Interpretable Fake News Detection With Attention Attribution as Evidence: Promises and Perils

ThanhVi Dang

Department of Computer and Information Science (CIS)

Major in Data Science

The Graduate School of Art and Science

Supervised by
Ruhul Amin, Ph.D.

Fordham University
New York, New York

January 2024

Abstract

In an era marked by the proliferation of social media platforms and internet usage, the pervasive issue of fake news and misinformation dissemination poses a significant threat to society. Current artificial intelligence (AI) and language understanding models often struggle to discern nuanced contextual and cultural elements in online communication, hindering the identification of fake news. This study experiments with a novel approach using self-attention attribution (ATTATTR) to enhance interpretability in fake news detection. ATTATTR assigns attribution scores to input tokens based on their contributions to attention heads within the Transformer model, offering insights into the underlying decision-making process. The research spans four diverse datasets, exploring the promises and perils of attention attribution for developing an interpretable fake news detection system. The paper addresses the lack of interpretability in existing models, particularly within the context of the Fake News dataset, and aims to assess the effectiveness of attention attribution in revealing features and patterns influencing classification decisions from the perspective of Fake News detection. The study emphasizes the need for methodological refinement or exploration of alternative approaches to yield satisfactory outcomes. Importantly, attention attribution, while exhibiting promising potential, is not a panacea, and for future work, collaboration with human annotation is suggested for achieving comprehensive interpretability.

Acknowledgments

As I reflect upon the journey of completing my Master's thesis, I want to express my sincere gratitude and appreciation to the people who have played pivotal roles in this academic endeavor - my family, my thesis advisor, my thesis committee, and my friends in and out of the program. This journey has been a challenging yet rewarding experience, and I am thankful for the support and guidance I have received along the way.

First and foremost, I extend my deepest appreciation to my thesis advisor, Dr. Ruhul Amin, for his unwavering support, invaluable insights, and continuous encouragement throughout the research process. Thank you for sharing your expertise and dedication - it played a pivotal role in shaping my thesis into a comprehensive piece of scholarly work. Thank you for everything that you have done for me throughout my journey to obtaining this Master’s degree.

I am also grateful to the members of my thesis committee, Dr. Zakirul Alam Bhuiyan and Dr. Mohamed Rahouti, for their constructive feedback and valuable suggestions that greatly enhanced the quality of this research.

Finally, I would like to express my deepest love to my family for their unwavering support, understanding, and encouragement throughout this academic endeavor. Their love and encouragement provided the emotional sustenance I needed to navigate through the challenges of graduate studies. Apart from my family, I am thankful to my friends in and out of the program, situated in various corners of the world, whose invaluable support has meant a lot to me.

This thesis stands as a collective effort, and I am thankful to all those who have played a part in its realization.

Table of Contents

Chapter 1: Introduction

In a period that is dominated by social media platforms and internet usage, the persistent issue of fake news and misinformation dissemination poses a substantial threat. The need for effective tools to identify and counteract such misleading content is paramount. Fourth-generation technologies, particularly artificial intelligence (AI), offer potential solutions to this challenge. However, current AI and language understanding models often struggle to discern the nuanced contextual and cultural elements of online communication, particularly in the context of identifying fake news on social media and online news outlets. This study introduces a novel approach by leveraging self-attention attribution (ATTATTR) as a promising method to enhance interpretability in fake news detection. ATTATTR assigns attribution scores to input tokens based on their contributions to attention heads within the Transformer model. These scores unveil which tokens the model prioritizes in its decision-making process, offering insights into the underlying mechanisms. By leveraging attribution tree construction for visualization and adversarial attacks on trigger words, our experiment seeks to provide a solution to the black-box natural language processing models problem. This exploration spans four diverse datasets, to test the efficacy of the existing model and to capture the varied contextual nuances inherent in the domains of fake news and misinformation. The research aims to contribute to the development of an interpretable fake news detection system using self-attention attribution, examining promises and perils across these datasets.

1.1. Problem Statement

In the face of the growing threat of fake news, existing machine learning models designed for detection often lack interpretability, impeding accountability and trust. Our research addresses this gap by examining how attention attribution compares to previous models in enhancing interpretability, particularly within the context of the Fake News dataset. We aim to assess the effectiveness of attention attribution in revealing the features and patterns influencing classification decisions. This investigation into the promises and perils of attention attribution contributes to the development of more transparent and effective fake news detection models, crucial for combating misinformation in the digital age. We also seek to use the extracted trigger words as evidence for fake news detection.

1.2. Research Contribution and Expected Impact

The proposed research on "Interpretable Fake News Detection With Attention Attribution as Evidence: Promises and Perils" aims to make significant contributions to the field of fake news detection and interpretation. By leveraging attention attribution interpretability techniques, our study seeks to provide a deeper understanding of the Transformer

mechanisms from the perspective of fake news datasets, unraveling the intricate decision-making processes employed by these algorithms. The anticipated impact of this research extends beyond the realm of improved model performance, as it holds the promise of enhancing transparency and interpretability in the fight against misinformation. Furthermore, the insights gained from our investigation into the promises and perils of attention attribution as evidence can inform the development of more robust and ethical fake news detection systems, fostering a more trustworthy information ecosystem. As the proliferation of fake news continues to pose serious challenges to public discourse and societal well-being, our research strives to pave the way for more effective and accountable solutions, thereby contributing to the broader goal of promoting information integrity in the digital age.

Chapter 2: Literature Review

2.1. Natural Language Processing

In the era of advancing technology and pervasive social media, the significance of enabling computers to comprehend human language has surged. Natural Language Processing (NLP) has seamlessly integrated into a spectrum of computer programs, spanning from e-commerce platforms to semantic analysis for deciphering customer feedback. Beyond traditional domains, NLP plays a crucial role in diverse fields. As technology continues to evolve, the pervasive influence of NLP underscores its pivotal role in bridging the gap between human communication and computational systems across various domains.

2.1.1. Transformer and its Architecture

The Transformer architecture, introduced by Vaswani et al. in 2017, revolutionized natural language processing and machine translation. It employs a self-attention mechanism to process input sequences in parallel, allowing it to capture long-range dependencies. The model consists of an encoder and a decoder (Figure 1.1), each comprising multiple layers. Within each layer, self-attention mechanisms enable the model to weigh different parts of the input sequence differently, facilitating effective information capture. Additionally, position-wise feedforward networks and layer normalization contribute to the model's ability to capture complex patterns. The Transformer's attention mechanism and parallel processing capabilities have since become fundamental components of various state-of-the-art models in natural language processing and beyond.

As mentioned, the Transformer model exclusively depends on the attention mechanism (Vaswani et al., 2017) for computing representations of both input and output. This characteristic inherently renders it a captivating subject for research. Additionally, numerous novel models, such as the Universal Transformers, OpenAI's GPT-3, and more recently, their Sparse Transformer (Amato et al., 2019), have been introduced, underscoring its significance as a sturdy neural foundation for intricate models.

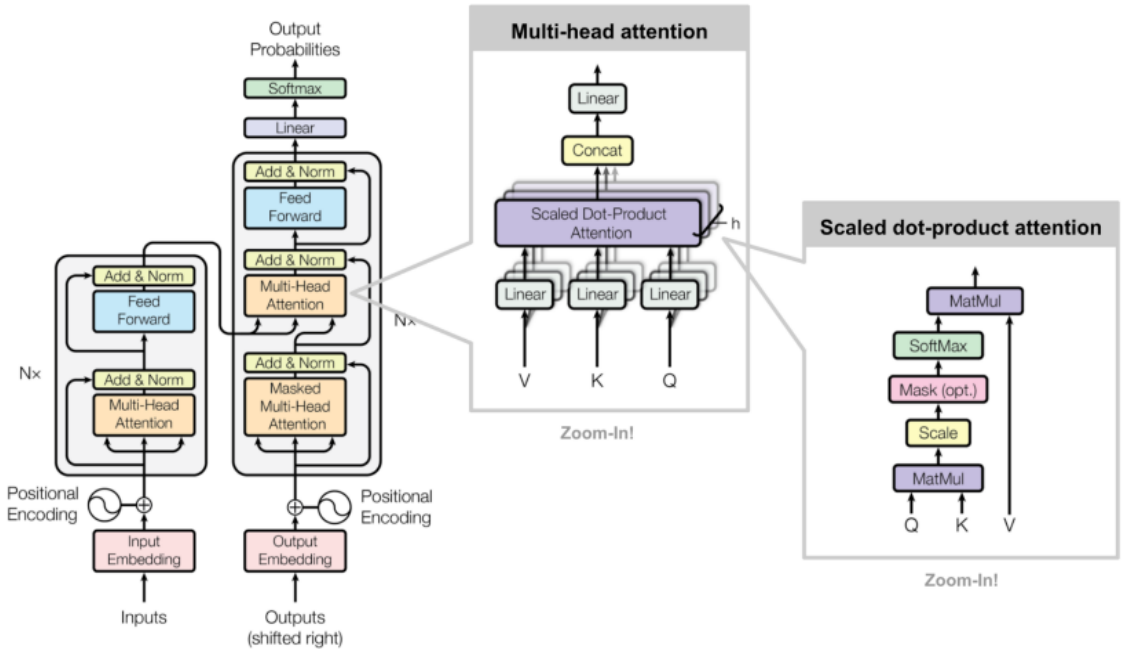

Figure 1.1. The Transformer Architecture paper "Attention is All You Need" by Vaswani et al. (2017) (source)

The Transformer model adheres to an encoder-decoder architecture, widely recognized for tasks involving sequence transduction, such as neural machine translation. Both the encoder and decoder consist of N identical layers. In the encoder, each layer includes a self-multi-head attention block, where queries (Q), keys (K), and values (V) are identical, followed by a feed-forward block, with both blocks featuring residual connections. The decoder's layers comprise a self-multi-head attention block, a multi-head attention block connecting to the encoder stack output (queries from the decoder, keys, and values from the encoder), and a subsequent feed-forward block. Residual connections augment all blocks. A final *softmax* classifier generates probabilities for the next token within the output vocabulary set. The model encompasses 65 million learnable parameters.

2.1.2. Self-Attention Mechanism

The concept of attention is extensively examined in neuroscience, psychology, cognitive science, and machine learning (Chun et al., 2011, Cho et al., 2015). It entails the selective focus on specific information while ignoring other perceivable data, with the acknowledged benefit of optimizing computational resources. Advances in understanding attention have given rise to various models aimed at clarifying its underlying mechanisms (Vaishnav, 2023)

In a self-attention layer, the input vector undergoes a learnable matrix transformation to yield K, Q, and V vectors, each with a dimension of 512 ($d_q = d_k = d_v = 512$) (Vaswani et al., 2017). Initially, a score (S) is computed to determine the attention focus on other words in a sequence while encoding the present word. This score is derived from the dot product of the query and key vectors ($S = Q \,.\, K^T$). To ensure stable gradients, the score is normalized ($S = S / \sqrt{d_k}$) and subsequently transformed into probabilities through a *softmax* function. The probability score indicates the relevance of the current word to others in the sequence. The computed score is then multiplied by the value vector (V), resulting in increased attention to relevant words and the neglect of irrelevant words in the subsequent layers. Dot-product attention can be written as:

$$Attention(Q,K,V)=Softmax(QK^T)V$$

The paper introduces scaled dot-product attention, wherein the dot product is divided by the dimensionality of the keys. However, a single instance of scaled dot-product attention proves inadequate for capturing diverse dependencies between distant positions in the sequence. To address this limitation, the paper advocates for multi-head attention, where attention is the concatenation of 8 parallel dot-product attention heads. This enables the model to collectively attend to information from various representation subspaces at different positions in the sequence.

Layer Type	Complexity per Layer	Sequential operations	Maximum path length
Self-Attention	$\mathcal{O}(n^2.d)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
Recurrent	$\mathcal{O}(n.d^2)$	$\mathcal{O}$(n)	$\mathcal{O}$(n)
Convolutional	$\mathcal{O}(k.n.d^2)$	$\mathcal{O}(1)$	$\mathcal{O}(log_k(n))$

Table 2.1. Comparison of different network complexity (Vaswani et al., 2017)

In terms of computational complexity, self-attention layers are generally more efficient than recursive or convolutional layers when dealing with sequences of length n and

dimensionality *d*, especially when *n* is smaller than *d,* which is often the typical scenario. The efficiency stems from the parallelization inherent in self-attention, allowing for faster processing of sequential data compared to the more sequential nature of recursive or convolutional operations.

2.1.3. Multi-Head Attention Mechanism

The self-attention mechanism presented by Vaswani et al. [2017] incorporates a key feature known as multi-head attention, contributing to enhanced performance in two ways. Firstly, it improves the network's capacity to concentrate on multiple positions in the sequence. Secondly, it assigns distinct representational subspaces to each word. For instance, when employing eight heads, eight sets of matrices for keys (K), queries (Q), and values (V) are created, each corresponding to a unique representational subspace. These sets are concatenated before being passed through the feedforward network.

The fundamental aspect of NLP tasks is the sequential order of words. Nevertheless, the operations discussed so far are permutation invariant. To address this, positional embedding vectors are introduced, and they are added to each input embedding vector. These vectors play a crucial role in enabling the model to discern the position of each word within a sequence in the projection space of K, Q, and V.

2.1.4. BERT

Bidirectional Encoder Representations from Transformers, otherwise known as BERT, is a breakthrough in natural language processing. It has revolutionized the way machines understand and generate human language. Developed by Google researchers in October 2018 (Devlin et al., 2018), BERT stands out for its *bidirectional* approach, enabling it to grasp the meaning of words in context more effectively than previous models. Unlike traditional unidirectional models, BERT considers both the left and right context of each word by employing the Masked Language Model (MLM). In this process, certain tokens in the input sequence are strategically masked, prompting the model to predict the missing tokens based on the surrounding context. This bidirectional understanding has proven invaluable in capturing intricate linguistic nuances, making BERT a pivotal tool in various NLP applications, such as sentiment analysis, question answering, and language translation (Kula et al., 2021; Jwa et al., 2019; Kula et al., 2019).

Furthermore, BERT's success lies in its pre-training and fine-tuning paradigm (Devlin et al., 2018). During pre-training, the model learns from vast amounts of unlabeled text, gaining a general understanding of language structures and patterns. This process helps BERT learn general language representations by predicting masked tokens within sentences and identifying the relationships between different sentences in a document. Once pre-training is complete, BERT is fine-tuned on specific downstream tasks, such as

question answering or named entity recognition. During fine-tuning, BERT is combined with task-specific layers on top of its pre-trained language representation layers. By fine-tuning task-specific datasets, BERT can adapt to different language-understanding tasks and achieve state-of-the-art performance on various benchmarks.

Overall, BERT's innovation lies in its bidirectional training approach and fine-tuning strategy, which enables it to capture rich contextual information and yield strong results on a wide range of natural language processing tasks. The versatility of BERT has led to its widespread adoption in diverse industries, ranging from healthcare and finance to e-commerce and customer support. Researchers and practitioners alike continue to explore and refine BERT-based architectures, extending their capabilities and pushing the boundaries of natural language understanding.

2.2. Explainable Machine Learning/Artificial Intelligence

Owing to the escalating influence of artificial intelligence (AI) and machine learning (ML) models in contemporary society, there has been a discernible surge in apprehensions concerning its reliability. The crux of these concerns lies in the prevalent deployment of 'black-box' models, notably Deep Neural Networks (DNNs), within AI solutions. These models, characterized by their expansive parameter space and algorithmic intricacies, present a formidable challenge for human interpretation due to their inherent lack of transparency. The decision-making processes of such models remain largely opaque to human understanding, thereby elevating the likelihood of biases rooted in unjust, outdated, or erroneous assumptions, which may evade detection through conventional model evaluation methodologies. Consequently, confidence in these opaquely functioning models is substantially compromised.

In response to the inherent limitations associated with conventional AI, there is a discernible imperative for the adoption of explainable artificial intelligence (XAI) methods. As articulated by BarredoArrieta et al. (2020), XAI endeavors to engender machine-learning techniques that yield models of exceptional efficacy while concurrently being interpretable, allowing for comprehension, management, and trust by human stakeholders. Fundamentally, xAI seeks to imbue models with a capacity to articulate their functioning when presented to an audience, furnishing details and explanations to enhance clarity (Choraś et al., 2020; Das and Rad, 2020). Through the integration of xAI principles, models can be imbued with heightened security, reduced susceptibility to errors, and, ultimately, a higher degree of trustworthiness.

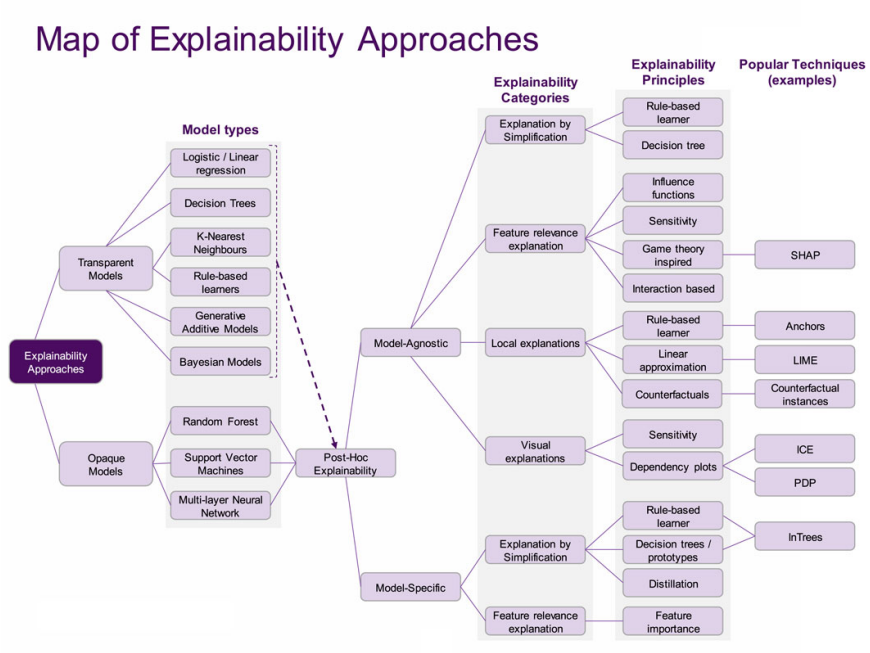

Figure 2.2. Map of Explainability Approaches by Bell and Papantonis (2021)

In the paper titled "Principles and Practice of Explainable Machine Learning," the authors present a comprehensive taxonomy framework that systematically categorizes models based on their degree of explainability (Bell and Papantonis, 2021). The taxonomy distinguishes between two overarching categories: transparent or white-box models and opaque or black-box models as seen in Figure 2.2. Further granularity is achieved by subcategorizing model types within these groups. Transparent models, including regression, decision trees, k-nearest neighbors, etc, are characterized by their interpretability, grounded in the statistical nature of their algorithms. In contrast, black-box models predominantly provide explanations and interpretations post hoc. Model-agnostic techniques, such as SHAP, Lime, and Anchors, exhibit a lack of dependency on specific machine learning models or architectures, offering greater generalizability and versatility. Conversely, model-specific approaches are tailored for particular machine learning models, leveraging their unique characteristics or structures. The subsequent discussion will delve into a more detailed exploration of SHAP, Lime, and Anchors in the upcoming segments.

2.2.1. LIME

Undoubtedly, one of the preeminent model-agnostic approaches to explainability is exemplified by the Local Interpretable Model-Agnostic Explanations (LIME) technique (Ribeiro et al., 2016). Methodologies categorized as model-agnostic are expressly designed with broad applicability as a guiding principle, emphasizing a requisite degree of flexibility to operate independently of the intrinsic architecture of a given model. These approaches are principally oriented toward establishing coherent relationships

between a model's input variables and their corresponding outputs. Prominent subtypes within this classification encompass model simplification, feature relevance, and visualizations.

In the specific case of LIME, the approach entails a localized approximation of an opaque model in the proximity of the targeted prediction. This involves the construction of either a linear model or a decision tree, centered around the specific prediction of interest. Subsequently, the derived model serves as a surrogate, facilitating a more intelligible explanation of the underlying complexities inherent in the opaque model. Additionally, the LIME methodology necessitates a transformation of the input data into an "interpretable representation," ensuring that the resultant features are comprehensible to human stakeholders, irrespective of the model's original feature set Madsen, Reddy, and Chandar, 2023). This transformative process is referred to as the "intermediate transformation" (Madsen, Reddy, and Chandar, 2023).

LIME offers distinct advantages, foremost among them being its reliance solely on black-box information and the dataset, obviating the need for gradient calculations (Miller, 2019). Additionally, it employs LASSO logistic regression, a variant of standard logistic regression incorporating an $L1$-regularizer. This renders its explanations selective and sparse, a quality that can be paramount in delivering human-friendly interpretations.

2.2.2. Anchors

A cognate methodology, denominated "anchors," is expounded upon in the study conducted by Ribeiro et al. (2018), the authors as the LIME (Ribeiro et al., 2016) method. In this approach, the overarching objective persists in locally approximating a model, albeit diverging from LIME by eschewing the application of a linear model. Instead, the method incorporates straightforward and interpretable "if-then" rules strategically devised to anchor the model's decision (Ribeiro et al., 2018). These rules are meticulously crafted to encapsulate pivotal features while judiciously excluding superfluous details, thereby engendering more parsimonious and succinct explanations. The anchors method is supposed to ensure that the prediction is always supposed to be the same, for the instances on which the anchor holds (Ribeiro et al., 2018).

Anchors, as an elucidative methodology, demonstrate a spectrum of commendable attributes. Its pronounced strength lies in interpretability, yielding lucid "if-then" rules that facilitate a nuanced grasp of the underlying decision-making processes. Positioned within the framework of local approximation, anchors adeptly concentrate on specific predictions, offering granular insights. The method's commitment to simplicity and conciseness ensures the consideration of only pivotal features shaping the model's decisions, thereby engendering more focused and streamlined explanations. Anchors'

reliance on a rule-based representation augments transparency, furnishing stakeholders with clear insights into the determinants of the decisional outcome. The technique's dedication to sparse explanations refines its interpretive efficacy, and its independence from gradient calculations enhances suitability for scenarios where computational complexities are formidable (Ribeiro et al., 2018). Moreover, as a model-agnostic paradigm, anchors exhibit versatility, rendering them applicable across diverse machine-learning models, thereby reinforcing their efficacy in cultivating interpretability, transparency, and trust within intricate decision systems.

2.2.3. SHAP

Within the domain of explainable artificial intelligence (XAI), Lundberg and Lee's (2017) seminal work on Shapley Additive exPlanations (SHAP) emerges as a notable contribution. SHAP endeavors to elucidate model predictions by constructing a local linear model around a specific instance, interpreting the resultant coefficients as indicators of feature importance. Distinguished from LIME (Ribeiro et al., 2016), SHAP is underpinned by a robust theoretical foundation rooted in coalitional Game Theory, specifically leveraging Shapley values (Shapley, 1952). These values encapsulate the average expected marginal contribution of a feature to the model's decision. However, challenges in real-world applications, characterized by high dimensionality, necessitate assumptions such as variable independence, introducing complexities in the computation of Shapley values. Furthermore, the impracticality of considering all conceivable variable orderings further compounds computational challenges in deriving Shapley values.

Concurrently, Strumbelj and Kononenko (2010) propose an alternative approach to measure feature importance using Shapley values, deviating from SHAP in terms of both the objective function and optimization strategy. Datta et al. (2016) introduce a comprehensive suite of measures, including Quantitative Input Influence (QII), to address the impact of correlated inputs on model outputs. These diverse methodologies underscore the versatility of Shapley values in quantifying the influence of variables on model outcomes.

2.3. Fake News Detection

“Fake News” is a term that has been increasingly popular as social media and online news sources rise (Wu et al., 2019; Granillo, 2020; Cerf VG, 2016). The nomenclature "fake news,” gaining its attention during the 2016 U.S. Election (Admed et al., 2017) has undergone a semantic transformation due to its prolific use, resulting in an intricate and nuanced concept (Lazer et al., 2018; Liu et al., 2014, Muhammed and Matthew, 2022). Presently, the term serves a dual purpose: (a) functioning as a pejorative designation to condemn media and journalistic practices, and (b) serving as a comprehensive umbrella

term encapsulating various manifestations of inaccurate, misguided, or fabricated information. In the latter interpretation, fake news encompasses a spectrum ranging from inadvertent misinformation, exemplified by instances of careless reporting, to deliberate disinformation, including propagandistic endeavors (Quandt et al., 2019). The concerns over this problem are global. However, much is unknown about how to reduce, eliminate, and detect “Fake News” to provide a safer and more accurate internet environment.

In response to the imperative of mitigating the impact and proliferation of misinformation, a multifaceted range of countermeasures has been developed. Among these, linguistic-based strategies predominate, frequently leveraging sophisticated technologies such as NLP and Deep Learning (DL). However, in light of recent strides in the field of AI, it is increasingly evident that achieving optimal model performance is no longer a singularly sufficient criterion. Therefore, previous research has attempted to find scientific answers to the system’s decision by exploring its explainability through XAI techniques such as Lime, SHAP, or Anchors (Ribeiro et al., 2016; Ribeiro et al., 2018; Lundberg and Lee, 2017). These techniques are used to evaluate fake news data on fake news detector models to ensure the quality and transparency of the deployed and published architectures.

The escalating concern over potential threats posed by fake news has prompted the development of various countermeasures, including those proposed and implemented by social media platforms themselves (Quandt et al., 2019; Tandoc et al., 2018; Ciampaglia, 2018). The landscape of fake news detection tools and methodologies can be broadly categorized into two main groups: network-based and linguistic-based (Zhang et al., 2019; Conroy et al., 2015). Hybrid approaches, amalgamating elements from both categories, are also prevalent.

Network-based approaches gauge the veracity of news by evaluating source credibility, and leveraging network properties such as authors, timestamps, or included links. These approaches focus on either heterogeneous or homogeneous networks. Heterogeneous networks encompass diverse node types, while homogeneous ones consist of a single type. An example of a network-based approach exploring homogeneous networks is presented by Zhou, X. & Zafarani, R. (2019), wherein news articles are represented as sociologically based patterns across different network levels. In contrast, linguistic-based methods center on scrutinizing the content of news articles to identify anomalies, assuming the existence of specific patterns unique to fake news. Statistical analysis, a subcategory of linguistic-based methods, focuses on assessing credibility through frequency anomalies. For example, the work of Ksieniewicz et al. (2019) employs a count vectorizer to obtain word occurrences, utilizing an ensemble of decision trees for classification.

The BERT model has also become instrumental in the realm of fake news detection. Numerous studies have leveraged BERT to enhance the accuracy of fake news identification. Jwa et al.'s exBERT model (2019) applied BERT for the first time in fake news detection, achieving commendable F1 scores by pre-training BERT with news-related data and fine-tuning with Linear and Softmax layers. Kula et al. (2021) proposed a hybrid architecture integrating BERT with a Recurrent Neural Network (RNN), showcasing comparable results for similar datasets. Additionally, Kaliyar et al.'s system combined three parallel blocks of single-layer Convolutional Neural Networks (CNNs) with BERT (2018), achieving a remarkable 98.90% accuracy on test data. These studies collectively underscore BERT's versatility and effectiveness in elevating the performance of fake news detection systems across various neural network architectures.

Chapter 3: Motivation

3.1. The Need for Explainability and Interpretability

The imperative for explainability and interpretability in machine learning models is underscored by their increasing prevalence and complexity. It is pivotal to establish a clear understanding of the definitions underpinning these terms. "Interpretability" refers to the model's capacity to elucidate and present its outcomes in a manner comprehensible to humans, prompting the query, "How is this result obtained?" Conversely, "Explainability" pertains to the ability to fathom and construe the decisions or outputs of a system, prompting the question, "Why is this result obtained?" (Madsen, Reddy, and Chandar, 2023).

Natural language models, exemplified by the Transformer architecture (Vaswani et al., 2017), predominantly operate as black-box models. The efficiency of these models corresponds with an escalation in complexity, thereby raising concerns regarding the responsibility associated with deploying such intricate systems. This gives rise to overarching questions concerning the safety, ethics, and accountability inherent in the use of advanced machine learning models. The imperative for achieving transparency and clarity in the decision-making processes of these models becomes paramount, particularly in contexts where human understanding and trust are pivotal considerations. The pursuit of responsible and ethical deployment of complex models necessitates a concerted effort to enhance their interpretability and explainability, fostering a harmonious integration of advanced technology within the broader socio-ethical landscape.

3.2. Contributing to Fake News Research

Identifying fraudulent news poses a formidable challenge, surpassing the intricacy associated with discerning deceitful product reviews. The expansive landscape of the internet and social media, coupled with recent advancements in computer technologies, facilitates the seamless generation and dissemination of deceptive news content. Distinguishing the intent and evaluating the impact of fraudulent reviews is comparatively straightforward; however, gauging the motives and ramifications of orchestrated propaganda through the dissemination of false news proves to be a considerably more intricate undertaking. The elusive nature of these intentions and impacts amplifies the complexity inherent in combatting the proliferation of misinformation.

In the context of fake news detection, the importance of Explainable Artificial Intelligence (XAI) becomes especially pronounced. The intricate nature of the phenomena surrounding fake news, often driven by nuanced linguistic subtleties and varying degrees of misinformation, necessitates a deeper understanding of model decisions. XAI serves as a critical tool in unraveling the decision-making processes of machine learning models, providing insights into why a particular news article is classified as fake or authentic. The transparency offered by XAI not only aids researchers and practitioners in comprehending model predictions but also fosters trust and accountability in the deployment of such models. As fake news continues to evolve dynamically, XAI empowers stakeholders to dissect and interpret the factors influencing model outputs, facilitating a more informed and responsible approach to mitigating the impact of deceptive information. By enhancing the interpretability of machine learning models, XAI becomes an invaluable asset in the ongoing efforts to fortify the accuracy and reliability of fake news detection systems.

3.3. Self-Attention Attribution in Fake News Detection Interpretability

Attention mechanisms have significantly elevated the efficacy of NLP tasks while preserving model interpretability. Despite the widespread adoption of self-attention, the challenge lies in achieving interpretability due to the multitude of attention distributions. Moreover, understanding the interpretability of self-attention in the specific context of fake news detection remains an open question. Current deep learning-based approaches for fake news detection lack interpretability and neglect the incorporation of external knowledge relevant to news articles. In the pursuit of establishing interpretable frameworks for fake news detection using attention-based methods, self-attention mechanisms play a pivotal role. They enhance interpretability by enabling the model to capture contextual dependencies, identify key information, generate contextualized

representations, offer attention attribution, and produce explainable features. This concerted effort aims to advance the transparency and interpretability of fake news detection models, addressing the challenges posed by the complexity of attention mechanisms in the context of misinformation identification.

Attention attribution in the realm of Fake News Detection is a pivotal aspect of model interpretation, providing valuable insights into the decision-making process. This attribution allows analysts to identify salient tokens within the input sequence, discerning words or phrases that hold significant relevance for the model's classification. By examining attention scores, one can pinpoint discriminative features indicative of misinformation, offering a quantitative measure of each token's impact. Visualizing attention scores creates a dynamic representation of the model's decision path, illustrating how attention shifts across the input sequence and highlighting the prioritized words at each step. This not only enhances interpretability but also facilitates a nuanced understanding of the linguistic features driving the classification. Attention attribution aids in comprehending the relationships between different tokens, shedding light on syntactic and semantic connections crucial for validation. Moreover, it allows for a deeper analysis of the model's behavior, scrutinizing instances where attention is assigned to specific tokens associated with biased language or propaganda. Ultimately, attention attribution contributes to model transparency, enabling effective communication of the decision rationale to end-users and stakeholders and fostering trust in the model's ability to discern between genuine and fake news. With this, we hope that researching attention attribution in Fake News Detection contributes to the development of more transparent, interpretable, and accountable models. It aligns the capabilities of AI systems with human expectations, ensuring that the technology is harnessed responsibly in the critical task of combating misinformation.

Chapter 4: Datasets

4.1. SST-2

The Stanford Sentiment Treebank 2 (SST-2) dataset is a binary sentiment classification dataset derived from the original Stanford Sentiment Treebank (Socher et al. 2013). Introduced as a simplified version, SST-2 focuses on a binary sentiment categorization task, where each sentence is labeled as either positive or negative. Introduced first by Pang and Lee (2005), the original SST dataset offers a diverse collection of sentences from movie reviews annotated with fine-grained sentiment labels. What sets SST apart is its hierarchical structure, where each sentence is associated with a sentiment label, and this label is recursively applied to every constituent phrase or subphrase within the sentence. This hierarchical annotation provides a nuanced perspective on sentiment,

enabling researchers to explore sentiment compositionality. The new and improved SST-2 dataset is a sentiment annotation structure that streamlines the dataset for a more straightforward sentiment analysis objective. While retaining the diversity of sentence structures and sentiments from various movie reviews, SST-2 provides a practical and efficient resource for training and evaluating sentiment analysis models. The dataset has been instrumental in advancing research on binary sentiment classification and serves as a valuable benchmark for assessing the performance of natural language processing models in capturing the positivity or negativity of textual content. The sentiment labels in SST-2 are binary, indicating whether the sentiment expressed in a sentence is positive or negative. It was parsed with the Standford parser and includes a total of 215,154 unique phrases from those parsed trees (using the Standford parser introduced by Klein and Manning (2003)) and each sentence is annotated by 3 human judges. During data cleaning and processing, we split the dataset into three sub-datasets. This dataset was used to to test the efficacy of the ATTATTR model as the authors also used this dataset in their paper.

4.2. IMDb

The IMDb dataset, often used for sentiment analysis, comprises a vast collection of movie reviews sourced from the Internet Movie Database (IMDb). These reviews are labeled with sentiment polarity, indicating whether the sentiment expressed is positive or negative (Maas et al., 2011). Typically utilized for training and evaluating sentiment analysis models, the IMDb dataset captures diverse opinions and expressions from users' reviews of films. The dataset's richness lies not only in its large size but also in the varied language styles and sentiments it encompasses, making it a valuable resource for researchers and practitioners seeking to develop and assess natural language processing models with a focus on sentiment analysis in the context of movie reviews. Unlike some curated datasets, IMDb itself is a user-generated platform that allows individuals to submit reviews and ratings for movies. As such, the dataset is not developed by specific authors or institutions but is a collection of reviews contributed by users on the IMDb platform. Researchers and practitioners often compile and preprocess subsets of these reviews for sentiment analysis tasks, utilizing the wealth of opinions and sentiments expressed by IMDb users for training and evaluating natural language processing models. Consisting of 50,000 movie reviews, this IMDb dataset was used to understand the ATTATTR model and the authors’ work in the context of sentiment analysis.

4.3. Fake News Constraints

The Fake News Constraints dataset comprises 10,700 meticulously annotated social media posts and articles, distinguishing between real and fake news related to COVID-19. Established in 2021 by Patwa et al., this dataset addresses the imperative of

countering the "infodemic" of misinformation prevalent during the COVID-19 pandemic. Aggregating data from diverse social media platforms such as Twitter, Facebook, and Instagram, as well as fact-checking websites, the shared task involves classifying social media posts as either authentic or deceptive. For example, the following two posts in Figure 4.3. belong to fake and real categories, respectively. Notably, the 'real' category encompasses tweets from 14 different verified sources, including official government accounts, medical institutions (e.g. World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), etc.), and reputable news channels, ensuring the dissemination of accurate and helpful information on COVID-19. Each tweet was reported to be read by humans and is marked as real news if contains useful information on COVID-19. Conversely, the 'fake' classification encompasses tweets, posts, and articles propagating unsubstantiated claims and speculations about COVID-19, diligently sourced from public fact-checking websites and social media. Rigorous manual verification involving scrutiny of original documents from web-based resources, including Facebook posts, tweets, news pieces, Instagram posts, public statements, press releases, etc., along with consultation of fact-checking websites like Politifact, Snopes, and Boomlive, underscores the dataset's reliability. In essence, the Fake News Constraints dataset stands as a meticulously curated and annotated resource for social media posts and articles, distinguishing between authentic and deceptive narratives surrounding COVID-19.

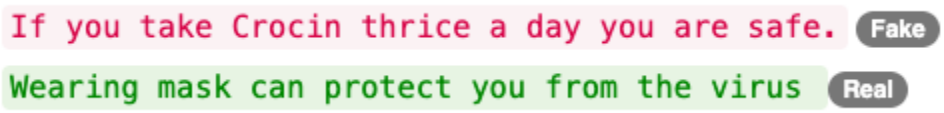

Figure 4.3. Example of Fake News Constraint Dataset

4.4. Fake/Real News

This second Fake News dataset has been meticulously assembled from real-world sources, underscoring its authenticity and relevance. To capture truthful opinions, a compilation of news articles sourced from *Reuters.com*, a reputable news website, was undertaken. For the category of fake news, a dataset available on *Kaggle.com* was employed, specifically featuring fake news originating from unreliable websites targeted by Politifact, a prominent fact-checking organization in the USA, in collaboration with Facebook to counter misinformation. The dataset incorporates 12,600 fake news articles and an equal number of truthful articles. The deliberate focus on political news articles stems from the heightened susceptibility of this genre to misinformation. Notably, both fake and truthful news articles are contemporaneous, sharing a timeline confined to the year 2016. Each article in the dataset surpasses a length threshold of 200 characters. Key information provided for each review includes the article text, article type, article label

indicating its authenticity (fake or truthful), article title, and the article date, ensuring comprehensive coverage for subsequent analysis and interpretation. Figure 4.4 shows the 20 most frequently used bigrams in fake and honest news from the dataset, respectively. As can be seen, both types of articles contain similar terms such as “Hillary Clinton” and “United States.” We can also see a variety of terms in both classes. However, the fake news category tends to have a skewed focus on the topic of “Clinton” than the opposing class.

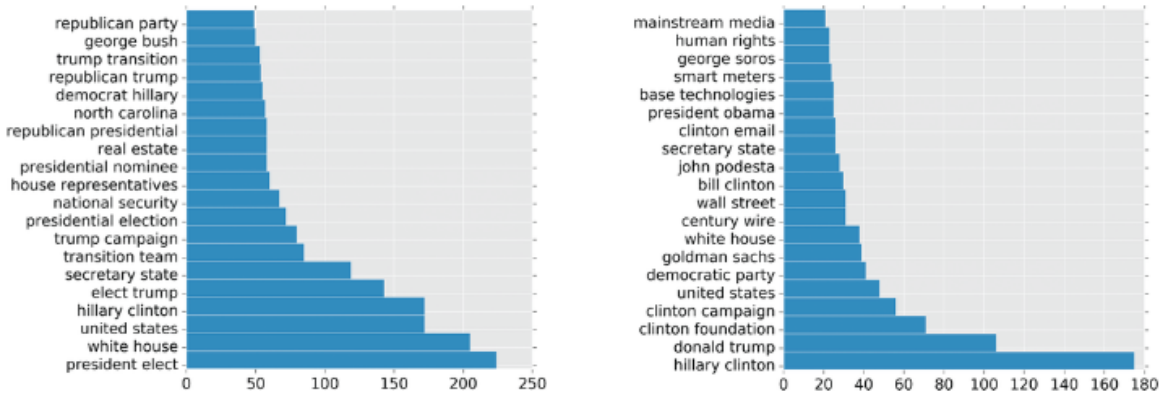

Figure 4.4. Top 20 bigram in real news (right) and fake news (left).

Chapter 5: Methodology

5.1. Self-Attention Attribution

Self-attention attribution (ATTATTR) emerges as a promising method for enhancing interpretability (Hao, Dong, Wei, and Xu, 2021). The authors propose ATTATTR as a technique that assigns attribution scores to input tokens based on their contributions to the attention heads within the Transformer model. These attribution scores provide insights into which tokens the model focuses on during its decision-making process. This approach holds the potential for greater interpretability compared to previous methods discussed. In light of that, we wanted to investigate whether this new method can be used as an interpretable method and whether it is a *good* interpretability method, especially for the Fake News dataset.

This method was first introduced to help interpret the information interactions inside the Transformer model, introduced by Hao, Dong, Wei, and Xu (2021). Due to the black-box-model nature of the Transformer model, the authors were inspired to address this issue to make self-attention more interpretable. The proposed method is to enhance

attention mechanism interpretability using ATTATTR to identify crucial attention connections and observe a lack of consistent correlation between attention weights and their impact on predictions. Introducing a heuristic algorithm for constructing self-attention attribution trees, they illuminate the information flow within the Transformer. Quantitative analysis assesses the contribution of attribution tree edges to predictions. Then, ATTATTR identifies significant attention heads, enabling competitive head pruning performance. Lastly, they leverage ATTATTR to extract interaction patterns as adversarial triggers, revealing the overemphasis of specific word patterns in fine-tuned models, diminishing robustness.

They believe that the attention score matrix, when visualized on only one head in the fine-tuned BERT model, appears dense, making it challenging to comprehend word interactions within the Transformer. Notably, a high attention score does not necessarily imply the importance of the word pair to model decisions. The focus is on attributing model decisions to self-attention relations, which assign higher scores when the interaction significantly contributes to the final prediction, which leads us to the next section about attribution score and how it is calculated.

5.2. Attribution Score

The attribution score in the paper refers to the score assigned to each input token based on its contribution to the attention heads within the Transformer model. It reflects the importance or relevance of each token in the decision-making process.

In the paper, the authors calculate the attribution score using the ATTATTR technique (Hao, Dong, Wei, and Xu, 2021). The ATTATTR method involves two main steps:

1. Attention-based feature importance scores: First, the authors compute the importance scores for each attention head in the model. These scores are calculated by measuring the attention weights between all pairs of tokens in the input sequence. The higher the attention weight between two tokens, the more important the interaction between them is considered.

2. Token-level attributions: The importance scores obtained in the previous step are then aggregated to create token-level attributions. To achieve this, the authors use a propagation algorithm that distributes the attention weight from each head to the tokens they attend to. By aggregating the attention weights from all attention heads, a combined attribution score is obtained for each token.

By following this approach, the authors can calculate the attribution scores for each token, providing insights into their contributions and relative importance within the Transformer model.

5.3. Attribution Tree Construction

The purpose of the attribution tree constructed is to provide interpretable insights into the information interactions happening inside a Transformer model. The attribution tree helps us understand how different input tokens contribute to the output tokens by visualizing the important connections (Hao, Dong, Wei, and Xu, 2021). It provides a structured representation of the model's attention mechanism, highlighting the tokens that have the most influence on the final prediction. With the attribution tree, one can identify information flow paths and observe token importance patterns across layers and self-attention heads. This enables researchers and practitioners to gain deeper insights into the inner workings of the Transformer model, understand its decision-making process, and potentially diagnose any biases or issues. Moreover, the interpretation provided by the attribution tree can help in various applications such as model debugging, model compression, and explaining the model's output to end-users. It adds transparency and explainability to the otherwise complex and black-box nature of Transformer models. The attribution tree is constructed using a bottom-up approach (Hao, Dong, Wei, and Xu, 2021), following these steps:

1. The first step is to compute the self-attention attributions for all the self-attention heads and layers in the transformer model. These attributions represent the importance of each input token in influencing the output token.
2. Once the attributions are computed, the next step is to construct an attribution tree based on the attributions. The tree is constructed by connecting the important input tokens to the corresponding output tokens.
3. Starting from the output layer, the attribution tree is built layer by layer in a bottom-up manner. For each layer, the tokens that have the highest attributions are connected to their corresponding output tokens.
4. The tree construction process continues until reaching the input layer, where the tokens are connected to the final output tokens.
5. At each layer, the tokens that are connected in the attribution tree are pruned to remove any irrelevant connections. This helps in simplifying the interpretation and reducing unnecessary connections.

By following this approach, the attribution tree represents the hierarchical structure of information interactions inside the transformer model, providing insights into how input tokens contribute to the output tokens.

5.4. Attention Head Pruning

As mentioned previously, a distinct aspect of the model architecture is its attention mechanism, where only a subset of attention heads actively contributes to the final prediction, while others exhibit comparatively diminished relevance. Because of that, the authors were motivated to study how to identify and prune attention heads that are not as important (Hao, Dong, Wei, and Xu, 2021).

The importance of attention heads is determined based on their impact on the model's performance and their contribution to the overall information flow within the Transformer model. The authors propose two metrics to assess the importance of attention heads:

1. Reconstruction Error: The reconstruction error measures the discrepancy between the original representations and the reconstructed representations obtained by excluding the attention weights of a particular head. If the reconstruction error is low, it implies that excluding the attention head has minimal impact on the model's ability to reconstruct the original representations, indicating that the head may be less important.

2. Contribution to Attribution Scores: This metric measures the contribution of each attention head towards the attribution scores obtained from the attribution tree. The authors calculate the average contribution of each head across all tokens, considering both positive and negative attributions. If an attention head consistently contributes a higher amount to the attribution scores, it is considered more important.

Combining these two metrics, the authors identify the attention heads that have a high reconstruction error and a significant contribution to the attribution scores as important heads. These heads play a crucial role in the information flow and decision-making process of the Transformer model.

5.5. Adversarial Attack

The adversarial attack serves as a means to assess the robustness and reliability of the attribution scores obtained from the self-attention mechanism. The purpose of the adversarial attack is to introduce perturbations or modifications to the input sequence to manipulate the attribution scores (Hao, Dong, Wei, and Xu, 2021). This helps evaluate the sensitivity of the attribution method to targeted perturbations and assess if it provides reliable interpretations. By crafting adversarial examples through perturbations, the authors seek to investigate whether the self-attention mechanism consistently assigns

high attribution scores to the original important tokens despite these modifications (Hao, Dong, Wei, and Xu, 2021). It aims to test the robustness of the attention mechanism in capturing meaningful interactions and detecting relevant information.

The model's decision-making process primarily relies on attention connections featuring higher attribution scores. The authors' observation reveals a tendency in the model to excessively highlight certain individual patterns when making predictions, often overlooking the majority of the input (Hao, Dong, Wei, and Xu, 2021). Subsequently, they leverage these over-emphasized patterns, identified as over-confident indicators, as adversarial triggers following the approach outlined by Wallace et al. (2019) to launch attacks on the BERT model.

A part of determining the adversarial attacks is trigger constructions. Trigger construction was performed to investigate the role of specific words or phrases in the context of a transformer model (Hao, Dong, Wei, and Xu, 2021). By constructing and evaluating triggers in this way, the paper aimed to analyze the effect of specific words or phrases on the attention mechanism of the Transformer model and gain insights into the interpretation of information interactions within the model.

Chapter 6: Experimental Results

6.1. Attribution Score

The attention score is a fundamental aspect of transformer models. It represents the level of importance or relevance assigned to each token in the input sequence by the self-attention mechanism. Attention scores determine how much the model attends to each token when processing the input and are used to aggregate information across the sequence. On the other hand, attribution scores in this paper refer to the interpretability of the attention mechanism. They aim to provide insights into why the model pays attention to certain tokens by attributing importance scores to each token. Attribution scores help understand which words or phrases contribute the most to the model's decision-making process. By comparing the attention scores obtained directly from the model with the attribution scores derived from interpretation techniques, we are assessing the reliability and robustness of the attention mechanism in transformer models in terms of interpretability.

SST-2

In Figure 6.1.1, the attention scores of a single head in the fine-tuned BERT are depicted for the SST-2 dataset. The sample sentence tested is "*plays as hollow catharsis, with lots of tears but very little in the way of insights*" belonging to the negative sentiment class.

Our observation indicates that the attention score matrix is notably dense, despite the representation of only one out of twelve heads. This density presents a significant challenge in comprehending the intricate interactions between words within the Transformer. Furthermore, it is crucial to note that a high attention score does not necessarily imply the importance of the corresponding pair of words in influencing the model's decision-making process.

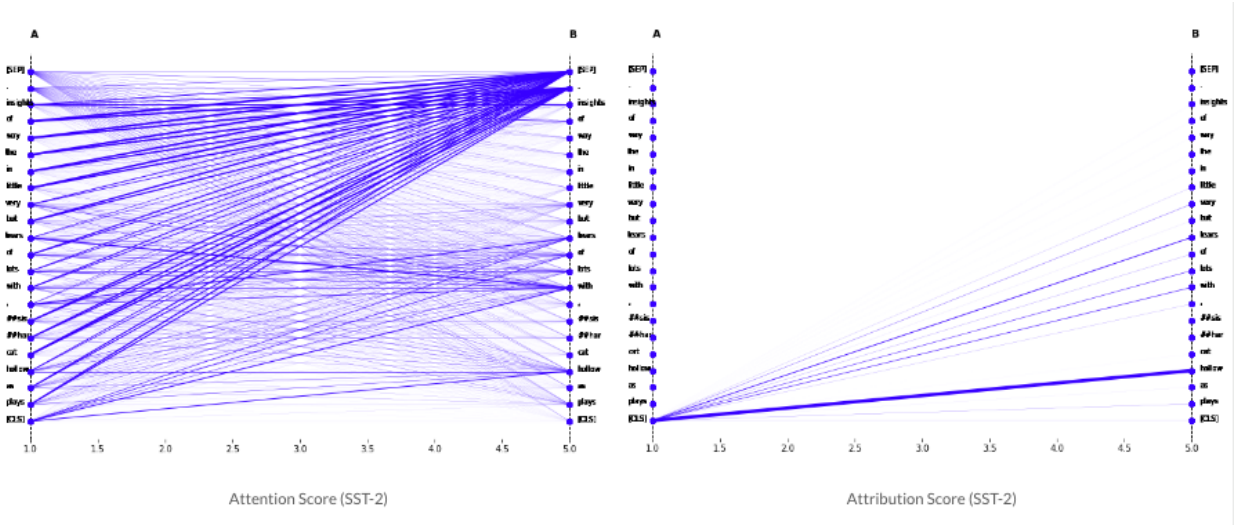

Figure 6.1.1. Attention Score (left) and Attribution Score (right) of a single head in BERT for SST-2 dataset.

The shading intensifies with higher values. We assert that augmented attention scores do not proportionately correlate with enhanced contributions to the ultimate prediction. Notably, the attention scores between the [SEP] token and other tokens register as relatively substantial; however, they yield minimal to negligible attribution scores. The predilection towards negative sentiment classification predominantly stems from the connections linking the [CLS] token in the initial segment with the term "hollow" in the subsequent segment, rendering the attribution process more readily explicable. Attention Attribution tends to discern sparse word interactions that substantively influence the conclusive model decision.

Fake News Constraints

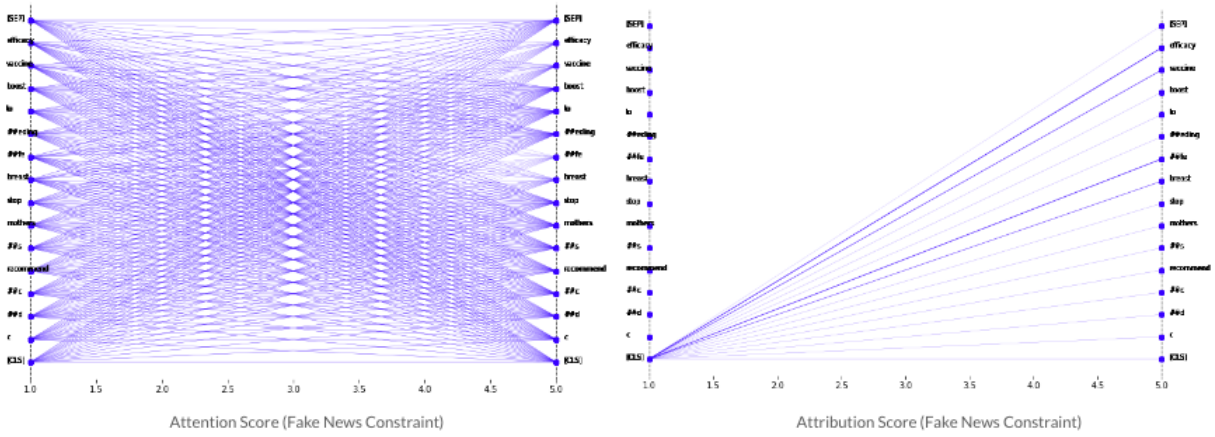

Figure 6.1.2. Attention Score (left) and Attribution Score (right) of a single head in BERT for Fake News Contraint dataset.

Similarly, we constructed a parallel plot for both the Fake news datasets. For the Fake News Contraint dataset, which consists of social media posts, we used the example sentence "*CDC recommends mothers stop breastfeeding to boost vaccine efficacy*" which belongs to the fake class, shown in Figure 6.1.2. Observing the attention scores alone, pinpointing the specific areas where the model directs its focus becomes a challenging endeavor. The figure essentially shows that each word in the example sentence is self-attending to all the words in the sentence. In contrast, employing the attribution score reveals that the primary contribution to the prediction of the fake classification emanates from the connections linking the [CLS] token in the initial segment with the terms "efficacy," "vaccine," and "feeding" in the second segment. This attribution pattern proves to be more readily explicable, enhancing interpretability in model behavior.

Fake/Real News

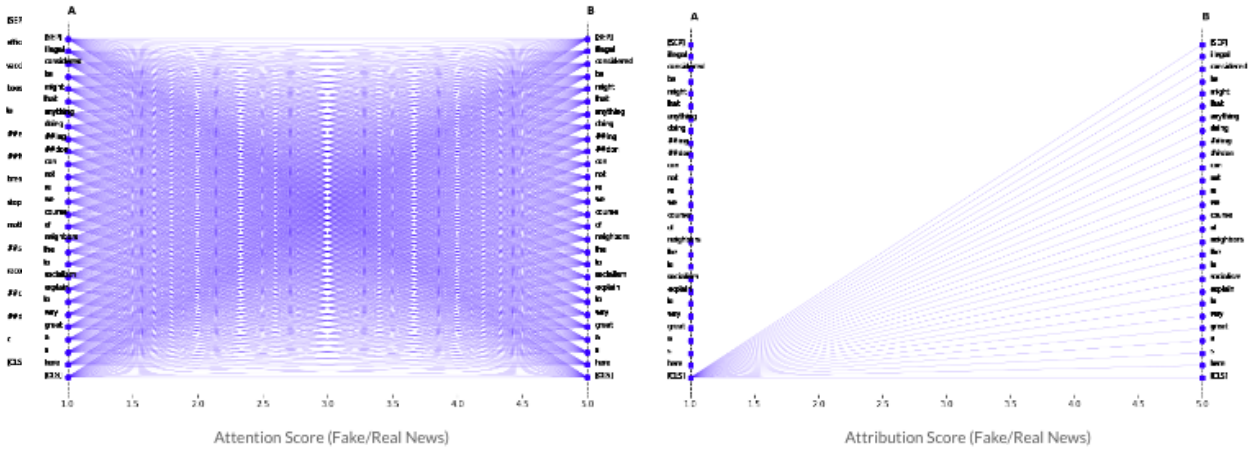

Figure 6.1.3. Attention Score (left) and Attribution Score (right) of a single head in BERT for Fake/Real News dataset.

Within the context of the fake/real news dataset, we present an additional parallel plot that juxtaposes the attention and attribution scores for an exemplary sentence: "*Here's a great way to explain socialism to the neighbors, of course, we're not condoning doing anything that might be considered illegal,*" assigned to the fake news class. This analysis unveils a notable constraint in the attention-attribution model's applicability to the fake news dataset. Specifically, ambiguity arises as to why the attribution scores exhibit a uniform similarity pattern from the [CLS] token in the first segment to all tokens in the second segment. This apparent uniformity poses a challenge in discerning nuanced word interactions that contribute differentially to the model's decision-making process, thereby emphasizing the necessity for enhanced interpretability and discriminative power in model evaluation.

6.2. Attribution Tree

Employing the devised heuristic algorithm for the construction of attribution trees, we have endeavored to delve comprehensively into the intricacies of deciphering the information flow within the Transformer model across our four datasets. This sophisticated approach affords us the capability to gain profound insights into the dynamic interactions among input words within a given sentence and their consequential attribution to the model's ultimate prediction. The visualization derived from this heuristic algorithm serves as a powerful tool, enabling us to capture the intricate dependencies that the Transformer model is inclined to apprehend during its

decision-making process. This nuanced understanding not only enhances our interpretative grasp of the model's inner workings but also positions us strategically to leverage post-interpretation insights for debugging and refining the training data. By elucidating the specific contributions of individual words and their interplay, this post-interpretation process becomes instrumental in refining the model, thereby elevating its robustness and performance across diverse scenarios.

SST-2

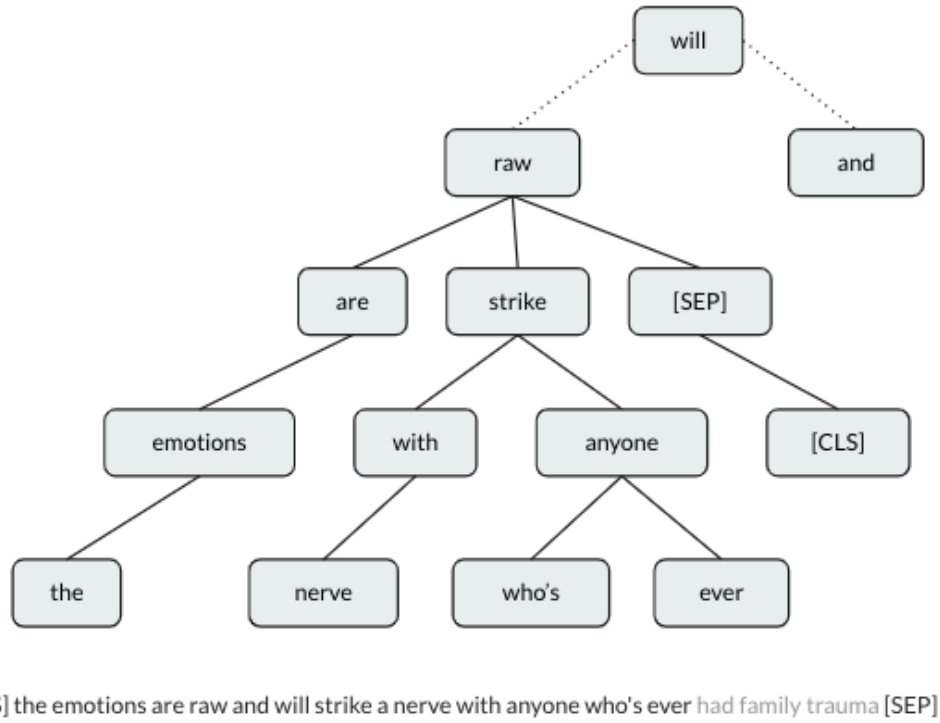

Figure 6.2.1. SST-2 example of Attribution Tree which is predicted as positive by BERT.

Within the same SST-2 dataset, we undertook an analysis using a distinct example and constructed an attribution tree for the sentence, "*The emotions are raw and will strike a nerve with anyone who's ever had family trauma*." This sentence belongs to the positive class. Our attribution tree construction algorithm facilitates the combination of interactions, essentially delineating the information flow from input tokens to the ultimate predictions. This visualization strategy serves as a powerful mechanism for unraveling the dependencies that the Transformer model is predisposed to capture during its decision-making process.

In the case of the attribution tree generated from the SST-2 dataset, where the golden label is positive, a closer examination reveals a locality in interactions at the bottom, with most information flows concentrated within a single sentence. Hierarchically, information aggregates from the term "raw" in the sentence, demonstrating a localized impact on the decision-making process. The attribution tree acts as an interpretive guide, elucidating how input words interact with each other to culminate in the final prediction, thereby enhancing the interpretability of model decisions. Notably, through the examination of feature interactions, we observe that all information aggregates toward the verb "will," providing a more nuanced understanding of why the model arrives at a specific decision.

It is pertinent to highlight that certain words at the bottom of the attribution tree, originating from the input, do not manifest in the tree, underscoring the selective nature of the model's information aggregation process. This nuanced analysis further contributes to a refined comprehension of the model's decision-making dynamics.

<u>Fake News Constraints</u>

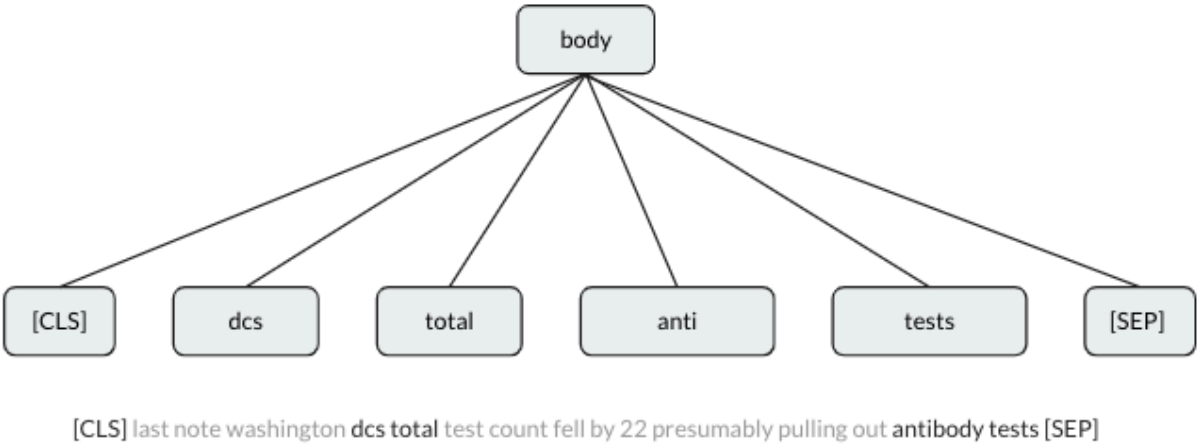

Figure 6.2.2. Fake News Constraint example of Attribution Tree which is predicted as real by BERT

Likewise, we sought to offer an elucidative perspective by presenting an attribution tree for the Fakenews Constraint dataset. The selected sentence, "*Last note: Washington DC's total test count fell by 22, presumably pulling out antibody tests,*" is classified under the golden label "fake." In examining this attribution tree, a discernible pattern emerges as the words "DCs," "total," "anti," and "tests" collectively contribute to the aggregation of the word "body." This convergence aligns logically with their contextual interconnections, thus enriching our understanding of the contextual dependencies that the model captures in the classification of deceptive news content.

In the attribution tree construction process, not all words necessarily appear in the tree visualization. This is primarily because the attribution tree aims to highlight the most influential and discriminative tokens based on their attribution scores, rather than providing a comprehensive representation of every single token. Specifically, tokens that contribute significantly to the attribution scores and have a noticeable impact on the self-attention mechanism are selected to appear in the attribution tree. These tokens are usually the ones that have a strong influence on the model's decision-making process. In the example discussed above, we see that out of all the words in the example sentence, only four words play a crucial role in the model's attention mechanism and they capture the most influential aspects of the input sequence.

Fake/Real News

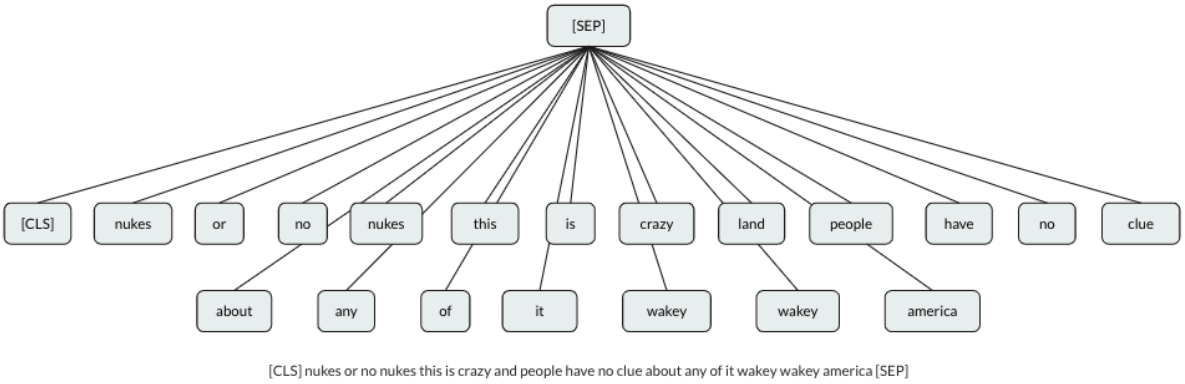

Figure 6.2.2. Fake/Real News dataset example of Attribution Tree which is predicted as fake by BERT

Once again, an attribution tree has been meticulously crafted for the Fake/Real News dataset, with a specific focus on the sentence, "*Nukes or no nukes, this is crazy, and people have no clue about any of it. Wakey wakey, America*," which pertains to the classification class fake.

Upon scrutiny of the attribution tree for this particular sentence, a notable observation emerges: all the words in the sentence aggregate towards the token [SEP]. This concentration of feature interactions at the [SEP] token complicates the understanding of information flow and the rationale behind the model's specific decisions. Although this sentence is presented as an illustrative example of the dataset's attribution tree results, a systematic exploration of various sentences from the dataset reveals a consistent trend. In a similar vein, other examples exhibit the aggregation of words towards either the [CLS]

or [SEP] tokens, thereby posing challenges in unraveling the nuanced decision-making processes of the model.

It is imperative to acknowledge that the chosen example is relatively succinct, while the majority of the dataset comprises more extensive paragraphs sourced from various news outlets. Notably, even in these longer examples, the aggregation of information towards specific tokens, such as [CLS] or [SEP], persists, exacerbating the complexity of deciphering the model's decision logic. This underscores a noteworthy limitation in the interpretability of the attribution tree approach when applied to lengthier textual inputs, necessitating a nuanced consideration of the dataset's composition and the implications for model understanding.

6.3. Self-Attention Head Pruning Evaluation

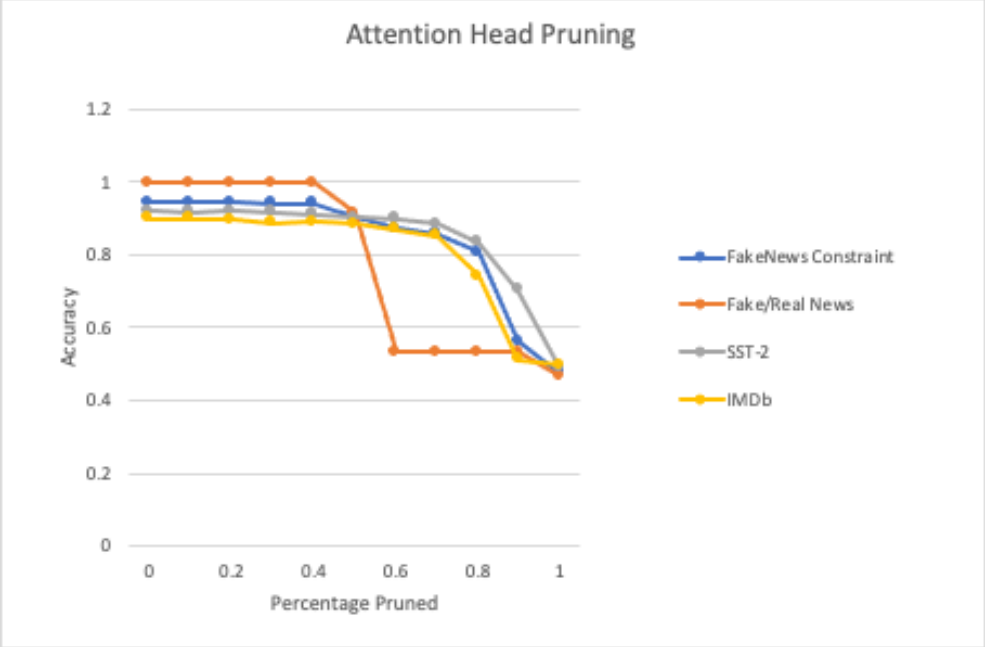

Figure 6.3. Comparison of evaluation accuracy as a function of head pruning proportion for all four datasets.

The exploration of self-attention head pruning represents a pivotal facet of the methodology. As delineated in the preceding section, our investigation revealed that only a fraction of attention heads significantly contribute to the final prediction, with others offering marginal utility. This observation prompted an in-depth inquiry into identifying and pruning attention heads deemed less consequential. The attribution scores, quantifying the contribution of self-attention edges to the ultimate model decision, played a pivotal role in guiding this process.

The presented graph in Figure 6.3 encapsulates the outcomes of the head pruning methodology across all four datasets. The scores depicted therein elucidate the performance of pruned models across various pruning rates, offering insights into the trade-off between model size (corresponding to a reduction in attention heads) and task performance. Generally, higher scores connote superior performance in the evaluation task.

Upon meticulous analysis, a discernible pattern emerges, highlighting a notable accuracy drop, particularly pronounced in the context of the fake/real news dataset. Notably, a reduction of approximately 40% in accuracy is observed when 60% of attention heads are pruned. Furthermore, a more drastic decline of approximately 50% in accuracy is evident when the pruning rate reaches 100%—indicative of a complete removal of attention heads. These results underscore the delicate balance between model size optimization and task-specific performance, emphasizing the nuanced considerations required in the strategic pruning of attention heads.

6.4. Adversarial Attacks/Triggers Evaluation

The methodology also evaluates the robustness of the attribution methods against adversarial attacks and studies the effectiveness of incorporating attribution-based triggers in the input sequence to deceive the model's prediction. The results indicate that incorporating attribution methods during model training can enhance the robustness against adversarial attacks. However, the susceptibility to trigger-based attacks highlights the need for robust defense mechanisms that consider and mitigate the influence of misleading attributions in the input sequence.

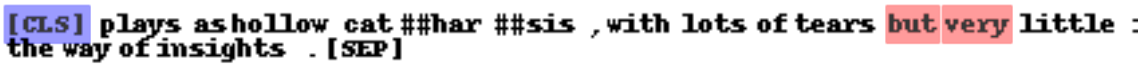

Figure 6.4.1. Using ATTATTR to extract the highlighted trigger words for example in the SST-2 dataset.

Leveraging attention attribution as a mechanism for extracting trigger words, we endeavor to employ these identified triggers as evidentiary elements in our study. To substantiate our findings, we present illustrative examples drawn from all four datasets. In each instance, the highlighted words correspond to the top two pairs of patterns that wield the most significant influence on the model's prediction within the negative sentiment class. Notably, the identified triggers, such as [CLS], [but], and [very], are considered key elements extracted from the model through attention attribution.

An inherent expectation arises from this extraction process—namely, that the deployment of these trigger words in analogous positions within other sentences should elicit model predictions indicative of negative sentiment. This hypothesis underscores the transferability of identified triggers across diverse examples within the same dataset. The identified triggers, once extracted, become instrumental in orchestrating controlled experiments to assess the model's sensitivity and responsiveness to these specific linguistic patterns.

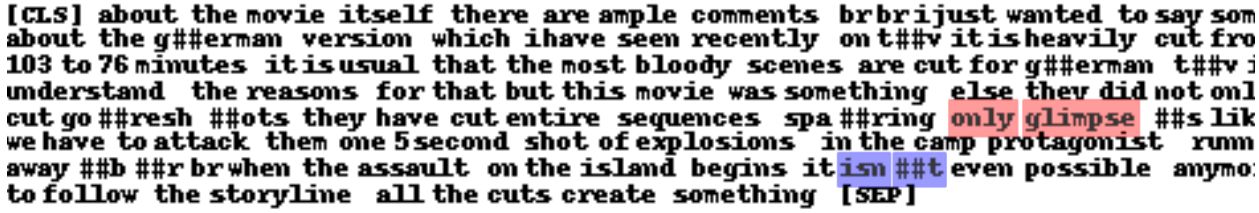

Figure 6.4.2. Using ATTATTR to extract the highlighted trigger words for example in the IMDb dataset.

This analytical approach is extended to the IMDB dataset, where a parallel investigation is conducted to ascertain the efficacy of the identified triggers in influencing sentiment predictions. By systematically applying the trigger words in analogous positions across various sentences within the IMDB dataset, we aim to assess the consistency of the model's response and validate the generalizability of our findings. This methodological strategy aligns with the overarching objective of discerning the robustness and reliability of trigger words identified through attention attribution across diverse contexts and datasets.

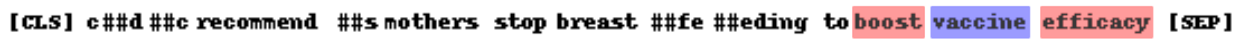

Figure 6.4.3. Using ATTATTR to extract the highlighted trigger words for example in the Fake News Constraint dataset.

The trigger words identified within the examples presented earlier for the fake news constraint dataset include "boost," "vaccine," and "efficacy." These terms were discerned through a meticulous analysis of attention attribution patterns, signifying their pronounced influence on the model's prediction, particularly within the context of deceptive news classification. The selection of these trigger words is strategic, as they encapsulate key thematic elements often associated with misleading information in the context of public health and societal concerns.

Figure 6.4.4. Using ATTATTR to extract the highlighted trigger words for example in the Fake News Constraint dataset.

In a parallel instance within the fake/real news dataset, a distinct set of trigger words emerges. Notably, the [CLS] token and "socialism" are identified as pivotal linguistic cues contributing significantly to the model's decision-making process. The [CLS] token, serving as a representation of the entire input sequence, signifies a holistic influence on the model's interpretation, while "socialism" points to a specific thematic element that resonates within the context of news classification.

These trigger words, meticulously extracted through attention attribution, embody the salient features that prompt the model's classification tendencies. Their identification is integral to understanding the linguistic nuances that the model deems indicative of deceptive or authentic news content. This strategic approach to trigger word identification contributes to the interpretability and explicability of the model's decision-making, shedding light on the key linguistic components that shape its predictions within the domain of news classification.

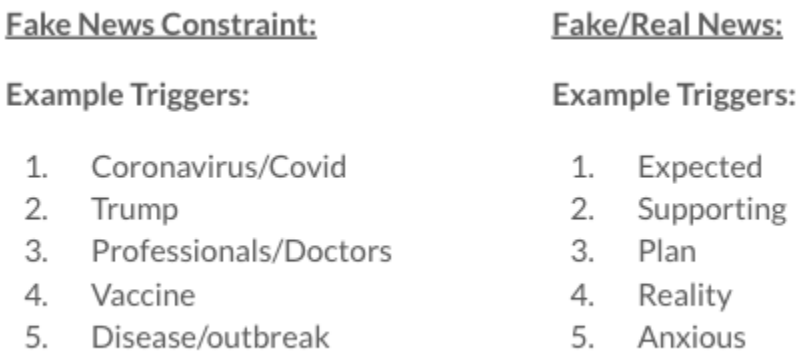

Figure 6.4.5. Examples of Trigger Words from Fake News Datasets

Our exploration of adversarial attack trigger words involved a meticulous evaluation process, culminating in the identification of noteworthy examples. Within the constraint dataset, the discerned trigger words exhibit a discernible thematic alignment, predominantly revolving around the ongoing pandemic. Notable instances include terms such as "coronavirus," "COVID," "Trump," "doctor," "vaccine," "disease," and "outbreaks." These trigger words are intricately linked to the prevailing context of the

datasets, comprising social media posts and tweets pertaining to the pandemic. The duality of these datasets, featuring both authentic and deceptive content, further emphasizes the contextual relevance of these triggers in capturing the nuances associated with pandemic-related information.

Conversely, in the other fake news dataset, the identified trigger words manifest a distinctly political orientation. In this context, the trigger words are predominantly sourced from news outlets, reflecting a news-centric thematic association. The discerned trigger words in this dataset align with the political discourse prevalent in news sources, thereby reflecting the model's susceptibility to linguistic cues rooted in political content.

This comparative analysis underscores the dataset-specific nature of adversarial triggers and highlights the model's sensitivity to thematic nuances. The divergence in trigger word characteristics between datasets illuminates the inherent biases and contextual considerations that influence the model's decision-making processes. Such nuanced insights contribute to a more profound understanding of the intricacies inherent in adversarial attacks within distinct domains.

Chapter 7: Discussion & Future Work

Our endeavor to assess the efficacy of attention attribution scores as a model interpretability method within the realm of Fake News detection has provided valuable insights. Despite conducting a meticulous application of the method on Fake News data and attempting to leverage trigger words as evidentiary elements, the results indicate a need for methodological refinement or exploration of alternative approaches to yield satisfactory outcomes.

While the model exhibits commendable accuracy on the validation dataset and successfully identifies trigger words, concerns arise regarding the reliability of these trigger words as substantive evidence for Fake News. The careful selection of trigger words, coupled with their rarity, poses challenges in obtaining a comprehensive set for evaluation purposes. The absence of an evaluation set further complicates the assurance of robust evidence. Consequently, uncertainties persist regarding the adequacy of optimizing attention for uncovering attribution as a means of identifying evidence for Fake News detection and enhancing interpretability, particularly given the difficulty in discerning meaningful trigger words.

In contrast to common interpretability methods such as Lime, SHAP, and ANCHORS, the attention attribution method exhibits promising potential for identifying pivotal words within Fake News data. Existing literature supports the interpretability achieved through common methods, and attention attribution aligns with these findings, albeit with certain limitations.

Importantly, the dynamic nature of the Fake News landscape underscores the inadequacy of relying solely on a single model for identification.

This study contributes to advancing Fake News research by shedding light on the challenges and nuances inherent in model interpretability for detection. As a suggestion for future work, incorporating a human-annotated evaluation set can universalize the attention mechanism for interpretability. Drawing inspiration from datasets like HateXplain (Mathew et al., 2020), which employs human rationales for training, the integration of human attention in the optimization process can enhance the contextual understanding of why individuals engage in creating and disseminating fake news.

In conclusion, while attention attribution holds promise, its current limitations necessitate collaborative efforts with human annotation to achieve comprehensive interpretability. Striving for a nuanced understanding of the motivations behind the creation and dissemination of fake news remains a complex yet crucial endeavor for advancing research in this dynamic field

References

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. SECURITY AND PRIVACY, 1(1), e9. https://doi.org/10.1002/spy2.9

Alharbi, R., Vu, M. N., & Thai, M. T. (2021). Evaluating Fake News Detection Models from Explainable Machine Learning Perspectives. ICC 2021 - IEEE International Conference on Communications, 1–6. https://doi.org/10.1109/ICC42927.2021.9500467

Baptiste Amato, Alexis Durocher, Gabriel Hurtado, Alexandre Jouandin, & Vincent Marois. (2019, March 1). Learning about the Attention Mechanism and the Transformer Model. https://deepfrench.gitlab.io/deep-learning-project/

BarredoArrieta, A. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inf. Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 (2020).

Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. Frontiers in Big Data, 4. https://www.frontiersin.org/articles/10.3389/fdata.2021.688969

Bhan, M., Achache, N., Legrand, V., Blangero, A., & Chesneau, N. (2023). Evaluating self-attention interpretability through human-grounded experimental protocol (arXiv:2303.15190). arXiv. http://arxiv.org/abs/2303.15190

Ciampaglia, G. L. Fighting fake news: A role for computational social science in the fight against digital misinformation. J. Comput. Soc. Sci. 1, 147–153. https://doi.org/10.1007/s42001-017-0005-6 (2018).

Chorąś, M., Pawlicki, M., Puchalski, D. & Kozik, R. Machine learning–the results are not the only thing that matters! what about security, explainability and fairness? In International Conference on Computational Science, 615–628 (Springer, 2020).

Conroy, N. K., Rubin, V. L. & Chen, Y. Automatic deception detection: Methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. 52, 1–4. https://doi.org/10.1002/pra2.2015.145052010082 (2015).

Das, A. & Rad, P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. CoRR (2020). arXiv:2006.11371.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. IEEE Symposium on Security and Privacy (SP). New York, NY: Institute of Electrical and Electronics Engineers, 598–617.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). Self-Attention Attribution: Interpreting Information Interactions Inside Transformer (arXiv:2004.11207). arXiv. http://arxiv.org/abs/2004.11207

Hoover, B., Strobelt, H., & Gehrmann, S. (2020). exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In A. Celikyilmaz & T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 187–196). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.22

Hosseini, M., Sabet, A. J., He, S., & Aguiar, D. (2022). Interpretable Fake News Detection with Topic and Deep Variational Models (arXiv:2209.01536). arXiv. https://doi.org/10.48550/arXiv.2209.01536

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. Npj Digital Medicine, 6(1), Article 1. https://doi.org/10.1038/s41746-023-00751-9

Jwa, H., Oh, D., Park, K., Kang, J. M. & Lim, H. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Appl. Sci. 9, (2019). https://doi.org/10.3390/app9194062.

Kaliyar, R., Goswami, A. & Narang, P. Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia Tools Appl. 80. https://doi.org/10.1007/s11042-020-10183-2 (2021).

Klein, D. & Manning, C. D. Accurate unlexicalized parsing. In ACL. (2003)

Kula, S., Choraś, M. & Kozik, R. Application of the bert-based architecture in fake news detection. In Herrero, Á. et al. (eds.) 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020), 239–249 (Springer International Publishing, Cham, 2021).

Kula, S., Kozik, R. & Choraś, M. Implementation of the bert-derived architectures to tackle disinformation challenges. Neural Comput. Appl. (2021). https://doi.org/10.1007/s00521-021-06276-0. arXiv:1902.10186

Ksieniewicz, P., Choraś, M., Kozik, R. & Woźniak, M. Machine learning methods for fake news classification. In Yin, H. et al. (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2019, 332–339 (Springer International Publishing, Cham, 2019).

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. IEEE Transactions on Multimedia, 17(11):1875–1886, 2015.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 142–150). Association for Computational Linguistics. https://aclanthology.org/P11-1015

Madsen, A., Reddy, S., & Chandar, S. (2023). Post-hoc Interpretability for Neural NLP: A Survey. ACM Computing Surveys, 55(8), 1–42. https://doi.org/10.1145/3546577

Marvin M Chun, Julie D Golomb, Nicholas B Turk-Browne, et al. A taxonomy of

external and internal attention. Annual review of psychology, 62(1):73–101, 2011.
Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection (arXiv:2012.10289). arXiv. https://doi.org/10.48550/arXiv.2012.10289
Mrini, K., Dernoncourt, F., Tran, Q., Bui, T., Chang, W., & Nakashole, N. (2020). Rethinking Self-Attention: Towards Interpretability in Neural Parsing (arXiv:1911.03875). arXiv. https://doi.org/10.48550/arXiv.1911.03875
Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. International Journal of Data Science and Analytics, 13(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6
Nadia Conroy, Victoria Lubin, Yimin Chen. (2016). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.
Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In ACL, pages 115–124.
Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2021). Fighting an Infodemic: COVID-19 Fake News Dataset (Vol. 1402, pp. 21–29). https://doi.org/10.1007/978-3-030-73696-5_3
Quandt, T., Frischlich, L., Boberg, S., & Schatto-Eckrodt, T. (2019). Fake News. In The International Encyclopedia of Journalism Studies (pp. 1–6). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118841570.iejs0128
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. In AAAI Conference on Artificial Intelligence (AAAI) (2018).
Shapley, L. S. (1952). A VALUE FOR N-PERSON GAMES. Defense Technical Information Center. Santa Monica, CA: RAND Corporatio.
Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods. AAAI/ACM Conference on Artificial Intelligence, New York, NY, February 2020. Ethics, and Society (AIES).
Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.
Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. Scientific Reports, 11(1), Article 1. https://doi.org/10.1038/s41598-021-03100-6

Strumbelj, E., and Kononenko, I. (2010). An Efficient Explanation of Individual Classifications Using Game Theory. J. Mach. Learn. Res. 11, 1–18. doi:10.1145/1756006.1756007

Tandoc, E. C., Lim, Z. W. & Ling, R. Defining fake news. Digital J. 6, 137–153. https://doi.org/10.1080/21670811.2017.1360143 (2018).

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (2 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Vaishnav, M. (2023). PhD Thesis: Exploring the role of (self-)attention in cognitive and computer vision architecture (arXiv:2306.14650). arXiv. http://arxiv.org/abs/2306.14650

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2153–2162. Hong Kong, China: Association for Computational Linguistics.

Wu L, Morstatter F, Carley KM, Liu H. Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explor. 2019;21(1):80–90. doi: 10.1145/3373464.3373475.

Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. European Journal of Operational Research, 279(3), 1036–1052. https://doi.org/10.1016/j.ejor.2019.06.022

Zhou, X. & Zafarani, R. Network-based fake news detection: A pattern-driven approach. ACM SIGKDD Explor. Newsletter 21, 48–60. https://doi.org/10.1145/3373464.3373473 (2019).

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset/activity