ManiTweet: A New Benchmark for Identifying Manipulation of News on Social Media

Anonymous ACL submission

Abstract

Considerable advancements have been made to tackle the misrepresentation of information derived from reference articles in the domains of fact-checking and faithful summarization. However, an unaddressed aspect remains - the identification of social media posts that manipulate information presented within associated news articles. This task presents a significant challenge, primarily due to the prevalence of personal opinions in such posts. We present a novel task, *identifying manipulation of news on social media*, which aims to detect manipulation in social media posts. To study this task, we have proposed a data collection schema and curated a dataset called ManiTweet, consisting of 3.6K pairs of tweets and corresponding articles. Our analysis demonstrates that this task is highly challenging, with large language models (LLMs) yielding unsatisfactory performance. Additionally, we have developed a simple yet effective framework that outperforms LLMs significantly on the ManiTweet dataset. Finally, we have conducted an exploratory analysis of human-written tweets, unveiling intriguing connections between manipulation and factuality of news articles.

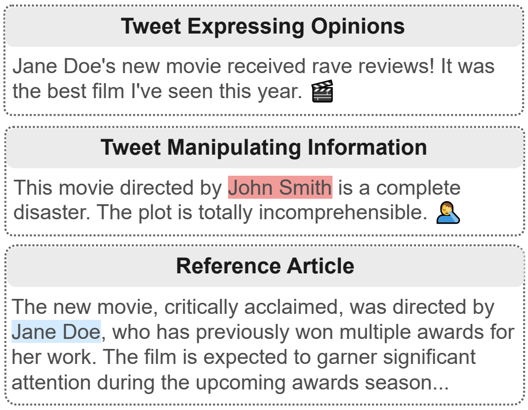

Figure 1: Two illustrative examples that highlight the challenge of identifying manipulation of news on social media. The first example expresses a personal opinion about watching a well-reviewed movie without distorting any facts from the associated article. Conversely, in the second example, the tweet falsely asserts that the movie is directed by John Smith instead of Jane Doe, thereby misrepresenting the information contained in the reference article. Hence, the second tweet misrepresents the information contained in the reference article.

1 Introduction

Detecting texts that contain misrepresentations of information originally presented in reference texts is crucial for combating misinformation. Previous research has primarily tackled this issue in the context of fact-checking (Thorne et al., 2018; Wadden et al., 2020), where the goal is to debunk unsupported claims using relevant passages, and in summarization (Kryscinski et al., 2020; Fabbri et al., 2022), where the focus is on assessing the faithfulness of generated summaries to the reference articles. However, none of the previous work has specifically addressed the identification of social media posts that manipulate information which was presented with a reference article from a news corpus. This poses a significant challenge due to *the prevalence of personal opinions in social media posts*. Our experiments demonstrate that state-of-the-art fact-checking and faithfulness assessment frameworks do not yield high performance in identifying social media posts that manipulate information (see §6). To effectively tackle this problem, models must be able to discern between personal opinions and sentences that distort information in social media posts. Examples of tweets that only express personal opinions and tweets that manipulate information can be found in Figure 1.

In this paper, we introduce a new task called *identifying manipulation of news on social media*. Given a social media post and its associated news article, models are tasked to understand whether and how the post manipulates information presented in the article. We define *manipulation* as cases where *a social media post intentionally misrepresents and distorts the content of the reference* *article*, following prior relevant studies (Shu et al., 2017; Fung et al., 2021). To explore this problem, we repurposed news articles from FakeNewsNet (Shu et al., 2020) and constructed a fully-annotated dataset, MANITWEET, consisting of 3.6K tweets accompanied by their corresponding news articles. To improve annotation cost-efficiency, we propose a two-stage data collection pipeline instead of naively requesting annotators to annotate a subset of human-written tweets from FAKENEWSNET. This approach tackles imbalanced tweet distributions, where the majority of tweets do not manipulate the associated article. It also addresses the challenge of verifying information between news articles and tweets, making the annotation process more efficient. In the first round, human annotators are assigned the task of validating tweets generated by large language models (LLMs) in a controllable manner. The data collected from these rounds is subsequently utilized to train a sequence-to-sequence model for identifying manipulation within tweets authored by humans. In the second round of annotation, these human-authored tweets are labeled accordingly. The 0.5K human-written tweets annotated in the second round are used as the test set for evaluation. Conversely, the 3.1K machine-generated tweets collected in the first round are used for our training and development set.

Our study aims to address three main research questions. First, we investigate the comparison between the fine-tuning paradigm and the in-context learning paradigm for this task. Using our curated dataset, we evaluate the performance of the fine-tuned sequence-to-sequence model discussed earlier in comparison to state-of-the-art LLMs. Surprisingly, we discover that our **much smaller fine-tuned model outperforms LLMs prompted with zero-shot or few-shot exemplars on the proposed task**. In fact, we find that LLMs do not achieve satisfactory performance on our task when only provided with a few exemplars. Second, we explore the impact of various attributes of a news article on its susceptibility to manipulation. To conduct this analysis, we employ the previously described sequence-to-sequence model to analyze a vast collection of over 1M tweets and their associated articles. Our findings reveal **a higher likelihood of manipulation in social media posts when the associated news articles exhibit low trustworthiness or pertain to political topics**. Finally, we investigate the role of manipulated sentences within a news article. To address this question, we perform discourse analysis on the test set of MANITWEET. Through this analysis, we uncover that **manipulated sentences within a news article often encompass the primary narrative or consequential aspects of the news article**.

Our contributions can be summarized as follows:

- We introduce and define the new task of identifying manipulation of news on social media.
- We propose a novel annotation scheme for this task. Using this scheme, we construct a dataset consisting of 3.6K samples, carefully annotated by human experts.
- We demonstrate that this dataset serves as a rigorous testbed for tackling identification of manipulation in social media. Specifically, we showcased the inadequate performance of LLMs in effectively addressing this challenge.
- Our proposed framework combines an LLM with a smaller fine-tuned model, utilizing opinion sentences extracted by the LLM as additional features. This achieves the best performance for our task.

2 Identifying Manipulation of News on Social Media

The goal of our task is to identify whether a social media post misrepresents information and what information is being manipulated given the associated reference article. Following prior work (Shu et al., 2017; Fung et al., 2021), we define the term *manipulation* as

Definition 1 *A social media post is deemed to manipulate information when it intentionally misrepresents and distorts the content of the reference article.*

The models are tasked to understand whether a tweet manipulates information in the reference article (§2.1), which newly introduced information in the tweet is used for manipulation (§2.2), and which original information in the reference article is manipulated (§2.3). In the following subsections, we provide detailed task formulation for each sub-task.

2.1 Sub-task 1: Tweet Manipulation Detection

Given a tweet and its associated news article, the first subtask is to classify the manipulation label l of this tweet, where $l \in \{\text{MANI}, \text{NOMANI}\}$. A tweet is considered MANI as long as there is at least one sentence that comments on the content of the associated article, and this sentence contains manipulated or inserted information. Otherwise, this tweet is NoMani.

2.2 Sub-task 2: Manipulating Span Localization

Once a tweet is classified as Mani, the next step is determining which information in the reference article was manipulated in the tweet. We refer to the information being manipulated as the *pristine span*, and the newly introduced information as the *manipulating span*. Both *pristine span* and *manipulating span* are represented as a text span in the reference article and the tweet, respectively. Identifying both information can help provide interpretability on model outputs and enable finer-grained analysis that provides more insights, as demonstrated in §6.2. Using Figure 1 as an example, the *manipulating span* is *John Smith*.

2.3 Sub-task 3: Pristine Span Localization

Similar to the second task, in this task, the model should output the *pristine span* that is being manipulated. In cases where the *manipulating span* is simply inserted, and no *pristine span* is manipulated, models should output a null span or an empty string. Using Figure 1 as an example, the *pristine span* is *Jane Doe*.

3 The ManiTweet Dataset

Our dataset consists of 3,636 tweets associated with 2,688 news articles. Each sample is annotated with (1) whether the tweet manipulates information presented in the associated news article, (2) which new information is being introduced, and (3) which information is being manipulated. We refer to this dataset as the ManiTweet dataset. An overview of the data curation process is shown in Figure 6. The following sections describe our corpus collection and annotation process.

3.1 News Article Source

To facilitate the analysis of human-written tweets, we created ManiTweet by repurposing a fake news detection dataset, FakeNewsNet (Shu et al., 2020). FakeNewsNet contains news articles from two fact-checking websites, PolitiFact and GossipCop, where each news article is annotated with a factuality label. In addition, for each news article, FakeNewsNet also consists of user engagement data, such as tweets, retweets, and likes, on Twitter. We reused the news content and the associated tweets from FakeNewsNet for our ManiTweet dataset.

During the early stage of the experiment, we observe that some news articles in FakeNewsNet are inappropriate for our study due to insufficient textual context. For example, some articles only contain a news title, a video, and a caption. To avoid such content, we remove news pieces containing less than 300 tokens.

3.2 Tweet Collection

Creating a high-quality dataset for our task using human annotators is extremely expensive and time-consuming primarily because the annotation task is challenging. Furthermore, real-world tweets authored by humans typically do not manipulate the associated articles. To address these issues, we have devised a two-stage pipeline to create training data. In the first round of annotation, we utilize ChatGPT[1] to generate both Mani and NoMani tweets in a controllable manner. Human annotators are then tasked with validating the generated tweets for their validity (§3.2.1). In the second round of annotation, we train a model on the data collected from the previous two rounds and employ this model to identify Mani human-written tweets for human annotation (§3.2.2). This approach ensures that annotators are not overwhelmed with a large number of NoMani tweets, resulting in significant improvements in time and cost efficiency compared to the aforementioned naive method.

3.2.1 Tweet Generation

We first used Stanza to extract Location, People, and Event named entities from all news articles. Then, we prompted ChatGPT to generate NoMani and Mani tweets for each news article. The span of these entities are denoted as $S = \{S_0, S_1, ..., S_n\}$. The prompts used for generating these tweets are as follows:

> **NoMani**: `This is a news article: NEWS_ARTICLE. Write a tweet that comments on this article. Keep it within 280 characters:`
>
> **Mani**: `This is a news article: NEWS_ARTICLE. Write a tweet that comments on this article but changes PRISTINE_SPAN to`

[1] `GPT-3.5-turbo`

```
NEW_SPAN and includes NEW_ENTITY
in your tweet. Keep it within 280
characters:
```

Here, PRISTINE_SPAN is a span randomly sampled from the spans of all named entities belonging to NEWS_ARTICLE , whereas NEW_SPAN is another span sampled from S with the same entity type as PRISTINE_SPAN. We have also experimented with other prompt templates. While the overall generation quality does not differ much, these prompt templates most effectively prevent ChatGPT from generating undesirable sequences such as "*As an AI language model, I cannot ...*".

In addition to generating MANI tweets where new information is manipulated from the original information contained in the associated article, we also produce MANI tweets where new information is simply inserted into the tweet using the following prompt:

```
This is a news article:
NEWS_ARTICLE. Summarize the
article into a tweet and comment
about it. Include NEW_SPAN in
your summarization but do not
include NEW_SPAN in the hashtag[2].
Keep it within 280 characters:
```

To further improve data quality and reduce costs in human validation, we only keep NOMANI tweets that contain at least one sentence inferrable from the corresponding article. Concretely, we use DocNLI (Yin et al., 2021), a document-level entailment model, to determine the entailment probability between the reference article and each tweet sentence. A valid consistent tweet must have at least one sentence with an entailment probability greater than 50%. Additionally, we remove MANI tweets that do not contain the corresponding NEW_SPAN specified in the corresponding prompts.

While we initially considered using various prompts to generate tweets in order to achieve greater diversity, our early experiments revealed that the resulting outputs did not exhibit significant variations in terms of styles and formats. Furthermore, ChatGPT possesses the capability to produce tweets with diverse styles even when the same prompt template is used. As a result, we have chosen to use a single prompt for all experiments.

[2] We instruct ChatGPT not to include NEW_SPAN in the hashtag. Otherwise, ChatGPT often does not insert NEW_SPAN into the main text of the tweet.

Split	# MANI	# NOMANI	# Doc	Tweet Author
Train	1,465	851	1,963	Machine
Dev	482	318	753	Machine
Test	294	226	299	Human

Table 1: Statistics of our MANITWEET dataset.

3.2.2 Our Proposed Annotation Process

We use Amazon's Mechanical Turk (AMT) to conduct annotation. Annotators were provided with a reference article and a corresponding generated tweet, along with labels indicating whether the tweet manipulates the article, and whether the predicted NEW_SPAN and PRISTINE_SPAN are accurate. In the first round of annotation, annotators were presented with tweets generated by ChatGPT. The labels for these tweets were naively derived from the data generation process, where we determined the manipulation label, NEW_SPAN, and PRISTINE_SPAN before prompting ChatGPT to generate a tweet. For efficient annotation, the annotators only need to validate whether the labels derived from the ChatGPT prompts are correct. We keep samples whose labels for all three sub-tasks are correct, while the others are discarded. In the second round of annotation, human-written tweets were annotated, and the predicted labels for these tweets were obtained from a model (see below paragraphs) trained on the data collected in the first annotation round. For detailed information regarding annotation guidelines and the user interface, please refer to Appendix D. The following paragraphs provide an overview of our annotation process.

First Round The first round of annotation is for curating machine-generated tweets, which are used as our training set and development set. Initially, for annotator qualification, three annotators worked on each of our HITs[3]. We used the first 100 HITs to train annotators by instructing them where their annotations were incorrect. Then, the next 100 HITs were used to compute the inter-annotator agreement (IAA). At this stage, we did not provide further instructions to the annotators. Using Fleiss' κ (Fleiss, 1971), we obtain an average IAA of 62.4% across all tasks, indicating a moderate level of agreement. Finally, we selected the top 15 performers as qualified annotators. These annotators were chosen based on how closely their annotations matched the majority vote for each HIT.

Since the annotators have already been trained,

[3] HIT refers to the Human Intelligence Task, which is the unit for an annotation task in Amazon Mechanical Turk.

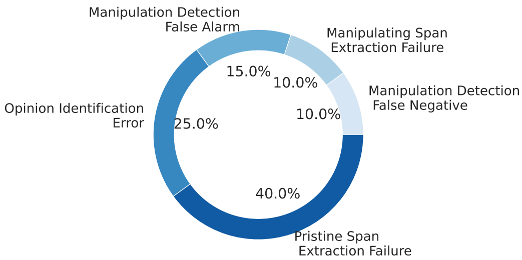

Figure 2: Distributions of errors. The error type definition is shown in Appendix H.

we assigned each HIT to a single annotator to improve annotation efficiency for the remainder of the machine-generated tweets. In addition to being annotated by an MTurk worker, each annotation is also re-validated by a graduate student. The average agreement between the graduate student and the MTurk worker is 93.1% per Cohen's κ (Cohen, 1960), implying a high agreement. We only keep samples where the validation done by the graduate student agrees with the annotation done by the worker. After two rounds of annotations, we collected 3,116 human-validated samples.

Second Round Using the 3K examples we collected, we train a sequence-to-sequence (seq2seq) model that learns to tackle all three tasks jointly. Concretely, we split the collected data into 2,316: 800 for training and validation. Model details are described in the next paragraph. Once the model was trained, we applied it to identify manipulation in the human-written tweets that are associated with the articles in FakeNewsNet. Then, we randomly sampled from predicted MANI and NOMANI examples to be further validated by MTurk workers. The inter-annotator agreement between the graduate student and the MTurk worker is 73.0% per Cohen's κ (Cohen, 1960). While the agreement is moderately high, it is much lower than that in the previous round. This suggests that manipulation in human-written tweets is more challenging to identify. The user interface of each round of annotation is shown in Appendix D.1. Finally, we have curated the MANITWEET dataset. The dataset statistics are shown in Table 1.

Baseline Model In this paragraph, we describe the model we used to facilitate the second round of annotation. Motivated by the advantages of generative models over sequence-tagging models (Li et al., 2021; Huang et al., 2021; Hsu et al., 2022), we trained a seq2seq model based on LongFormer-

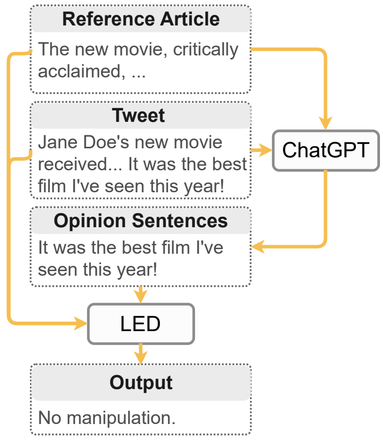

Figure 3: An overview of the proposed framework, **LLM + LED-FT**. We first use ChatGPT to identify sentences that express opinions from the tweet. Then, the opinion sentences are fed to a LED as additional features to help discern between sentences that express personal opinions and sentences that manipulates information.

Encoder-Decoder (LED)[4] (Beltagy et al., 2020) that learns to solve the three tasks jointly. We name this model **LED-FT**.

Formally, the input $x = [t||a]$ to our model is the concatenation of a tweet t and the corresponding article a. The objective of the model is maximum likelihood estimation,

$$\mathcal{L} = -\sum_i p(y_i | y_{<i}, x), \tag{1}$$

where y_i denotes the i-th token in the decoding targets. Concretely, if the article is NOMANI, the model should output "`No manipulation`". Otherwise, the model should output "**`Manipulating span`**: `NEW_SPAN` \ **`Pristine span`**: `PRISTINE_SPAN`". For cases where `NEW_SPAN` is merely inserted into the tweet, the model will output "None" for `PRISTINE_SPAN`. Details of inputs, outputs, and training hyper-parameters can be found in Appendix B.

4 Methodology

We conducted an error analysis on the **LED-FT** model discussed in the previous section. Our analysis revealed that a significant portion of errors occurred due to the model's inability to distinguish between tweet sentences that express personal opinions and those that manipulate information from the associated article, as depicted in Figure 2 (refer to Appendix C for further details). To address this issue, we propose a pipeline approach that involves

[4] `https://huggingface.co/allenai/led-base-16384`

Model	Learning Method	Sub-task 1	Sub-task 2			Sub-task 3		
		F1	EM	F1	RL	EM	F1	RL
Human	-	89.92	44.23	67.93	68.82	42.88	65.29	66.31
Vicuna	Zero-shot	47.09	1.35	5.11	6.07	4.04	6.21	7.06
ChatGPT	Zero-shot	52.49	1.54	13.30	15.96	4.42	7.46	8.35
ChatGPT	Two-shot ICL	65.28	0.96	7.62	8.87	12.50	13.91	14.18
ChatGPT	Four-shot ICL	54.69	3.07	12.79	15.15	1.54	4.99	5.95
ChatGPT	Two-shot CoT	52.92	1.54	7.70	9.21	4.42	5.86	6.12
ChatGPT	Four-shot CoT	53.88	0.96	7.93	9.66	3.46	5.24	5.70
CONCRETE	Zero-shot	57.88	-	-	-	-	-	-
DocNLI	Zero-shot	62.26	-	-	-	-	-	-
QAFactEval	Zero-shot	62.56	-	-	-	-	-	-
LED-FT (Ours)	Fine-tuned	72.62*	26.73*	29.25*	29.68*	13.65*	14.46	14.53
LLM + LED-FT (Ours)	Zero-shot + Fine-tuned	**73.46***	**28.85***	**31.72***	**32.32***	**15.19***	**16.21***	**16.41***

Table 2: Performance (%) of different models on the MANITWEET test set. EM denotes Exact Match, and RL denotes ROUGE-L. Statistical significance over best-performing LLMs computed with the paired bootstrap procedure (Berg-Kirkpatrick et al., 2012) are indicated with * ($p < .01$).

utilizing ChatGPT to identify personal opinions within the tweet. This extracted opinions is then incorporated into our seq2seq model during both training and testing stages. An overview of the framework is shown in Figure 3.

More specifically, we denote the identified opinion sentences in the tweet t as $o = p_{\text{LLM}}(t, a, d)$, where d represents the instruction provided to ChatGPT for opinion identification. The input to our fine-tuned model becomes $x' = [t||a||o]$, and the loss function remains as MLE:

$$\mathcal{L}' = -\sum_i p(y_i|y_{<i}, x'). \quad (2)$$

By incorporating this framework, we aim to enhance the model's ability to differentiate between personal opinions and instances where information is manipulated from the associated article. We name this pipeline **LLM + LED-FT**.

5 Experimental Setup

5.1 Evaluation Metrics

Subtask 1 involves a binary classification problem, and thus, the Macro F1 score serves as the evaluation metric. For subtasks 2 and 3, in addition to Exact Match, we use Macro Overlap F1 score (Rajpurkar et al., 2016) and ROUGE-L (Lin, 2004) as the metrics to more accurately assess model performance by allowing models to receive partial credit for correctly identifying some parts of the information, even if they fail to output the entire text span.

5.2 Baselines

We compare our proposed framework with various recently released large language models (LLMs), including Vicuna[5] (vic, 2023) and ChatGPT, which have demonstrated superior language understanding and reasoning capabilities. ChatGPT is an improved version of InstructGPT (Ouyang et al., 2022) that was optimized for generating conversational responses. On the other hand, Vicuna is a LLaMA model (Touvron et al., 2023) fine-tuned on ShareGPT[6] data, and has exhibited advantages compared to other open-source LLMs, such as LLaMA and Alpaca (Taori et al., 2023). We tested the zero-shot, two-shot, and four-shot performance of ChatGPT in both in-context learning (ICL) and chain-of-thought (CoT) (Wei et al., 2022) settngs , where the in-context exemplars are randomly chosen from our training set. For Vicuna, we only evaluated its zero-shot ability as we found that it often outputs undesirable texts when exemplars are provided. The details of our prompts for these LLMs can be found in Appendix E. In addition, we also evaluate one fact-checking framework, CONCRETE (Huang et al., 2022), and two faithfulness evaluation frameworks, QAFactEval (Fabbri et al., 2022) and DocNLI (Yin et al., 2021) on our subtask 1. Similar to previous studies, we establish the faithfulness thresholds for both frameworks by selecting the values that yield the highest performance on our development set.

6 Results

6.1 Performance on MANITWEET

Table 2 presents a summary of the main findings from our evaluation on the MANITWEET test set. We have made several interesting observations:

[5]Vicuna-13b is evaluated in our experiment.
[6]https://sharegpt.com/

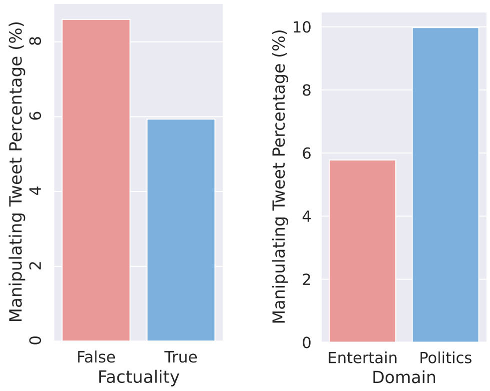

Figure 4: The percentage of tweets that manipulate the associated articles across different levels of factuality and domains.

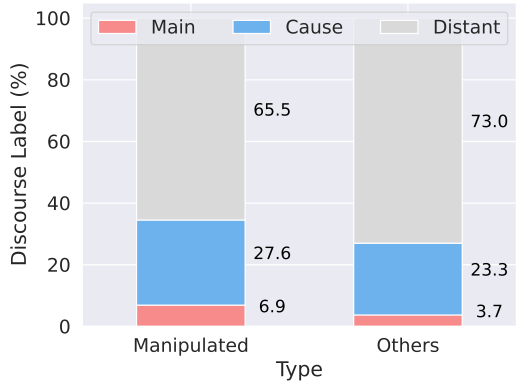

Figure 5: Results of discourse analysis. Manipulated sentences within news articles tend to encompass the main story (*Main*) or convey the consequential aspects (*Cause*) of the corresponding news story.

First, all LLMs we tested performed poorly across the three proposed tasks. This indicates that simply prompting LLMs, whether with or without exemplars, is not sufficient to effectively address the problem of identifying manipulation of news on social media. We also found that providing more exemplars do not work well on our task as the performance drop when we increase the number of in-context exemplars from 2 to 4. This is likely caused by the long-context nature of our task. Indeed, the average number of tokens per article is 2609.6 in the test set. Secondly, despite its simplicity and smaller size compared to the LLMs, **LED-FT** outperforms all baseline models significantly in identifying social media manipulation across all three tasks. This outcome highlights the value and importance of our training data and suggests that a fine-tuned smaller model can outshine larger models when tackling challenging tasks. Finally, the proposed **LLM + LED-FT** outperforms all other models, including **LED-FT** significantly. This implies that LLMs can complement smaller fine-tuned models by identifying opinions and that the ability to identify opinion sentences from social media posts is critical for our task. Examples of how the opinions extracted by ChatGPT help correct errors can be found in Appendix F.

In order to gauge the feasibility of the task, we enlisted the assistance of a graduate student to tackle our test set. While this may not necessarily represent the upper bound of performance, it provides a preliminary approximation of human performance. As depicted in Table 2, there remains a discernible gap between **LLM + LED-FT** and human performance. This highlights great opportunities in our task for future research.

6.2 Exploratory Analysis

The proposed **LED-FT** model enables us to perform a large-scale study of manipulation on the MANITWEET test set and the 1M human-authored tweets associated with the news articles from the FakeNewsNet dataset. In this section, we explore how an article is MANI and how different properties of a news article, such as domain and factuality affect manipulation.

Insight 1: Low-trustworthiness and political news are more likely to be manipulated. Figure 4 shows the percentage of the 1M human-written tweets that are manipulated across 2 domains and factuality levels.[7] We first observe that tweets associated with *False* news are more likely to be manipulated. One possible explanation is that audience of low-trustworthy news media may pay less attention to facts. Hence, they are more likely to manipulate information from the reference article accidentally when posting tweets. In addition, we also see that tweets associated with *Politics* news are more frequently manipulated than those with *Entertainment* articles. This could be explained by the fact that people have a stronger incentive to manipulate information for political tweets due to elections or campaigns.

Insight 2: Manipulated sentences are more likely to contain the main story or consequence of a news story. To discover the role of the sentence being manipulated in the reference article, we conducted discourse analysis on these sentences. We only conducted the analysis on our test set instead of the entire 1M human-written

[7]The domain and factuality labels of each news article are already annotated in the FakeNewsNet dataset.

tweets for this analysis. Concretely, we formulate the discourse classification task as a sequence-to-sequence problem and train a LED-based model on the NewsDiscourse dataset (Choubey et al., 2020) using a similar strategy discussed in §3.2.2. The learned discourse classification model achieves a Micro F1 score of 67.7%, which is on par with the state-of-the-art method (Spangher et al., 2021). Upon the discourse classification model being trained, we applied it to all the sentences in the reference article to analyze the discourse distribution. As shown in Figure 5, compared to other sentences, sentences that were manipulated are much more likely to contain *Main* or *Cause* discourse, which corresponds to *the primary topic being discussed* and *the underlying factor that led to a particular situation*, respectively. Examples of the manipulated sentences with a *Main* or *Cause* discourse can be found in Appendix G.

7 Related Work

7.1 Faithfulness

Faithfulness is often referred to as the factual consistency between the inputs and outputs. This topic has mainly been studied in the field of summarization. Prior work on faithfulness can be divided into two categories: evaluation and enhancement, the former of which is more relevant to our study. One line of faithfulness evaluation work developed entailment-based metrics by training document-sentence entailment models on synthetic data (Kryscinski et al., 2020; Yin et al., 2021) or using traditional natural language inference (NLI) models at the sentence level (Laban et al., 2022). Another line of studies evaluates faithfulness by comparing information units extracted from the summaries and input sources using QA (Wang et al., 2020; Deutsch et al., 2021; Fabbri et al., 2022).

Our task differs from faithfulness evaluation in two key ways. Firstly, for our task to be completed effectively, models must possess the additional capability of distinguishing tweet sentences that relate to the reference article from those that simply express opinions. In contrast, models evaluating faithfulness only need to identify whether each sentence in the output is inferable from the input. Secondly, we require models to not only identify which original information is being manipulated by the new information, but also to provide interpretability as to why a tweet has been manipulated.

7.2 Fact-checking

Fact-checking is a task that determines the veracity of an input claim based on some evidence passages. Some work assumes the evidence candidates are provided, such as in the FEVER dataset (Thorne et al., 2018) and the SciFact dataset (Wadden et al., 2020). Approaches for this category of fact-checking tasks often involve a retrieval module to retrieve relevant evidence from the given candidate pool, followed by a reasoning component that determines the compatibility between a piece of evidence and the input claim (Yin and Roth, 2018; Pradeep et al., 2021). Other work focuses on the *open-retrieval* setting, where evidence candidates are not provided, such as in the Liar dataset (Wang, 2017) and the X-Fact dataset (Gupta and Srikumar, 2021). For this task formulation, one of the main challenges is to determine where and how to retrieve evidence. Some approaches determine the veracity of a claim based solely on the claim itself and the information learned by language models during the pre-training stage (Lee et al., 2021), other methods leverage a retrieval module to look for evidence on the internet (Gupta and Srikumar, 2021) or a set of trustworthy sources (Huang et al., 2022). Similar to the faithfulness task, the key distinction between fact-checking and our proposed task lies in the additional requirement for models to possess the capability of discerning between tweet sentences that pertain to the reference article and those that merely express opinions.

8 Conclusion

In this study, we have introduced and defined a novel task called *identifying manipulation of news on social media*, which aims to determine whether and how a social media post manipulates the associated news article. To address this challenge, we meticulously collected a dataset named ManiTweet, composed of both human-written and machine-generated tweets. Our analysis revealed that existing large language models (LLMs) prompted with zero-shot and two-shot exemplars do not yield satisfactory performance on our dataset, highlighting avenues for future research. We believe that the resources presented in this paper can serve as valuable assets in combating the dissemination of false information on social media, particularly in tackling the issue of news manipulation.

9 Limitations

Using LLMs for data creation. LLMs, such as ChatGPT, are instrumental in crafting entire tweets that are not only coherent but also conditioned on the specifics of the given news article, ensuring a level of fluency that mimics that of human writers. Moreover, the tweets fashioned by ChatGPT showcase a distinct superiority in quality when compared to more traditional methods of data synthesis, such as those that are rule-based or template-based. These earlier approaches often resulted in output that was both stilted and monotonous, falling short in fluency and variety, a fact substantiated by references(Goyal and Durrett, 2021; Utama et al., 2022). By leveraging the capabilities of ChatGPT, we can generate machine-authored tweets that not only boast a broad diversity but also maintain a convincingly realistic quality, thereby providing an enriched dataset for scalable human annotation.

LLM prompts. In our experiments involving prompting LLMs, we only explored ICL and CoT for prompting LLMs. There is a possibility that LLMs can achieve better performance when provided with more in-context exemplars and when prompted in a more refined manner.

10 Ethical Considerations

The primary ethical consideration in our work pertains to the presence of false information in two aspects: tweets that manipulate the associated news articles and the inclusion of false news from the FakeNewsNet dataset. As with other fact-checking and fake news detection research, it is important to acknowledge the dual-use concerns associated with the resources presented in this work. While our resources can contribute to combating false information, they also possess the potential for misuse. For instance, there is a risk that malicious users could utilize the manipulating tweets or fake news articles to train a text generator for creating deceptive content. We highlight appropriate and inappropriate uses of our dataset in various scenarios:

- **Appropriate**: Researchers can use our framework to study the manipulation issue on social media and develop stronger models for identifying social media posts that manipulate information.
- **Inappropriate**: The fake news and manipulating tweets in MANITWEET cannot be used to train text generators for malicious purposes.
- **Inappropriate**: Use the manipulation prompts discussed in this paper to generate tweets and spread false information.
- **Inappropriate**: The fake news in MANITWEET should not be used as evidence for fact-checking claims.

Furthermore, the privacy of tweet users is another aspect that warrants consideration, given that we are releasing human-written tweets. However, we assure that the dataset does not pose significant privacy concerns. The tweets in our dataset are anonymized, and it is important to note that all the associated news articles were already publicly available. Therefore, the release of this dataset should not have adverse implications for privacy.

References

2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality.

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

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of* *the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1621–1630, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 230–239.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

A Additional Discussios

If real-world tweets typically do not manipulate associated articles (§3), how practical and relevant is the proposed task? While manipulated tweets that distort information from news articles may not be extremely common on social media, they can still have an outsized impact when they do occur. Even a small number of tweets that deliberately misrepresent the facts around a news story have the potential to spread wildly on social media and shape public discourse (Allcott and Gentzkow, 2017; Starbird, 2017). We would argue that the harm caused by manipulated tweets warrants research efforts into detecting and combating them, even if the absolute number of such tweets is low. A few viral manipulated tweets can still reach millions of users and significantly skewed perceptions around news events and issues. Identifying and fact-checking these tweets is key to limiting the spread of misinformation.

Discrepancies between the training set and the test set. Despite our best efforts to minimize the gap between the training set and test set of MANITWEET, some discrepancies remain due to the training set being generated by machines and the test set being produced by humans. This limitation is primarily attributed to budget constraints. In fact, synthetically generating training data is a common strategy in relevant fields where extensive human annotation poses significant challenges, such as fake news detection (Huang et al., 2023; Fung et al., 2021) and factual inconsistency detection (Kryscinski et al., 2020; Utama et al., 2022). In the future, with additional resources, we aim to create an additional training set consisting entirely of human-written tweets. By comparing the performance of models trained on this human-written training set with those trained on the machine-generated training set, we can gain further insights. However, we wanted to emphasize that our test set exclusively consists of tweets authored by humans, which ensures the relevance of our techniques and dataset for real-world applications in handling tweets produced by actual Twitter users. While our data collection method may introduce discrepancies in the distribution between the training and test sets, the fundamental purpose of our dataset remains consistent: to investigate the manipulation of news articles on social media.

Manipulation types. Our approach focuses on manipulations of three types of entities: LOCATION, PEOPLE, and EVENT. This approach may fail in cases where the manipulation is complex, beyond entity-level perturbations or involving multiple entities. However, it is important to highlight that following a meticulous examination of 100 manipulated examples from our dataset, we found that an overwhelming **85% of them involve named entity manipulations only**. Through this analysis, we categorized manipulations based on their intent and the nature of the information distortion, identifying three additional manipulation types in addition to entity-level manipulation:

- Misattribution of Quotes or Actions (10%): Where social media posts attribute incorrect quotes or actions to individuals or entities not associated with them in the referenced news articles.
- Exaggeration/Understatement (3%): Manipulations that inflate or diminish the severity or importance of the facts presented in the articles.
- Temporal Distortion (2%): Tweets misleadingly suggest that certain events happened at a different time than reported in the article, affecting the perceived relevance or cause-effect relationships.

Based on this analysis, we have established stronger support for our claim in the paper and enriched our understanding of various manipulation types for future research. This highlights that our formulation is still relevant and can handle the vast majority of real-world manipulations.

How `PRISTINE_SPAN` is mapped to `NEW_SPAN`? `PRISTINE_SPAN` refers to a text span within the reference article that is associated with a particular named entity and is relevant to the news narrative. `NEW_SPAN`, on the other hand, is a different text span associated with the same type of entity but is randomly sampled from the set of all named entities extracted from the news articles.

The intention behind replacing `PRISTINE_SPAN` with `NEW_SPAN` is to create a manipulated piece of text by altering entity-related information found in the original article. By ensuring that the `NEW_SPAN` shares the same entity type as the `PRISTINE_SPAN`, we maintain the semantic plausibility of the generated tweet.

For example, consider the following:

> **Reference Article**: "President Smith advocated for environmental policies in the recent summit held in Geneva, emphasizing the need for sustainable development." (PRISTINE_SPAN: "President Smith")

By extracting named entities, we might get a list like ["President Smith", "Geneva", "Prime Minister Johnson", "Paris"]. Suppose we choose "Prime Minister Johnson" as the NEW_SPAN to replace "President Smith". The manipulating tweet could then be:

> **Manipulating Tweet**: "Prime Minister Johnson pushed for new economic measures in the conference that took place in Paris, expressing urgency for financial reform." (NEW_SPAN: "Prime Minister Johnson")

Here, the NEW_SPAN provides alternative, yet topically coherent, entities to create misinformation while preserving the sentence structure and general subject matter of the original article.

The prompts given to ChatGPT are pretty lengthy and may not be well articulated to the desired answers, and more shots given even result in worse performance. Our study aimed to explore the baseline effectiveness of LLMs such as ChatGPT and Vicuna in the task of identifying news manipulation on social media without extensive prompt engineering. This choice was deliberate to mirror a more *generalizable and accessible* use case, where users of varying technical backgrounds rely on LLMs.

The prompts were carefully designed to reflect the task's complexity, ensuring clarity in instructions to produce relevant and accurate responses. Our aim was not to maximize the performance through prompt engineering but to **establish a fundamental understanding of LLM capabilities in this novel task domain under relatively straightforward conditions**.

To clarify, the drop in performance with more in-context examples suggests that this task likely requires additional abilities beyond simply providing more examples, which is an insightful result in itself, indicating areas for future research in improving LLMs' handling of complex and long-context relations in texts.

Model	Person (%)	Location (%)	Event (%)
ChatGPT Two-shot ICL	64.5	58.68	68.14
LED-FT (Ours)	71.01	66.46	73.16
LLM + LED-FT (Ours)	73.21	72.21	72.33

Table 3: Breakdown F1 scores w.r.t. different entity types.

Model	Prompts	Sub-task 1	Sub-task 2	Sub-task 3
GPT-4	Zero-shot	70.23	22.92	10.56
GPT-4 Turbo	Zero-shot	72.21	19.56	12.43

Table 4: F1 scores of GPT-4 and GPT-4 Turbo on the MANITWEET test set.

Is it true that the unsatisfying performance of LLMs is due to the capability of the language model or the prompt engineering? We tested models that have stronger long-context reasoning ability, such as GPT-4 (with a context window of 8K tokens). If these models show increased performance compared to ChatGPT and Vicuna, we can better conclude that the poor performance of ChatGPT and Vicuna is caused by their insufficient long-context reasoning abilities. In Table 4, we show the performance of GPT-4 and GPT-4 Turbo on our task. Based on our findings, we can confirm that models with stronger long context reasoning ability are better at identifying manipulating tweets as well as manipulated and inserted information. This validates our hypothesis that the poor performance of ChatGPT and Vicuna is caused by the long-context nature of our task and their limited ability in modeling long-form texts.

Are some entities more difficult to identify than others? We ran an additional analysis to understand the performance breakdown for each error type. The results are summarized in the Table 3.

Overall, we can see that manipulation of location-related entities is the most challenging to identify. We also found that by utilizing opinion sentences identified by LLM, we achieve significant performance gain on manipulations involving Person and Location entities. This highlights the effectiveness of the proposed framework.

B Training Details

B.1 LED-based Fine-tuned Model

The input to our LED-based model is a concatenation of a tweet and a reference article:

```
Tweet: TWEET \
Reference article: REF_ARTICLE
```

If the article is NoMani, the model should output:

```
No manipulation
```

Otherwise, the model should output the following:

```
Manipulating span: NEW_SPAN \
Pristine span:
PRISTINE_SPAN
```

For cases where NEW_SPAN is merely inserted into the tweet, the model will output "None" for PRISTINE_SPAN. Using this formulation, our model is learned to optimize the maximum likelihood estimation loss. We set identical weights for all tokens in the outputs.

B.2 ChatGPT Prompts

The prompt to ChatGPT for identifying opinions is as follows:

> **Tweet**: TWEET \
> **Reference article**: REF_ARTICLE
> Given the above tweet and article. List the sentences in the tweet that merely express opinions instead of manipulating information from the article. If there is none, answer "None". Do not provide explanations.

B.3 Training Hyper-parameters

To learn the model, we use a learning rate of 5e-5. The maximum input and output sequence length are 1024 and 32 tokens, respectively. The model is optimized using the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 4 and a gradient accumulation of 8. During inference time, we use beam search as the decoding method with a beam width of 4.

B.4 Training Discourse Analysis Model

For this discourse analysis model, the input is a concatenation of the reference article and a sentence from the same reference article, while the output is one of the discourse labels defined in News-Discourse. We then compare the discourse label distribution for sentences that contain text span (PRISTINE_SPAN) that are manipulated by a tweet versus that for other sentences, as shown in Figure 5.

C Error Analysis

To gain insights into the additional modeling and reasoning capabilities required for effectively addressing the task of social media manipulation, we manually compare 50 errors made by the LED-based model with ground-truth labels and analyze the sources of errors. The distribution of errors is illustrated in Figure 2. Notably, the most prevalent error arises from the model's inability to extract the correct pristine span from the reference article that underwent manipulation. Among the 18 erroneous predictions in this category, 16 cases result from the model producing an empty string. This indicates that the model considers the manipulating information to be inserted when, in reality, it is manipulated from the information present in the reference articles. This could be attributed to the presence of 368 instances where the original information is an empty string, while the alternative answers for the original information only occur 1-2 times in other instances. This can be solved by scaling down the loss for these samples with an empty string as the label for original information. Additionally, another common type of error involves the model's failure to identify opinions expressed in the tweet. In these instances, the model considers the tweet to be manipulating information from the article, whereas the tweet primarily expresses opinions. Examples of these errors are presented in Appendix F.

D Annotation Details

In this section, we describe the details of our annotation process. We show an overview of our data curation process in Figure 6. For better control of the annotation quality, we required that all annotators be from the U.S. and have completed at least 10,000 HITs with 99% acceptance on previous HITs. The reward for each HIT is $1 U.S. dollar, complying with the ethical research standards outlined by AMT (Salehi et al., 2015). Annotation interfaces are shown below.

D.1 User Interface

Figure 7 and Figure 8 display the annotation interface for the first round and the third round of annotation, respectively. The only difference is that for the second round of annotation, we asked annotators to correct errors made by our basic model discussed in §3.2.2. Samples that do not receive "yes" on all three questions for the first round of

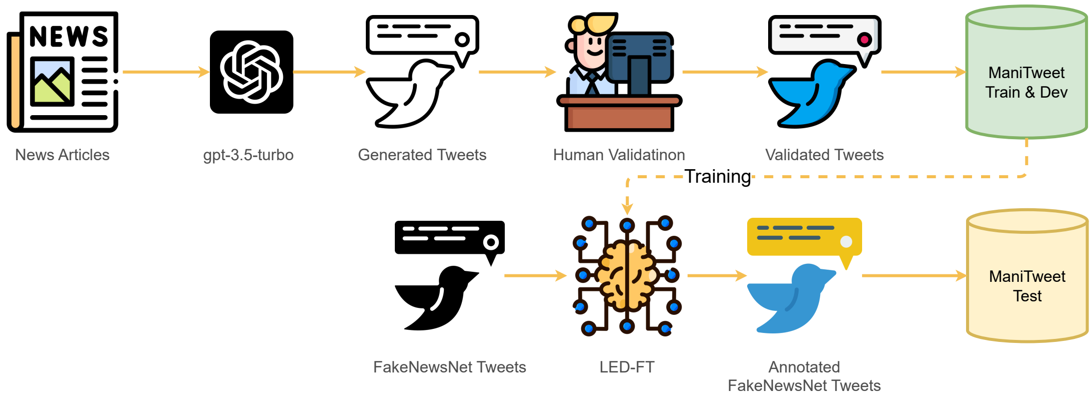

Figure 6: An overview of our data curation process.

annotation will be discarded. The rationale behind this design stems from three key reasons: Firstly, the data for the first round of annotation is automatically generated, enabling a relatively cost-effective approach to discard invalid samples and generate new ones, as opposed to requesting annotators to correct errors. Secondly, the data generated in these two rounds is predominantly valid, which eliminates the need for annotators to rectify errors and consequently accelerates the annotation process. Lastly, in the second round of annotation, by instructing annotators to identify errors made by our model, we can effectively identify the challenges faced by the model.

E Prompts for LLMs

The zero-shot and two-shot prompt template to LLMs for the experiments discussed in §5.2 is shown in Table 6. The in-context exemplars for the two-shot experiments are randomly sampled from the training set of MANITWEET.

F Additional Qualitative Examples

Table 7 presents two instances where our baseline model makes errors. In the first example, our model was not able to identify that "Inspired Our Next Trip To The Salon" is an expression of opinion, resulting in the model incorrectly classifying this sample as MANI. In the second example, although our model accurately predicts the example as MANI and extracts the correct manipulating span, it fails to extract the pristine text span correctly, likely due to the nature of the training set, as discussed in Appendix C.

Table 8 shows an example where extracting opinion sentences from the tweet by ChatGPT enables

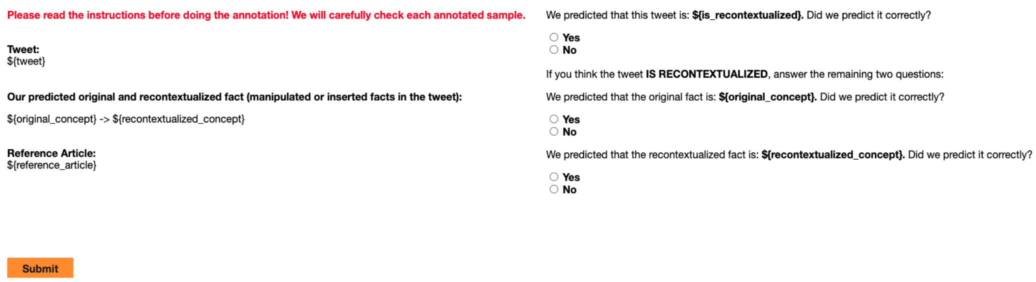

Figure 7: MTurk user interface for the first round of data annotation.

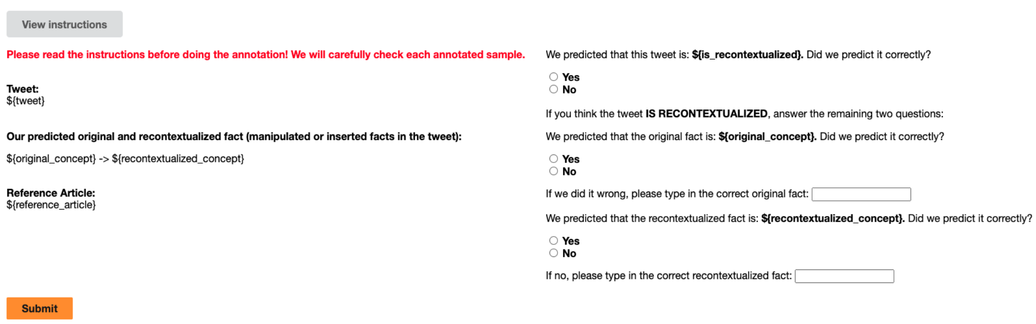

Figure 8: MTurk user interface for the second round of data annotation.

Role	Utterance
User	You are tackling a social manipulation problem. You will be giving a tweet and an article, and your task is to identify which information from the article is misrepresented by which information in the tweet. You should answer in the following format “Manipulating span: manipulating_span Pristine span: pristine_span” in a single line. Here, {manipulating_span} is the new information introduced in the tweet and original_concept is the original information in the article. If the tweet simply inserts information, {original_concept} should be "None". If the tweet does not manipulate the article, answer “No manipulation”. You do not need to output other information such as an explanation. You don’t need to provide code. In the following utterances, you will be presented a pair of tweet and news article.
LLM	I understand the task. Please provide me with the tweet and the corresponding article, and I will identify the manipulation.
User	Tweet: Sad news for Ben Affleck and Lindsay Shookus as they split after a year of dating. It looks like Ben has been getting cozy with Playboy model Shauna Sexton, but sources say it’s just light and fun. Best of luck to all involved. #BenAffleck #LindsayShookus #ShaunaSexton Article: Ben Affleck and Lindsay Shookus Split After a Year of Dating Ben Affleck and Lindsay Shookus have broken up after over a year of dating, a source tells ET... == For this example, you should answer “No manipulation”
LLM	No manipulation
User	Tweet: Meghan’s Lush Malibu Home Worth $16M, which comes with 8 bedrooms, 8 bathrooms, a private beach, a pool, a home cinema - they even want to spend an extra $1 million to decorate it to their own taste. Unfortunately Caboolture scores 0 here! Article: Take A Look Inside Prince Harry And Meghan Markle’s Lush Malibu Home If you’re set to marry royalty, like Meghan Markle you would expect to move into a palace, and live out your Disney Princess fantasy... == For this example, you should answer “Manipulating span: Caboolture Pristine span: None”
LLM	Manipulating span: Caboolture Pristine span: None
User	Tweet: *Input Tweet* Article: *Input Article*

Table 5: The zero-shot and two-shot prompt template to the LLMs for producing the results in Table 2. Texts marked in gray indicate that they are only used in the two-shot experiments.

our model to correctly identify the tweet as not manipulating the associated article.

G Discourse Analysis Examples

Table 9 shows examples of manipulated sentences associated with a *Main* or *Cause* discourse. A *main* discourse implies that the sentence conveys the main story of an article, whereas a *cause* discourse indicates that the sentences discuss the consequential aspect of the main story.

H Error Type Definition

In this section, we provide illustrations for each error type:

- **Opinion Identification Error**: The tweet predicts that a tweet manipulates the reference article. However, the manipulating span predicted by the model is in fact merely opinions and not trying to manipulate the content. An example is shown in Table 4 in the appendix. It is true that no annotator has specified the ground truth for opinion sentences. All the error analyses were performed manually by the authors.

- **Manipulation Detection False Alarm**: This is effectively the “Manipulation Detection False Positive” in which the model predicts a tweet manipulates the reference article but the label is NoMani (no manipulation). Note that “Opinion Identification Error” is considered a special case of “Manipulation Detection False Alarm” where the manipulating span overlaps with opinions expressed by the tweet author.

- **Manipulation Detection False Negative**: The model predicts there is on manipulation within a tweet but the label is Mani (manipulating).

- **Manipulating Span Extraction Failure**: The model successfully predicts the manipulation label for a manipulating tweet but fails to iden-

Role	Utterance
User	You are tackling a social manipulation problem. You will be giving a tweet and an article, and your task is to identify which information from the article is misrepresented by which information in the tweet. You should answer in the following format: "Opinion sentences: opinion_sentences Rationale: rationale Manipulating span: manipulating_span Pristine span: pristine_span". Here, {opinion_sentences} is the sentences in the tweet that express opinions, {rationale} is your reasoning process that determines whether the tweet manipulates the article, {manipulating_span} is the new information introduced in the tweet and {original_concept} is the original information in the article. If the tweet simply inserts information, {original_concept} should be "None". If the tweet does not manipulate the article, answer "No manipulation". You do not need to output other information such as an explanation. The tweet does not manipulate the article if it is simply expressing opinions. In the following utterances, you will be presented a pair of tweets and news articles.
LLM	I understand the task. Please provide me with the tweet and the corresponding article, and I will identify the manipulation.
User	Tweet: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon ... Article: Ariana Grande Dyed Her Hair, And This Is Our Favorite Color Transformation Yet Ariana Grande is giving us whiplash with her hairstyles lately, and we honestly love it. On July 18th, Grande took to Instagram to debut her latest hair transformation. She's now sporting pastel lavender locks and good god (is a woman), it looks amazing on her... === For this example, you should answer "Opinion sentences: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon": Rationale: The tweet is only expressing opinion and therefore there is no manipulation. Manipulating span: No manipulation. Pristine span: No manipulation"
LLM	Opinion sentences: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon": Rationale: The tweet is only expressing opinion and therefore there is no manipulation. Manipulating span: No manipulation. Pristine span: No manipulation
User	Tweet: *Input Tweet* Article: *Input Article*

Table 6: The one-shot CoT prompt template to the LLMs. For two-shot and four-shot prompts, the prompt includes more examples.

tify the specific text spans that manipulate the content of the reference article.

- **Pristine Span Extraction Failure**: The model successfully predicts the manipulation label for a manipulating tweet but fails to identify the specific text span from the reference article that was manipulated.

Opinion Identification Error	
Input	**Tweet**: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon ... **Article**: Ariana Grande Dyed Her Hair, And This Is Our Favorite Color Transformation Yet Ariana Grande is giving us whiplash with her hairstyles lately, and we honestly love it. On July 18th, Grande took to Instagram to debut her latest hair transformation. She's now sporting pastel lavender locks and good god (is a woman), it looks amazing on her...
Prediction	**Is manipulated**: Yes ✗ **Manipulating span**: Salon ✗ **Pristine span**: None
Pristine Span Extraction Failure	
Input	**Tweet**: Transcript: Democratic Presidential Debate in Brooklyn view more ... **Article**: The Democratic Debate in Cleveland This is rightly a big issue in Ohio. And I have laid out my criticism, but in addition my plan, for actually fixing NAFTA. Again, I have received a lot of incoming criticism from Senator Obama. And the Cleveland Plain Dealer examined Senator Obama's attacks on me regarding NAFTA and said they were erroneous. So I would hope that, again, we can get to a debate about what the real issues are and where we stand because we do need to fix NAFTA. It is not working. It was, unfortunately, heavily disadvantaging many of our industries, particularly manufacturing. ...
Prediction	**Is manipulated**: Yes **Manipulating span**: Brooklyn **Pristine span**: None ✗

Table 7: Example outputs from our baseline model where it produces erroneous outputs.

Input	**Tweet**: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon ... **Article**: Ariana Grande Dyed Her Hair, And This Is Our Favorite Color Transformation Yet Ariana Grande is giving us whiplash with her hairstyles lately, and we honestly love it. On July 18th, Grande took to Instagram to debut her latest hair transformation. She's now sporting pastel lavender locks and good god (is a woman), it looks amazing on her...
Prediction	Is manipulated: Yes ✗ Manipulating span: Salon ✗ Pristine span: None
Input	**Tweet**: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon ... **Predicted Opinions**: Ariana Grande's New Lavender Hair Color Just Inspired Our Next Trip To The Salon **Article**: Ariana Grande Dyed Her Hair, And This Is Our Favorite Color Transformation Yet Ariana Grande is giving us whiplash with her hairstyles lately, and we honestly love it. On July 18th, Grande took to Instagram to debut her latest hair transformation. She's now sporting pastel lavender locks and good god (is a woman), it looks amazing on her...
Prediction	**Is manipulated**: No ✓ **Manipulating span**: None ✓ **Pristine span**: None

Table 8: Example outputs from our LED-FT and LLM + LED-FT. The predicted opinion extracted by ChatGPT allows the fine-tuned model to predict the manipulation label correctly.

	***Main* Discourse**
Tweet	#Zuckerbergtestimony Mark Zuckerberg's testimony before the House Energy and Commerce Committee is over.
Article	... U.S. Rep. Joe Barton, R-Texas, chairman of the House Energy and Commerce Committee, made the following statement today during the full committee hearing on the Administrationś FY Ó7 Health Care Priorities: "Good afternoon.. **Let me begin by welcoming Secretary Michael Leavitt today to the Energy and Commerce Committee.** We look forward to hearing him testify about the Administrationś Fiscal Year 2007 Health Care Priorities ...
	***Cause* Discourse**
Tweet	Thank you, Rep. Johnson, for your service! Weekly Republican Address: Rep. Sam Johnson (R-TX) ... via @YouTube
Article	... In the address, Boehner notes that this is a new approach that hasn't been tried in Washington – by either party – and it is at the core of the Pledge to America, a governing agenda Republicans built by listening to the people. **Leader Boehner recorded the weekly address earlier this week from Ohio, where he ran a small business and saw first-hand how Washington can make it harder for employers and entrepreneurs to meet a payroll and create jobs.** Following is a transcript ...

Table 9: Examples of manipulated sentences with a *Main* discourse and a *Cause* discourse. The manipulated sentences are marked in **boldface**. The manipulating and pristine spans are marked in red and blue, respectively.