Half Truth Detection

Satyam Shukla
M-tech(TA) 23M0817
Computer Science and Engineering
Indian Institute of Technology Bombay
satyamshukla@iitb.ac.in

Abstract

In today's digital age, the proliferation of misinformation poses a significant challenge to discerning truth from falsehood. Among the plethora of misleading information, half-truths stand out as insidious, blending truth elements with falsehood to deceive and manipulate. This research presents a comprehensive framework for detecting half-truths in online content. Our approach begins with the systematic summarization of evidence relevant to a claim. Subsequently, we categorize the claims into three distinct classifications: true, half-true, or false. We demonstrate the efficacy and reliability of our approach in accurately discerning the veracity of claims. This research contributes to the ongoing efforts in combating misinformation by offering a systematic and scalable approach to half-truth detection. By conducting a series of experiments, including utilizing BERT-base (Devlin et al., 2018) and LLAMA-2-7B (Touvron et al., 2023) models on the PolitiFact dataset (Misra, 2022). Additionally, we preprocessed the PolitiFact dataset. We empower users with the knowledge and tools necessary to navigate the increasingly complex landscape of information dissemination.

1 Introduction

The widespread distribution of misinformation, particularly in the guise of half-truths, carries considerable adverse consequences, as it harbors the capacity to unsettle societal and economic cohesion (Allcott and Gentzkow, 2017)(Su et al., 2020). In today's digital age, the proliferation of misinformation has emerged as a formidable challenge, casting a shadow over the integrity of information dissemination and undermining the foundations of trust and credibility. As individuals navigate the vast expanse of online content, they are confronted with a deluge of information, ranging from news articles and social media posts to advertisements and user-generated content. Within this sea of information lies a troubling phenomenon – the propagation of half-truths.

Fake news [1] refers to fabricated or misleading information presented as legitimate news. Often intending to deceive or manipulate public opinion.

Half-truth sentences[2][3] can be partially true or completely true, but they leave out the important context of the information, leading to ambiguous/-multiple interpretations of the sentence. Unlike outright falsehoods, which can often be easily identified and debunked, half-truths operate on a subtler level, exploiting the vulnerabilities of human cognition and perception. By selectively disclosing certain facts while omitting or distorting others, purveyors of half-truths seek to manipulate perceptions, shape beliefs, and influence behavior. Figure 1 is an example of how half-truth is deeply rooted in the advertisement.

In the realm of advertising, half-truths are employed to create persuasive narratives that resonate with consumers while obscuring inconvenient truths or shortcomings. Moreover, the digital ecosystem has facilitated the rapid dissemination and amplification of half-truths, fueling their proliferation and impact. Social media platforms, in particular, serve as breeding grounds for the spread of misinformation, with half-truths often going viral and spreading rapidly across networks of users.

Traditional fact-checking methods, while valuable, are often labor-intensive and time-consuming, making them ill-suited to the rapid pace and scale of online misinformation. Moreover, half-truths present unique challenges, requiring a sophisticated understanding of language, context, and human psychology to discern and refute effectively.

Recent advances in computational techniques and natural language processing (NLP) have offered promising avenues for addressing the chal-

[1] https://ndtv.com/
[2] https://www.flipkart.com/
[3] https://twitter.com/

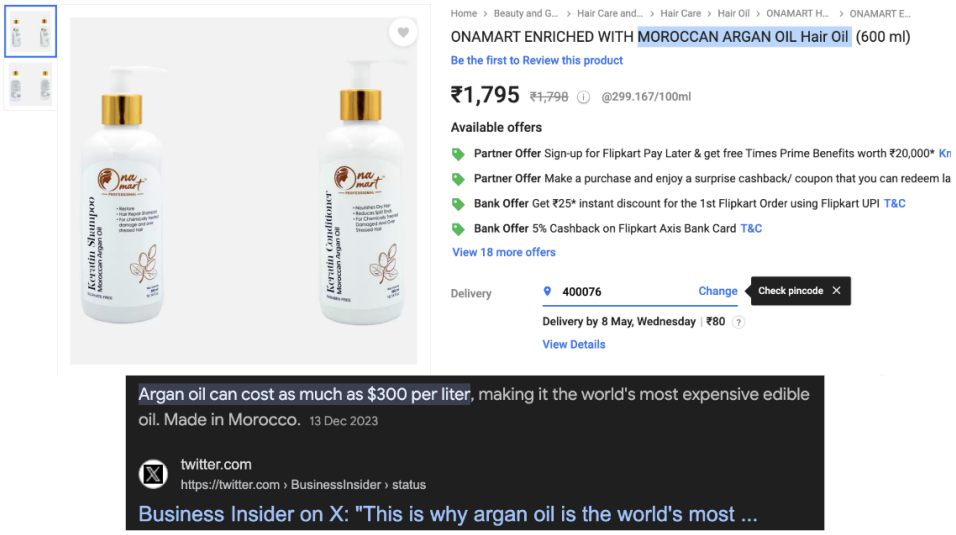

Figure 1: The Above product is an example of a half-truth as the product details do not contain the amount of argon oil used in making this 600ml product.

lenge of half-truth detection. Machine learning models, including deep neural networks and language models, have shown remarkable capabilities in analyzing textual data and identifying patterns indicative of deception. These models leverage large-scale datasets to learn complex language representations, enabling them to detect subtle cues and inconsistencies that may signal the presence of half-truths.

One notable development in this regard is the emergence of language models such as GPT (Generative Pre-trained Transformer) (Radford et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which have achieved state-of-the-art performance across a range of NLP tasks. These models leverage transformer architectures to capture contextual information and semantic relationships within text, allowing them to understand and generate human-like language with remarkable fluency and coherence.

However, despite these advancements, detecting half-truths remains a multifaceted and evolving challenge. The inherent ambiguity and subjectivity of language, coupled with the ever-changing landscape of online communication, necessitate ongoing innovation and adaptation in the field of misinformation detection. Moreover, the ethical implications of automated content analysis and moderation raise important questions about the balance between free expression and the need to combat harmful misinformation.

In light of these challenges and opportunities, this research endeavors to contribute to the growing body of literature on half-truth detection by proposing a comprehensive framework for identifying and classifying half-truths in conferences and debate content. Building upon the theoretical foundations of linguistics and cognitive science, our approach seeks to leverage language models for combating the spread of half-truths.

This research draws upon experimentation with state-of-the-art language models, including Mistrel-7B, LLama-2-7B, and BERT-base, utilizing the Hugging Face transformer library to classify claims within the PolitiFact dataset. Through empirical analysis and validation, the performance of these models in classifying claims into true, half-true, and false categories will be evaluated.

2 Background and Related Work

PolitiFact website[4] has emerged as a prominent player in the landscape of fact-checking initiatives, known for its Truth-O-Meter ratings that evaluate the accuracy of political statements and claims. Founded in 2007 by the Tampa Bay Times, PolitiFact employs a team of journalists and researchers to assess the truthfulness of statements made by politicians, pundits, and public figures. The Truth-O-Meter, PolitiFact's signature rating system, assigns statements a rating ranging from "True" to "Pants on Fire" based on their accuracy.

PolitiFact's methodology involves thorough in-

[4] https://www.politifact.com/

Figure 2: The Above product is an example of fake information as they were fooling people by saying their product is healthy.

vestigation and analysis of statements, drawing upon a diverse range of sources and evidence to determine their veracity. Statements are fact-checked using a combination of journalistic rigor, expert analysis, and reference to primary sources. Moreover, PolitiFact operates with transparency, providing detailed explanations and citations for its rulings to empower readers to evaluate the evidence for themselves.

In recent years, advancements in natural language processing (NLP) and machine learning have revolutionized the field of misinformation detection. One of the key developments in this regard is the Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art NLP model introduced by Google in 2018. BERT leverages a transformer architecture to process contextual information bidirectionally, enabling it to capture intricate linguistic nuances and semantic relationships within the text.

The SentimentalLiar (Upadhayay and Behzadan, 2020) dataset utilized a comprehensive framework for sentiment analysis, encompassing five primary classes of emotions: joy, sadness, anger, fear, and surprise. Each of these emotions represents a distinct aspect of human sentiment and can be expressed through various linguistic cues and textual features.

In addition to the five classes of emotions, the dataset employed sentiment scores to quantify the overall sentiment expressed in a piece of text. These sentiment scores are numerical values assigned to each text sample, indicating the degree of positivity or negativity conveyed in the content. By analyzing the sentiment scores, researchers can assess the overall emotional tone of the text and classify it into different sentiment categories, such as positive, negative, or neutral.

In a study by (Devlin et al., 2018), the authors introduced BERT, showcasing its ability to achieve state-of-the-art performance on a wide range of NLP tasks, including sentiment analysis, named entity recognition and question answering. Researchers have achieved remarkable results in misinformation detection by pre-training BERT on large corpora of text data and fine-tuning it on task-specific datasets, leveraging BERT's contextual understanding of language to discern deceptive content. In addition to BERT, convolutional neural networks (CNNs) (O'shea and Nash, 2015) have been widely utilized (Upadhayay and Behzadan, 2020) for text classification tasks, including misinformation detection. CNNs excel at capturing local patterns and features within textual data by applying convolutional filters. By learning hierarchical representations of text, CNNs can discern subtle cues and patterns indicative of deceptive content.

Recurrent Neural Networks (RNNs) represent a pivotal advancement in the realm of natural language processing and sequential data analysis. Unlike traditional feedforward neural networks, RNNs possess the unique ability to capture temporal de-

pendencies and sequential patterns within data, making them particularly well-suited for tasks involving sequential inputs, such as text analysis and time series prediction. The seminal work by (Ma et al., 2016) titled "Detecting Rumors from Microblogs with Recurrent Neural Networks" underscores the effectiveness of RNNs in detecting misinformation and rumors from microblogging platforms. By leveraging the inherent sequential nature of microblog data, the authors employed RNNs to model the temporal dynamics of information propagation and identify patterns indicative of rumor dissemination. Through their research, Ma and colleagues demonstrated the efficacy of RNNs in discerning factual information and rumors, thereby offering a valuable tool for combating the spread of misinformation in online social networks. Their work not only highlights the potential of RNNs in rumor detection but also underscores the importance of leveraging advanced machine learning techniques to address contemporary challenges in information verification and credibility assessment.

To address the growing issue of misinformation, the natural language processing (NLP) community has dedicated considerable effort to developing automatic fact-checking tools. However, existing research faces limitations that hinder its widespread adoption by real fact-checking organizations like PolitiFact. Many studies have concentrated on claims authored by crowdsourcing, but these claims do not accurately reflect the complexities encountered by fact-checkers when dealing with real-world claims. Additionally, other research that tackles real-world claims either depends on access to a specific set of documents containing the "gold" evidence or conducts unconstrained retrieval, which may lead to the retrieval of articles authored by fact-checkers about the claim. Notably, prior work has not yet implemented a system to retrieve evidence from diverse sources in the wild (Ferreira and Vlachos, 2016) (Alhindi et al., 2018) (Atanasova et al., 2020).

3 Dataset

The dataset utilized in this research originates from PolitiFact, a renowned fact-checking website, and is publicly available on Kaggle. This high-quality dataset comprises 21,152 statements that have undergone thorough fact-checking by experts in various domains. Each statement in the dataset is meticulously categorized into one of six verdict categories: "true," "mostly true," "half true," "mostly false," "false," and "pants on fire," representing a spectrum of veracity levels. Details about each class are given in Appendix A1.

The dataset is structured with eight attributes associated with each record:

- **Verdict:** The outcome of the fact-checking process, categorized into one of the six predefined categories.
- **Statement Originator:** The individual or entity responsible for making the statement being fact-checked.
- **Statement:** The actual statement subjected to fact-checking.
- **Statement Date:** The date when the statement being fact-checked was made.
- **Statement Source:** The source where the statement was originally made, categorized into various types such as speeches, television, news, social media, etc.
- **Factchecker:** The name of the person or organization responsible for fact-checking the claim.
- **Factcheck Date:** The date when the fact-checked article was published.
- **Factcheck Analysis Link:** A link to the detailed analysis article published by PolitiFact, providing additional context and insights regarding the fact-checking process.

The dataset underwent significant refinement and augmentation processes to enhance its quality and suitability for research purposes. Through a series of preprocessing steps, including content extraction via web scraping, filtering of short claims, removal of instances lacking URLs, and elimination of evidence sentences specifying labels for the claim, the dataset was meticulously curated. Moreover, instances where the claim token size exceeded the evidence token size were filtered out to ensure coherence and relevance. Additionally, to streamline the classification task, the original verdict categories were reorganized, with "true" and "mostly true" merged into a single category labeled "true," while "mostly false," "false," and "pants on fire" were consolidated into "false." As a result of

these enhancements, the updated dataset now comprises a total of 18,574 instances, distributed across "true," "half-true," and "false" categories.

- **True:** 6,105 instances in total, with 4,294 instances in the training set, 601 instances in the testing set, and 1,210 instances in the validation set.
- **Half-True:** 4,769 instances in total, with 4,064 instances in the training set, 572 instances in the testing set, and 1,133 instances in the validation set.
- **False:** 7,700 instances in total, with 5,445 instances in the training set, 884 instances in the testing set, and 1,371 instances in the validation set.

This curated dataset serves as the foundation for training and evaluating half-truth detection models, enabling comprehensive analysis and assessment of proposed methodologies in the context of combating misinformation.

4 Methodology

4.1 Finetuning of LLama-2-7B

We leverage the power of the Llama-2-7B model (Touvron et al., 2023), a transformer-based autoregressive causal language model developed by Meta AI. Llama-2-7B represents a significant advancement in natural language processing (NLP), capable of processing human inputs, generating text, and engaging in natural conversations with users. Fig 3 depicts the architecture; we fine-tune LLama 2 7B by giving the claim and evidence text input separated with <SEP> token and asking the model to generate a label. The fine-tuning process involves training the model on our dataset, which consists of instances with attributes such as ID, claim, extracted evidence, speaker, and label.

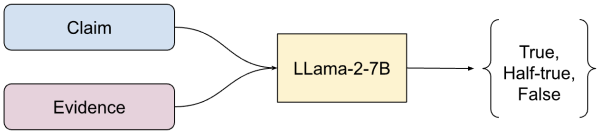

Figure 3: Architecture used for Half-truth detection using LLama-2-7B

4.2 Finetuning of Bert-base

We harness the capabilities of the BERT-base model (Devlin et al., 2018), a widely-used transformer-based model, for our half-truth detection task. BERT-base has demonstrated remarkable performance across various NLP tasks and provides a strong baseline for our classification task.
Fig 4 depicts the architecture, we fine-tune the BERT-base model on our updated dataset derived from the Politifact check dataset; We fine-tuned for half-truth detection; our approach with BERT-base involves a simpler classification task, aiming to classify claims into one of three classes: true, half-true, or false. The input to the BERT-base model consists of both the claim and evidence, enabling the model to capture the context necessary for accurate classification.

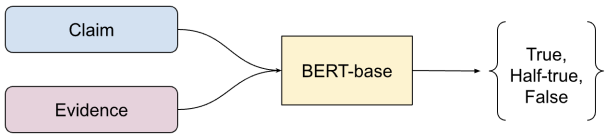

Figure 4: Architecture used for Half-truth detection using BERT-base

5 Experiments and Results

In our experiments with both the Bert-base and LLama-2-7B models, we observed distinct performance outcomes. Table 1 shows while the Bert-base model exhibited satisfactory results, LLama-2-7B faced challenges attributed to its inherent properties. LLama-2-7B, primarily designed for generation tasks, encountered difficulties due to its fixed window size issue. Despite its potential, LLama-2-7B's constrained context window, capped at 4096 tokens, posed limitations, especially concerning in-context learning. Given that the maximum token count in the evidence of our dataset hovered around 2k, LLama-2-7B struggled to process the information within this context window effectively. Conversely, the Bert-base model, tailored for classification tasks, delivered promising outcomes. Its adeptness at generating embeddings for both the claim and evidence facilitated superior classification performance.

Assumption 1 Evidence were containing label itself and in real world scenario this is not going to be the case. So, we removed all sentences from the evidence which were having any of the label(true, mostly true, half true, mostly false, false, or pants fire).

Result Analysis After removing those label, model is not able to predict that good because earlier in

Table 1: Accuracy Test Results where Politifact: Dataset, After Fliter: Dataset after filtering discussed in Dataset section, and Speaker: Adding speaker information to the claim, shown in the LLama-2-7B prompt in 7

Model (Dataset)	**Accuracy (%)**
Bert-base (Politifact)	64
Bert-base (After filter)	62
Bert-base (Speaker)	55
LLAMA-2-7B (Politifact)	44.73
LLAMA-2-7B (After filter)	30.5

dataset there were labels in evidence.

Assumption 2 Adding the name of the speaker to the claim will give give more context about the claim to the model and model will be able to classify claim better.

Result Analysis Further adding speaker names also creates problem may be the model is learning mapping of speaker names with the label of the claim.

6 Summary and Conclusion

The importance of half-truth detection is underscored by its potential societal impacts. Deceptive content, particularly when related to popular figures or sensitive topics, is more likely to go viral, posing risks of national issues and harm to individuals. We analyzed one of the major challenges that we have come across is the lack of adequate data in the LIARPLUS dataset. So, creating and annotating a dataset specifically for gathering a larger quantity of high-quality data on half-truths from news articles.

In today's digital landscape, the prevalence of misinformation, particularly in the form of half-truths, presents a significant challenge. Unlike outright falsehoods, half-truths blend elements of truth with deception, making them difficult to discern and refute. Through a series of experiments utilizing state-of-the-art language models, including LLama-2-7B and BERT-base, on the PolitiFact dataset, the research evaluates various methodologies for half-truth detection. Despite facing challenges, the models demonstrate good performance, particularly BERT-base, tailored for classification tasks. The experiments highlight the importance of context and evidence summarization in accurately discerning the veracity of claims.

7 Future Work

There are several critical directions for future research in half-truth detection. One imperative task is the creation of a comprehensive half-truth dataset. Existing datasets often lack balance and fail to represent the complexities of half-truths adequately. Furthermore, datasets extracted from fact-checking websites such as PolitiFact may suffer from ambiguous classifications, hindering the development of accurate detection models.

To address these challenges, future research should focus on curating a balanced and representative half-truth dataset, drawing upon a diverse range of sources and domains. This dataset should encompass a variety of linguistic expressions and deceptive tactics commonly employed in crafting half-truths. By incorporating techniques based on linguistic interpretation, including semantic, syntax, and pragmatics analysis, researchers can develop more nuanced detection models capable of discerning subtle shades of deception.

Expanding upon considering advertisement datasets, it's crucial to recognize the prevalence of half-truths within advertising contexts. Simple half-truths, especially common in political campaigns and commercial advertisements, often exploit the line between fact and fiction to persuade and influence audiences.

Additionally, researchers should explore the integration of translation datasets, toxicity datasets, and advertisement datasets to enhance the diversity and richness of the training data. Leveraging state-of-the-art language models, such as Mistral 7B, researchers can generate synthetic half-truths to create a more comprehensive and challenging dataset for model evaluation.

Moreover, future research should investigate the virality aspect of half-truth dissemination and its potential societal impacts. By analyzing the virality patterns of deceptive content, researchers can gain insights into the mechanisms driving its spread and develop proactive strategies for mitigating its harmful effects on public discourse and societal well-being.

In summary, future research in half-truth detection should prioritize the creation of a comprehensive half-truth dataset and explore innovative techniques for dataset augmentation and model evaluation. By leveraging advanced computational techniques and interdisciplinary insights, researchers can advance our understanding of half-truths and

develop more effective strategies for combating their spread in the digital age.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Rishabh Misra. 2022. Politifact fact check dataset.

Keiron O'shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bibek Upadhayay and Vahid Behzadan. 2020. Sentimental liar: Extended corpus and deep learning models for fake claim classification. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

A Politifact Truth-o-meter

1. **True:** The statement is accurate, and there are no significant omissions.
2. **Mostly True:** The statement is accurate but may require clarification or additional information.
3. **Half True:** The statement is partially accurate but lacks important details or may be taken out of context.
4. **Mostly False:** The statement contains an element of truth but disregards crucial facts that could alter the perception.
5. **False:** The statement is not accurate.
6. **Pants on Fire:** The statement is inaccurate and makes a ludicrous claim.

B Snapshot of Dataset


```
{
        "id": 15978,
        "label": "half-true",
        "speaker": "Facebook␣posts",
        "claim": "Children␣ages␣5␣to␣9␣are␣not␣
            affected␣by␣the␣coronavirus.",
        "evidence": "In␣a␣viral␣video␣clip,␣a␣
            suburban␣Atlanta␣mother␣who␣implored␣
            her␣local␣school␣board␣to␣stop␣
            mandating␣masks␣for␣children␣claimed␣
            that␣kids␣ages␣5␣to␣9␣are␣not␣
            affected␣by␣the␣coronavirus.␣The␣
            post␣was␣flagged␣as␣part␣of␣
            Facebook's␣efforts␣to␣combat␣false␣
            news␣and␣misinformation␣on␣its␣News␣
            Feed.␣(Read␣more␣about␣our␣
            partnership␣with␣Facebook.)␣Children␣
            are␣less␣likely␣than␣adults␣to␣
            contract␣COVID-19,␣and␣it␣is␣rare␣
            for␣them␣to␣become␣seriously␣ill␣
            with␣the␣virus,␣but␣to␣say␣they␣are␣
            not␣affected␣goes␣too␣far.␣More␣than␣
            3.6␣million␣children␣in␣the␣United␣
            States␣have␣tested␣positive,␣and␣at␣
            least␣297␣have␣died,␣according␣to␣
            the␣latest␣report␣from␣the␣American␣
            Academy␣of␣Pediatrics␣and␣the␣
            Children's␣Hospital␣Association.␣One␣
            reason␣schools␣require␣masks␣is␣to␣
            try␣to␣prevent␣kids␣from␣giving␣the␣
            virus␣to␣other␣people,␣which␣can␣
            happen␣even␣if␣kids␣become␣infected␣
            but␣don't␣show␣symptoms.␣Moreover,␣
```

```
            because adults are being immunized
            and new variants of the virus are
            more likely to infect children,
            children are rapidly becoming the
            major reservoir of COVID in the
            United States, said Dr. Mark
            Schleiss, professor of pediatrics at
            the University of Minnesota Medical
            School. The two-minute video clip
            was widely shared by Facebook users,
            including conservative commentator
            and Fox Nation host Tomi Lahren, who
            has 4.8 million Facebook followers.
            She wrote: This badass stuck her
            neck out and said what so many are
            thinking but are afraid to say! You
            go girl!!! BURN THE MASKS! The clip
            shows Courtney Taylor on April 15
            speaking to the Board of Education
            in Gwinnett County, outside of
            Atlanta. Every month I come here and
            I hear the same thing:
            social-emotional health. If you
            truly mean that, you would end the
            mask requirement tonight. Tonight,
            said Taylor, who identified herself
            as the mother of a 6-year-old
            student. This is not March 2020
            anymore. We have three vaccines.
            Every adult in the state of Georgia
            that wants that vaccine is eligible
            to get it right now and everyone one
            of us knows that young children are
            not affected by this virus. They're
            not. We chose you to make decisions
            that would be in our children best
            interest, and forcing 5-, 6-, 7-, 8-
            and 9-year-old little children to
            cover their noses and their mouths,
            where they breathe, for seven hours
            a day every day for the last nine
            months for a virus that you know
            doesn't affect them that is not in
            their best interest. And this has to
            stop. The district requires masks
            for students, staff and visitors.
            Students can remove masks during
            meals; they can also remove them
            outside when social distancing can
            be achieved.Masks cloth face
            coverings are meant to protect other
            people in case the wearer is
            unknowingly infected but does not
            have symptoms yet, the district
            says."
    }
```

C Prompt of Llama-2-7B

```
def generate_prompt(data_point):
    label_description = {
        "true": "The claim is supported
            by evidence, and it doesn't
            distort the context.",
        "half-true": "Claim suggests
            uncertainty, looking at
            evidence and leave out
            important facts that could
            distort the context of the
            claim.",
        "false": "it's simply inaccurate
            or presents an absurd claim
            ."
    }
    data_point["text"] = f"""
        Given a claim and its
            corresponding evidence,
            classify the claim into one
            of the following three
            classes: "true," "half-true
            ," or "false."

        The claim is true when {
            label_description['true']}
        The claim is half-true when {
            label_description['half-true
            ']}
        The claim is false when {
            label_description['false']}

        Claim: f"{data_point['speaker
            ']}, said, {data_point['
            claim']}"
        Evidence: {data_point["evidence
            "]}
        Label: {data_point["label"]}
    """.strip()

    return data_point
```

Listing 1: LLama-2-7B prompt used for the fine-tuning, added speaker to the claim

```
def generate_prompt(data_point):
    label_description = {
        "true": "the claim is supported
            by evidence, and it doesn't
            distort the context.",
        "half-true": "claim suggests
            uncertainty, looking at
            evidence and leaving out
            important facts that could
            distort the context of the
            claim.",
        "false": "it's simply inaccurate
            or presents an absurd claim
            ."
    }
    data_point["text"] = f"""
        Given a claim and its
            corresponding evidence,
            classify the claim into one
            of the following three
            classes: "true," "half-true
            ," or "false."

        The claim is true when {
            label_description['true']}
        The claim is half-true when {
            label_description['half-true
            ']}
        The claim is false when {
            label_description['false']}

        Claim: {data_point['claim']}
        Evidence: {data_point["evidence
            "]}
```

```
        Label: {data_point["label"]}
    """.strip()

    return data_point
```

Listing 2: LLama-2-7B prompt used for the fine-tuning, no speaker added to the claim