analysis shows that there are common errors in the HEDGE explanations for the two classification tasks and that HEDGE in its current state may not be suitable for accurately explaining fake news classification decisions.

Introduction

As of 2019, 72% of U.S. adults use at least one social media platform with this proportion continuing to trend upwards. Of those who use social media, the majority use it daily ([Pew Research Center, 2019](#Pew_2019)) Social media enhances their user’s ability to connect with one another and engage with news and entertainment. In recent years, the spread of fake news has increasingly become a focus of discussion on news and social media outlets. Social media is the primary source of news for nearly one in five Americans. Those who use social media as their primary source of news are more likely to get wrong information about coronavirus and politics and more likely to hear unproven claims ([Pew Research Center, 2020](#Pew_2020))

Left unchecked, these claims can spread quickly. In the month following the 2016 Presidential Election, there were over one million tweets related to the “Pizzagate” conspiracy theory ([Douglas, 2016](#Douglas_2016)). The spread of the fake news on social media can lead people to act on misinformation in real life. Due to the “Pizzagate” conspiracy theory, a man fired shots in the restaurant associated with the conspiracy theory as a result of his beliefs. More recently, the siege on the Capitol Building following the spread of fake news about the results of the 2020 Presidential Election further demonstrates the need to curb the spread of fake news.

Many of the major social media sites have some form of rules intended to prohibit or limit the spread of false information and fake news on their platform. To moderate content being posted on their platforms, some of the largest sites, like Facebook, have thousands of employees tasked with the review of posts that have been flagged by AI or reported by users ([Barrett, 2020](#Barrett_2020)). However, these moderators frequently make mistakes. Facebook CEO Mark Zuckerberg admitted that their moderators “make the wrong call in more than one out of every 10 cases” ([Koetsier, 2020](#Koetsier_2020)). Therefore, it is clear that the decision-making process involved with content moderation is in need of improvement.

One way the moderation process could be improved would be through the implementation of AI powered explanations about why a particular post has been flagged. This could help the moderators more accurately assess content by giving them an additional factor by which they can interpret and review a post. Furthermore, the use of explanations in content moderation could be used to give posters more specific information about why one of their posts has been removed. This could help to reduce the sentiment among the users whose posts are removed that the content moderation is not arbitrary.

One way to go about creating explanations for text classification would be to use hierarchical explanations using feature interaction detection. HEDGE (Hierarchical Explanation via Divisive Generation) is a model introduced in the paper Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. HEDGE helps to explain the decisions of text classification decisions [(Chen, et al., 2020).](#Chen_2020) The HEDGE model constructs hierarchical explanations for binary text classification based on feature interaction detection. HEDGE takes in a trained text classification model’s weights and breaks down the text input into separate phrases and words one step at a time. The text is broken down into phrases and words in order of the weakest interactions it finds within the text segment.

Chart, bar chart

Description automatically generated

Figure 1: HEDGE for BERT on a sentence from the IMDB dataset used in the HEDGE paper. The selection reads: “read the book, forget the movie!”. The sample is broken down in each level until only the individual tokens remain. The values of darker red values indicate words or phrases that contribute to a negative classification and the darker blue values indicate words or phrases that contribute to a positive classification.

In this research we will analyze the HEDGE model by looking for error patterns in its hierarchical explanations so that we can better understand its interpretability and its potential for applications in fake news detection. HEDGE is model-agnostic, so while it was originally presented in the use of explaining binary sentiment classifications, it should also be applicable in the explanation of other binary language classification tasks, such as fake news detection after it is given model weights for fake news classification.

Previous works about explainable fake news detection on social media content do not utilize hierarchical explanations. Many of the works on the explainability of the classification of social media content use the weights of words in a post and display the words in a post that had the greatest impact in the classification decision. While this does help identify individual words that are important to the classification decision, it does not provide enough information to create a phrase-level understanding of why such decisions are made. By incorporating hierarchical explanations into a classification model, we should be able to come closer to automating explanations for fake news detection.

Literature Review

Existing literature on the topic of explainable fake news detection is limited and does not make use of hierarchical explanations on classification decisions for fake news detection within the posts themselves. Furthermore, explainable fake news detection often lacks interpretability. Interpretability is important because it would help content moderators perform their job functions more accurately and would help social media users understand why their post would have been marked as fake news.

dEFEND: A System for Explainable Fake News Detection, approaches fake news detection by incorporating the analysis of the post itself with the comments associated with it ([Cui, et al., 2019](#Cui_2019)). The dEFEND paper explains that accounting for comments posted alongside the original post improves the rate of successfully detecting fake news. The explanatory part of that project is encapsulated by returning scores that rate how likely a certain post has non-factual claims. While this project showed that there can be successful explanatory features in fake news detection, explanations of this kind lack the level of interpretability that would be beneficial for social media users and moderators.

The paper Explainable Machine Learning for Fake News Detection takes a similar approach to their explainability ([Reis, et al., 2019](#Reis_2019)). Their models considered 172 features, and from the generated models they selected their top performing models. From the top performing models, they determined the most significant features for predicting if a story contained fake news. The explainability of this model is brought about by selecting the features that had the greatest impact on the classification decision and displaying how much of an impact they had. Among their top 10% of performing models, they were able to achieve AUC values between 0.855 and 0.885. Even though their fake news classification models are relatively successful, the interpretability of their explanations leave room for improvement.

A task similar to the detection of fake news is the detection of toxic language. In the paper Explainable AI approach towards Toxic Comment Classification ([Mahajan, et al., 2020](#Mahajan_2020)), the weights of words in their model were used to provide an explanation for the classification of toxic posts. This project was able to output the specific words in a post that had the greatest influence in the classification decision. These results are highly interpretable because its users would be able to see the words that the model considered to be toxic. Despite the model’s good interpretability, it fails to consistently classify posts with low word counts.

There has also been work on the causal spread of fake news. The paper Towards Causal Understanding of Fake News Dissemination ([Cheng, et al., 2020](#Cheng_2020)) studies the users who spread fake news and attempts to determine the causal effect of an account’s attributes on the probability that a user disseminates fake news.

Method

The paper that introduced HEDGE demonstrated HEDGE’s effectiveness on three text classifiers: LSTM, CNN, and BERT. Their results showed that, while LSTM and CNN classifiers produced more interpretable results, BERT was the more accurate classifier. Due to the higher ranking of importance of an accurate classifier compared to interpretable results, this paper will focus on the usage of the BERT classifier. Having an accurate classification takes priority over interpretability if the end goal is to improve content moderation. Further, BERT models have been shown to work well in fake news classification tasks ([Jin, et al., 2020](#JIn_2020)). Therefore, analysis in this paper relies on the assumption that BERT is also the best model for fake news detection.

To document observations on the generations from the original HEDGE model for sentiment analysis, 100 hierarchical explanations were generated from the test set of the IMDB dataset ([Maas, et al., 2011](#Maas_2011)). To document observations of HEDGE generations for a fake news classifier, the 100 hierarchical explanations were generated from the test set of the LIAR dataset ([Wang, 2017](#Wang_2017)). The total number of tokens, the level which provided the “best” explanation, and notes about errors in the hierarchical explanation were recorded for these 100 observations. To evaluate the quality of the explanation of a particular level, I consider whether the construction of the phrases make sense as stand-alone phrases, and whether the weights of each selected phrase make sense for their impact in the classification decision. Looking over the entire hierarchical explanation of each samples, I also take note of the order of the creation of the hierarchical explanations, and if the weakest feature interaction that this decision is based on makes sense.

Chart

Description automatically generated

Figure 2: HEDGE for BERT on a sentence from the IMDB dataset used in the HEDGE paper. The selection reads: “i hope this group of film-makers never re-unites.”

Based off of Figure 2, I will present an example of the annotations and the thought process behind the annotation that was used on the 100 hierarchical explanations that were generated from the IMDB dataset. This sample is made up of 13 tokens. I determined that level 3 provided the best stand-alone explanation for the classification decision. At this level, HEDGE shows that the phrase “hope this group of film-makers never re-unites” is what is causing the model to decide that this review is a negative review. Looking at the entire hierarchical explanation in Figure 2, there are some errors. Due to the tokenization algorithm, the model recognizes the word “unites” as two separate tokens (“unite” and “s”). The model detected this and made the decision that the weakest interaction in the entire sample was between “I hope this group of film-makers never re-unite” and “s.”. This does not make much sense and I would have expected a weaker interaction to have been detected elsewhere, perhaps between “I hope this” and the rest of the selection.

Results

The length of the 100 observations from both the IMDB and LIAR datasets ranged from 7 tokens to 72 tokens. The level that provided the best explanations tended to be between level 3 and level 8. The level that created the best overall explanation seemed to be highly related to how long the text sequence was. This makes sense because longer text sequences have a greater number of ideas that would need to be separated from one another in the hierarchical explanations. The paper that presented HEDGE had their annotators evaluate HEDGE generations up to level 5, with the reasoning being that the longer text sequences simply have too many features and can be difficult to work with. For the purpose of maximizing the interpretability of the HEDGE explanations in a fake news classification context, it may be most helpful to limit HEDGE generations to a certain level. Presenting a full HEDGE generation, rather than limiting the generation, to a social media moderator could lead to a problem of too much information, which would reduce the utility of HEDGE in assisting them in their tasks.

A common issue that I encountered in the HEDGE generations on the IMDB dataset was that the hierarchical breakdowns often failed to separate phrases by punctuation in the early levels. As punctuation often mark the beginning or end of ideas, it would make sense to see punctuation at the beginning or end of the first phrases that get separated. An example of this error in the hierarchical breakdown can be seen in Figure 1.

Another common issue that I found in the HEDGE generations on the IMDB dataset was that names, when given as a first and last name, would frequently get separated at early to middle levels. This problem could potentially be remedied by the incorporation of named entity recognition techniques and by recognizing names as a single token.

HEDGE generations on the IMDB dataset also show that words that are made up of multiple tokens appear to be likely to be split up in early layers. This reduces the interpretability of some of the explanations.

Chart

Description automatically generated with medium confidence

Figure 3: HEDGE for BERT on a sentence from the IMDB dataset used in the HEDGE paper. Here, BERT made a wrong classification decision. The selection reads: “this is a good film. this is very funny. yet after this film there were no good ernest films!”

5 of the 100 observations studied from the IMDB dataset made a prediction to the wrong class. Figure 3 shows the hierarchical explanation of one of these wrong predictions. This text was predicted by the BERT model to be a negative review rather than the correct label of a positive review. While the class prediction was wrong, the hierarchical explanation provides valuable insight into what went into that decision. At level 4 of the set of explanations, it is made clear that the phrase “were no good ernest” was one of the primary contributors to the negative review classification. At level 4, we can also see that HEDGE detects the phrase “is a good film. this is very funny. yet after this film there” as having a positive impact. From this example and it’s hierarchical explanations specifically, we can see that a potential reason why this review was misclassified is because this review contains negative descriptions of other films. While the BERT classification did not pick up on this detail, the HEDGE explanation generated from it would help those who need to make classification decisions, as it provides a structure to understand and identify why classification mistakes are made. HEDGE’s ability to help reveal mistakes in classification could potentially make it valuable in the space of fake news moderation, which could reduce the number of false positives when a moderator has access to these explanations.

35 of the 100 observations studied from the IMDB dataset made a prediction to the wrong class. HEDGE generations on the LIAR dataset share some similar errors with the HEDGE generations on the IMBD dataset. The generations on the LIAR dataset also failed to separate phrases by punctuation at the early levels. Further, first and last names also tended to be separated at early levels. Finally, the HEDGE generations also spilt up words that consist of multiple tokens in early levels.

Chart

Description automatically generated

Figure 4: HEDGE for BERT on a sentence from the LIAR dataset. Here, BERT correctly classified the statement as fake. The selection reads: “Bill White has a long history of trying to limit or even disenfranchise military voters.”

For HEDGE generations on the LIAR dataset, blue signifies a fake news classification, and red signifies a true classification. Figure 4 shows one such example. In Figure 4 we see some of the same issues with the first name, last name and punctuations being separated out at early levels. One of the potentially more useful traits in the HEDGE generations on the LIAR dataset is that they seem to separate out the claim being made at early levels. In this example, “long history of trying to limit or even disenfranchise military voters” is what they are trying to attribute to the topic, Bill White. This observation could potentially be useful for users of HEDGE generations for fake news because they may be able to quickly identify what the claim being made is. However, the weights attributed to each text sequence in the explanation may be less interpretable in the case of fake news discrimination, when compared to its use in sentiment analysis. In the example presented in Figure 4, BERT correctly classified the text as fake news. However, the claim of “long history of trying to limit or even disenfranchise military voters” contributed to a true classification, as seen in colors of the hierarchical explanation. Even though this claim should be seen as what is false in the given text, HEDGE is showing that the BERT classifier sees the claim as being true. This reduces the utility of HEDGE explanation for fake news discrimination. However, this could just suggest that the overall BERT classifier is primarily picking up on stylistic traits when making its classification decision over the entire text sequence. Even so, it is an overall reduction in interpretability.

Graphical user interface, chart, application, bar chart

Description automatically generated

Figure 5: HEDGE for BERT on a sentence from the LIAR dataset. Here, BERT incorrectly classified the statement as true. The selection reads: “State revenue projections have missed the mark month after month.”

Figure 5 is an example of an incorrect classification of the BERT classifier on the LIAR dataset. In this example, the BERT classifier incorrectly classified the text as true. Much like the previous example, the topic of the claim, “State” gets separated at an early level. The claim “missed the mark month after month” also gets separated at an early level. As explained earlier, the separation of the claim at early levels could be useful in fake news explainability. In this case, the claim pushed the classification towards a fake news classification even though the overall prediction of the text input was that the selection was true. Across the 100 observations on the LIAR dataset, it seemed that there was no clear relationship between the fake news vs true classification weights given to the claim that was separated out, and the overall classification of the text.

There are some additional observations made on the HEDGE generations on the LIAR dataset that do not necessarily have relevance to whether the hierarchical generations were made in error or not. The first of which is that the mentioning of names tends to push the model to a fake news classification. This is not necessarily a mistake on the part of the model. The model must assign a weight to each word. Perhaps this means that the use of names is more likely in fake news. Another observation relates to the sentiment of fake news. It has been well studied that fake news tends to have a greater probability of negative sentiment ([Zaeem, et al., 2020](#Zaeem_2020)). However, the HEDGE generations on the LIAR dataset seem to the point to an opposite conclusion. Negative phrases generally seem to push towards a true classification in fake news discrimination in our model.

Another observation made on the HEDGE generations is that they can be computationally expensive to create when given long text inputs. There is an exponential increase in number of possible ways to split up the text segment when adding more tokens to a sequence. Generating the explanations on a system with 64 GB in RAM approached 10 minutes for a sequence of 72 tokens. This further suggests that HEDGE may be best suited for shorter texts.

Limitations and Future Work

One of the biggest challenges with using HEDGE is that it was only designed for univariate models for binary classification. While this may be sufficient for high accuracy in tasks like sentiment analysis, it severely underperforms in tasks like fake news classification, where state of the art models utilize many factors to make their classification. Further, even in most fake news discriminators, ground truth is not utilized, which leads to potential biases in its implementation. The fake news classifier primarily detects stylistic features and ignores network structures in social media interactions and ground truth, which could greatly improve accuracy and interpretability. One possible way to continue to use HEDGE generations in fake news explanations could be to condition the hierarchical explanations on the classification decision of a more accurate, multivariate classifier. Furthermore, HEDGE generations could have improved interpretability if they had less awkward splits in phrases. Incorporating constituency parsing into the HEDGE algorithm to place greater importance on the splits made from a parse tree could help mitigate such issues.

Conclusion

In this paper, we identified a few common errors in the HEDGE explanations on IMDB sentiment classifications. We found that names, punctuations, and words made up of multiple tokens tended to be features that the HEDGE model performed poorly on. We also evaluated HEDGE on the LIAR dataset for fake news discrimination. On the LIAR dataset, the hierarchical explanations had errors similar to the errors on the errors made on the IMDB dataset. The explanations generated on the LIAR dataset also appeared to separate out the claim being made in early layers, which could have utility for its users. However, inaccurate classifications and the lack of the use of ground truth make it hard to justify using HEDGE for fake news explanations in it’s current state.

# **References**

Barrett, P. (2020). *Who Moderates the Social Media Giants?* Center for Business.

Chen, H., Zheng, G., & Ji, Y. (2020). Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Cheng, L., Guo, R., Shu, K., & Liu, H. (2020). Towards Causal Understanding of Fake News Dissemination. *arXiv preprint arXiv:2010.10580*.

Cui, L., Shu, K., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: A System for Explainable Fake News Detection. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2961-2964.

Douglas, W., & Washburn, M. (2016, December 6). *Religious zeal drives N.C. man in ‘Pizzagate’*. Retrieved from The Courier Tribune: https://www.courier-tribune.com/news/20161206/religious-zeal-drives-nc-man-in-8216pizzagate8217

Jin, D., Jin, Z., Tianyi Zhou, J., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 05, pp. 8018-8025.

Koetsier, J. (2020, June 9). *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*. Retrieved from Forbes: https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=67e517d554d0

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142-150.

Mahajan, A., Shah, D., & Jafar, G. (2020). Explainable AI approach towards Toxic Comment Classification. *EasyChair Preprint*.

Pew Research Center. (2019, June 12). *Social Media Fact Sheet*. Retrieved from Pew Research Center: https://www.pewresearch.org/internet/fact-sheet/social-media/

Pew Research Center. (2020, July 30). *Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable*. Retrieved from Pew Research Center Journalism and Media: https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/

Reis, J., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable Machine Learning for Fake News Detection. *Proceedings of the 10th ACM conference on web science*, 17-26.

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422-426.

Zaeem, R. N., Li, C., & Barber, S. (2020). On Sentiment of Online Fake News. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 760-767.