BanMANI: A Dataset to Identify Manipulated Social Media News in Bangla

Mahammed Kamruzzaman[1], **Md. Minul Islam Shovon**[2], **and Gene Louis Kim**[3]

[1,3] University of South Florida , Tampa, FL, USA 33620
[2] Rajshahi University of Engineering & Technology , Rajshahi-6204, Bangladesh
[1,3]{kamruzzaman1, genekim}@usf.edu
[2]mainulislam588@gmail.com

Abstract

Initial work has been done to address fake news detection and misrepresentation of news in the Bengali language. However, no work in Bengali yet addresses the identification of specific claims in social media news that falsely manipulates a related news article. At this point, this problem has been tackled in English and a few other languages, but not in the Bengali language. In this paper, we curate a dataset of social media content labeled with information manipulation relative to reference articles, called BanMANI. The dataset collection method we describe works around the limitations of the available NLP tools in Bangla. We expect these techniques will carry over to building similar datasets in other low-resource languages. BanMANI forms the basis both for evaluating the capabilities of existing NLP systems and for training or fine-tuning new models specifically on this task. In our analysis, we find that this task challenges current LLMs both under zero-shot and fine-tuned settings.[1]

1 Introduction

Misinformation is an increasingly pressing concern in the current social and political landscape where information frequently spreads through social media platforms with few constraints to reflect the information in reliable sources. This is further exacerbated by the presence of "bots" made by malicious actors that are designed to artificially spread ideas that distort reality (Ferrara, 2020; Lei et al., 2023). In order to mitigate this issue, considerable work has been done to identify fake articles (Shu et al., 2020), verifying scientific and encyclopedia claims (Wadden et al., 2020; Thorne et al., 2018), and identifying claims on social media that distort news from trusted sources (Huang et al., 2023). However, most such work is limited to only English.

Bangla, with the fifth-most L1 speakers worldwide, at 233.7 million[2], only has prior work in detecting fake articles (Hossain et al., 2020). More work in this direction is needed

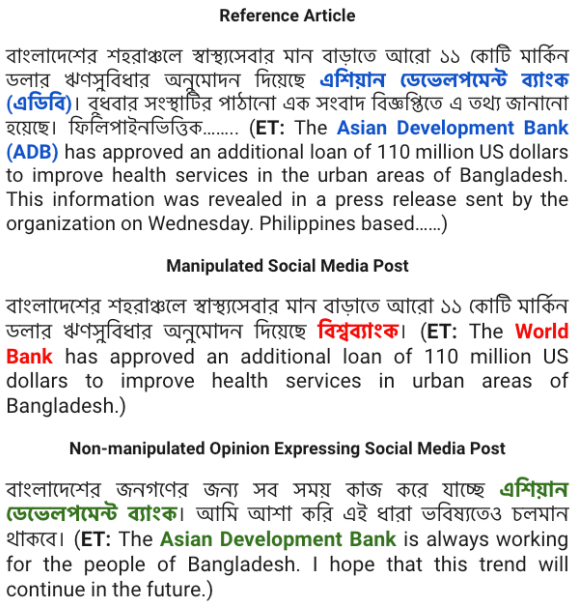

Figure 1: Example of manipulated and non-manipulated social media post with the corresponding reference article. **ET** denotes the English Translation of the given Bangla sentences. In the given example, the Asian Development Bank (highlighted in blue color) is incorrectly referred to as the World Bank (highlighted in red color) in the manipulated post. In the non-manipulated opinion-expressing post, the Asian Development Bank (highlighted in green color) is correctly referred to.

[1]Our dataset is available at https://github.com/kamruzzaman15/BanMANI.

[2]https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, pages 51–58
Varna, Sep 7, 2023
https://doi.org/10.26615/978-954-452-090-8_007

for Bangla on social media platforms, as demonstrated by the 2012 Ramu incident. In the Ramu incident, a Facebook post from a fake account led to the destruction of a Buddhist temple and dozens of houses in Bangladesh by an angry mob of almost 25,000 people (Ahmed and Manik, 2012). In this vein, we construct a dataset of news-related social media content for identifying news manipulation in social media, `BanMANI`. This dataset is the comparable Bangla counterpart to the ManiTweet dataset (Huang et al., 2023) in English. Figure 1 shows an example of a reference news article alongside both a manipulated and a non-manipulated social media post.

This paper's contributions are the following.

- We construct a publicly available Bangla dataset of 800 news-related social media items that are annotated as manipulated or not relative to 500 reference news articles.
- We present a semi-automatic method for generating such a dataset, which allows scalable dataset collection using annotators efficiently for languages with few available NLP tools.
- We demonstrate that current SOTA LLMs struggle on this task, both in zero-shot and fine-tuned settings.

2 Related Work

This paper is most closely related to fact-checking and fake news detection tasks. While much work in this direction has been done in English, Hossain et al. (2020) only recently started work in this domain in Bangla by releasing a dataset for fake news detection.

In English, Huang et al. (2023) released a dataset for identifying news manipulation in Tweets. In order to supplement fully-human data, they used a semi-automatic approach of generating Tweets using ChatGPT and using human annotators to validate and label the results. They found that ChatGPT and Vicuna failed to solve this new task, even after fine-tuning. In their work, they used `FakeNewsNet` (Shu et al., 2019) dataset to seed their reference articles.

Fact-checking tasks closely resemble our task in that claims must be compared against reference evidence, such as in the FEVER (Thorne et al., 2018) and SCIFACT (Wadden et al., 2020) datasets. Techniques for this kind of fact-checking work often use a retrieval module that pulls relevant data from the supplied candidate pool. The degree of consistency between a piece of evidence and the input claim is then evaluated using a reasoning component (Pradeep et al., 2021).

While our task compares text against a reference article, models must be able to separate social media news related to the reference article from those that only convey opinions to ensure the successful completion of our task. This is the key difference between these (i.e., fact-checking and fake news detection) and our work.

3 Task Definition

Our goal is to identify whether a news-related *social media item* (a post or a comment) is manipulated. If the social media item is manipulated then furthermore to determine what particular information is being manipulated relative to a related reliable reference article. We divide this task into three parts.

Subtask 1. First, we identify whether a particular social media item is manipulated. This part is a binary classification task and we consider an item as manipulated if there is at least one manipulated excerpt.

Subtask 2. Second, if a social media item is classified as manipulated then we need to identify which particular excerpt is manipulated. The task then is to identify the excerpt of the social media item which is not consistent with the original reference news article. In our dataset, we refer to any manipulated or newly introduced span as an *altered excerpt*.

Subtask 3. The third subtask is to identify the part of the original news article which is manipulated in the social media item. In our dataset, we define the information being manipulated as *original excerpt*. Models must produce an empty string or "none" as the output when the *altered excerpt* is inserted without modifying any *original excerpts*.

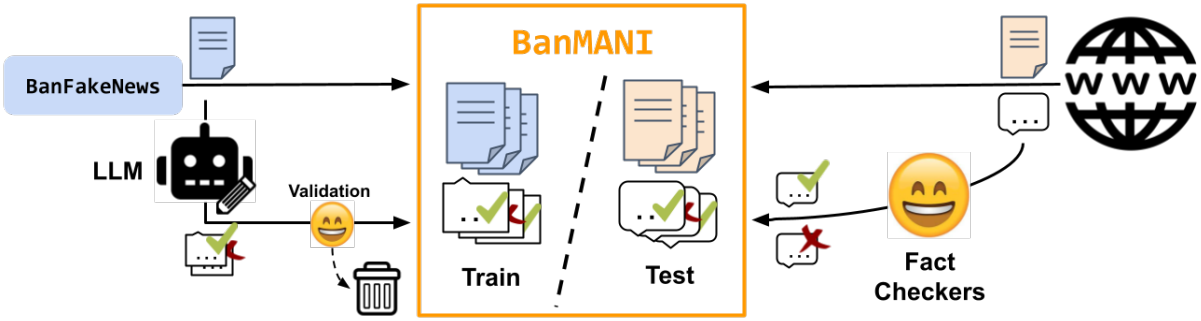

Figure 2: A diagram of the dataset collection procedure. The left side shows the semi-automatic data collection procedure for the training set, seeded by the BanFakeNews dataset (Section 4.3). The right side shows the collection of human-fact-checked items for the test set (Section 4.4).

4 BanMANI: Dataset Creation

We confined our test data collection to Facebook since this platform is more commonly used by Bangla speakers compared to typically studied media platforms for English speakers (e.g., Twitter, Instagram, etc.).[3] We have created a dataset that contains 800 news-related social media items with 500 associated news articles. Our dataset contains 530 manipulated items and 270 non-manipulated items. The breakdown of our dataset is shown in Table 1.

4.1 Bangla-specific Challenges

The task of constructing a Bangla version of the ManiTweet dataset is complicated by several factors. First and foremost, the availability and efficacy of NLP tools in Bangla are much more limited than in English. This means that some reliably automated steps in the English data collection process may be impossible or unreliable in Bangla. In addition to this, a Bangla version of FakeNewsNet (Shu et al., 2020), the dataset that Huang et al. (2023) use as a basis for their ManiTweet dataset, does not exist. FakeNewsNet contains news articles with associated Twitter data which can be directly annotated with any identifiable manipulation. In our dataset construction process, we must identify news articles and corresponding social media posts ourselves since no such seed dataset exists in Bangla.

[3]According to StatCounter.com (https://gs.statcounter.com/), Twitter held 22.01% of thie social media market share in the US in June 2023, but only 1.41% in Bangladesh. On the other hand, Facebook held 48.2% of the market share in the US but 78.84% in Bangladesh.

4.2 Source of News Articles

We collected our news article from BanFakeNews (Hossain et al., 2020), a dataset for Bangla fake news detection. From that dataset, we selected 6 domains where we expect the most social media manipulation to occur: National, International, Politics, Entertainment, Crime, and Finance. From those categories, we selected 2.3k seed news articles, which were used to generate manipulated and non-manipulated social media news. We furthermore upsampled the Politics and Entertainment domains as these were singled out in Huang et al.'s (2023) analysis. For more details on the initial data selection, see Appendix A.

4.3 Social Media Item Generation

No suitable dataset of social media items with corresponding news articles exists in Bangla. In order to efficiently use our limited annotator resources, we deploy a semi-automated data collection process using ChatGPT[4]. We use ChatGPT to generate both manipulated and non-manipulated social media items from a seed news article, which is then validated by human annotators.

4.3.1 Collection of Substitutable Sets

In order to generate manipulated social media items using ChatGPT, we first must identify plausible but incorrect substitutions that can be made in social media items. We collect such possible substitutions through a named entity recognition (NER) tagger. This mirrors the procedure used by Huang et al. (2023). We

[4]GPT-3.5-turbo

Split	Manipulated		Non-manipulated		
	Post	Comment	Post	Comment	Total
Train	370	100	130	50	650
Test	40	20	60	30	150

Table 1: `BanMANI` Dataset Statistics

collect news-relevant substitutable sets by running a Bangla NER tagger on 2,300 news articles from the `BanFakeNews`. We consider any two entities with the same NER label as substitutable with each other. We collected all `PERSON`, `ORGANIZATION`, and `LOCATION` named entities from the NER results, following the NER label choices used by Huang et al. (2023).

Based on preliminary experimentation of available Bangla NER systems, we found mBERT-Bengali-NER[5], a BERT-based multilingual Bengali NER system, to perform the best in our use case. Due to the high error rate of Bangla NER taggers, we perform a human filtering step to remove mistakes in the automatic NER labeling. Details of this step are provided in Appendix B.

We supplement the automatically collected entity sets with manually constructed sets of common entity substitutions that were identified in the data construction process. For example, some people write এশিয়ান ডেভেলপমেন্ট ব্যাংক (Asian Development Bank) in their post, when the original news article contains বিশ্বব্যাংক (World Bank). The same interchange also happens for বাংলাদেশ ব্যাংক (Bangladesh Bank) and এশিয়ান ইনফ্রাস্ট্রাকচার ইনভেসমেন্ট ব্যাংক (Asian Infrastructure Investment Bank). So we created a substitutable subset inside the `ORGANIZATION` entity label that contains these four together (এশিয়ান ডেভেলপমেন্ট ব্যাংক, বিশ্বব্যাংক, বাংলাদেশ ব্যাংক, এশিয়ান ইনফ্রাস্ট্রাকচার ইনভেসমেন্ট ব্যাংক). Members of these hand-curated sets can similarly be substituted with each other to create manipulated news.

4.3.2 Item Generating Prompts

We use the `content` attribute of the news articles from the `BanFakeNews` to create the social media posts and `headline` attribute of the news articles to create comments. Since comments are generally shorter than posts, we use a different approach to generate them. Social media item generation prompt templates are given in Figure 3.

Non-manipulation prompt for post: You are given a Bangla news article: **BENGALI_NEWS_ARTICLE.** Summarize the article without changing its original meaning and comment about it. Keep it within 250 characters.

Non-manipulation prompt for comment: You are given a Bangla news article: **BENGALI_NEWS_ARTICLE** Summarize the article without changing its original meaning and generate a short headline about it. Keep it within 100 characters.

Manipulation prompt for post: You are given a Bangla news article: **BENGALI_NEWS_ARTICLE.** Summarize and comment on this article but change the **ORIGINAL_EXCERPT** to **ALTERED_EXCERPT** and include **ALTERED_EXCERPT** in your comment. Keep it within 250 characters.

Manipulation prompt for comment: You are given a Bangla news article: **BENGALI_NEWS_ARTICLE.** Summarize and generate a short headline about it but change the **ORIGINAL_EXCERPT** to **ALTERED_EXCERPT** and include **ALTERED_EXCERPT** in your comment. Keep it within 100 characters.

Figure 3: Prompt templates for social media item generation. Here, the "ALTERED EXCERPT" and "ORIGINAL EXCERPT" bear the same meaning as described in Subtask 2 and 3 respectively.

After generating the manipulated and non-manipulated social media items using ChatGPT, we assign human annotators to validate the generated data. The total number of generating manipulated and non-manipulated items using ChatGPT are 2.3k. The generated social media items from ChatGPT are not always coherent or related to the seed news articles. So the human annotators discarded 1.65k generated data during the validation stage. We use the remaining social media items generated by ChatGPT as our training data. In this project, graduate and undergraduate students are working as human annotators. The inter-annotator agreement between the involved annotators is 92.2% per Cohen's kappa (Cohen, 1960). The detailed data annotation process, including screenshots of the annotation inter-

[5] https://huggingface.co/sagorsarker/mbert-bengali-ner

Domain	Manipulated Articles
National	16
International	14
Politics	19
Entertainment	5
Crime	1
Finance	5

Table 2: Manipulated News Articles in Test Data

faces, is available in Appendix C.

4.4 Test Data Collection

We collected 150 human-generated social media items for our test set. These items were collected manually from Facebook using two distinct strategies. In the first strategy, items were sourced from media and news company pages on Facebook, such as Prothom Alo.[6] From these pages, we collected posts that shared a news article with accompanying post text that add commentary as well as comments from the comment sections under news articles on the page. In the second strategy, we collected posts from pages such as BD FactCheck [7] and Rumor Scanner [8] which specialize in identifying fake news published on other platforms.

5 Exploratory Data Analysis

From Table 2, we see most of the manipulated news is political. Some people spread manipulated news on social media to influence public opinion, promote a particular political party, etc., and these might be the reasons behind the manipulated political news. Also, we notice that national and international news are manipulated in a bigger amount. Sites and pages with low trustworthiness are most likely to spread manipulated news. The followers of those sites and pages are most likely unaware of the fact and accidentally post manipulated news.

6 Experimental Setup

6.1 Models

Zero-shot ChatGPT. We use ChatGPT for the zero-shot setting experiments. For details prompt about the zero-shot experiment, see Appendix D.

Fine-tuned. Fine-tuning allows the user to get more out of the available models through provided API. As a result, it can achieve higher quality results than traditional prompt design, train on more examples beyond the limit of traditional prompt, and saves token due to shorter prompts. Fine-tuning improves on few-shot learning by training on much more examples that can fit in a prompt. Which lets you achieve better results in fine-tuned tasks. In general, fine-tuning involves preparing and uploading training data, training the new fine-tuned model with prepared data, and using the fine-tuned model. For our work, we used GPT-3 (Brown et al., 2020) `ada`[9] as our base model due to the unavailability of fine-tuning for the latest models. Also, `ada` is capable of handling simple tasks and is the fastest model in the GPT-3 series. We used a prompt-completion format for our training data and later fine-tuned our model with this data, resulting in competitive outputs.

6.2 Evaluation Metrics

For subtask 1, we use F1 score as this is simply a classification task. Since subtasks 2 and 3 involve span extraction, we use Exact Match (EM) and ROUGE-L (RL).

7 Results & Analysis

The result of the zero-shot ChatGPT and our fine-tuned model is presented in Table 3 and Table 4 respectively. From Table 3 and Table 4, we can see that our fine-tuned model outperforms the zero-shot ChatGPT model for subtask 1, where the F1 score of zero-shot ChatGPT and fine-tuned model is 57.02% and 65.77% respectively. In terms of EM, we can see that our fine-tuned model performs better for both subtask 2 and subtask 3. For subtask 2, if we look at the RL value of our fine-tuned model, we can see that the precision of RL is 69.26%, which is 33.2% more than the zero-shot model. That is also the case for the F1 score of RL. In the same way, for sub-task 3, we can see that the precision and F1 score of RL outperforms the zero-shot model.

[6] https://www.facebook.com/DailyProthomAlo
[7] https://www.facebook.com/bdfactcheck
[8] https://www.facebook.com/RumorScanner
[9] https://platform.openai.com/docs/models

Metric	Subtask 1	Subtask 2	Subtask 3
F1	57.02	--	--
EM	--	8.2	12.3
RL (r, p, f)	--	(79.72, 36.06, 46.83)	(64.78, 41.04, 49.94)

Table 3: Evaluation results of ChatGPT with Zero-shot. Here, EM denotes Exact Match, RL denotes ROUGE-L, which is broken down into (r, p, f) denoting recall, precision, and F1 score respectively.

Metric	Subtask 1	Subtask 2	Subtask 3
F1	65.77	--	--
EM	--	11.9	13.34
RL (r, p, f)	--	(61.95, 69.26, 64.75)	(63.65, 50.74, 56.46)

Table 4: Evaluation result of our fine-tuned GPT-3 model. The table labeling conventions match those of Table 3.

8 Limitations & Future Work

Due to our budget limitation, we were not able to collect a large set of human-written social media items. This means that there exists a gap between the quality of the training and test data; the training set was automatically created, unlike the test data. In the future, we will collect more human-written items from social media to create an entirely human-written training dataset. Our prompts are also purposefully simple, as this was the first step in creating such a dataset. We expect to get qualitative gains in the automatically generated data with more careful prompt engineering. Finally, our experiments were limited to only a single popular LLM for each setting. Expanding the experiments to cover other LLMs, especially open-source LLMs would lead to more robust experimental results and better replicability. We also leave the few-shot method as our future work.

9 Conclusion

In this paper, we presented the `BanMANI` dataset, the semi-automatically constructed dataset of news manipulation in social media. This dataset extends Huang et al.'s 2023 ManiTweet dataset to Bangla. Our semi-automatic collection process generates social media posts from seed articles using a multilingual LLM and a Bangla NER system. These results are filtered using human annotators for efficient use of annotator time. We find that both zero-shot and fine-tuned LLMs struggle on this dataset, pointing to important directions of future work. Surprisingly, we find that LLMs perform similarly effectively on this dataset when compared to the English variant. We hope that this new resource can help with combating information manipulation in Bangla-speaking social media communities. Furthermore, we believe that the technique laid out here can act as a basis for similar work in other under-served languages in NLP.

References

Inam Ahmed and Julfikar Ali Manik. 2012. A hazy picture appears. [Online; posted 03-October-2012].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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

Emilio Ferrara. 2020. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, 25(6).

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. BanFakeNews: A dataset for detecting fake news in Bangla.

In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France. European Language Resources Association.

Kung-Hsiang Huang, Hou Pong Chan, Kathleen McKeown, and Heng Ji. 2023. Manitweet: A new benchmark for identifying manipulation of news on social media.

Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Jundong Li, Qinghua Zheng, and Minnan Luo. 2023. BIC: Twitter bot detection with text-graph interaction and semantic consistency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10326–10340, Toronto, Canada. Association for Computational Linguistics.

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.

K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. 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, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

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.

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.

A Initial Data selection Details

Initially (before the first round of human validation) we took 2.3k news articles and generated news-related social media items. In Table 5, we show the details of each category data.

Domain	No. of Articles
National	288
International	288
Politics	690
Entertainment	460
Crime	287
Finance	287

Table 5: Initially Taken News Articles Based on Each Category

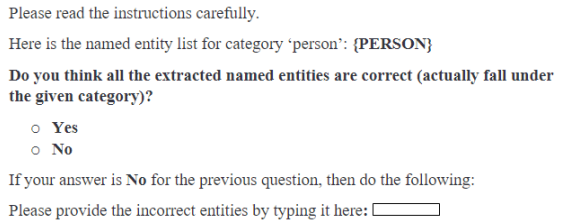

Figure 4: NER Annotation Interface

B NER Annotation Process

Since the performance of the Bangla NER system is not accurate, we need to discard some of the named entities after extracting them. We presented our NER annotation details in Figure 4. In Figure 4, we only show the annotation process for `PERSON` and we do this for every other category (i.e, `ORGANIZATION, and LOCATION`).

C Data Annotation Process

In our research, we perform a two-stage data annotation process for our data. To ensure data quality and consistency, we have selected only those annotators whose mother tongue is Bengali. In this project, all the annotators are graduate and undergraduate students from different institutions. In this project, we have selected a total of 5 students as annotators and kept the data that got at least three annotators' votes.

Stage 1. In the first stage, we asked each annotator to read the generated social media items carefully and see whether it makes sense to them or not and this stage is only limited to our train data. We need to introduce this round because sometimes ChatGPT generates very poor data that doesn't make any sense or totally unrelated to the corresponding news article. Especially the Bangla data generation

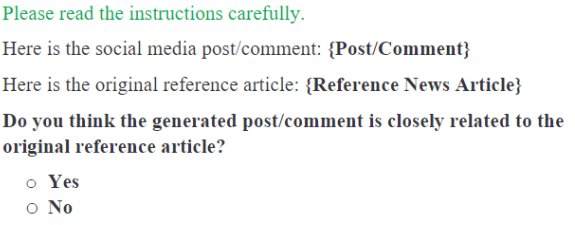

Figure 5: Stage 1 Annotation.

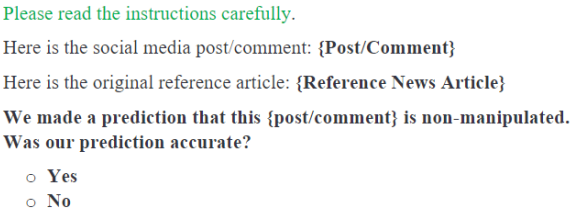

Figure 6: Stage 2 Annotation for Non-manipulated Social Media Items.

performance of ChatGPT is bad compared to English. So this round of annotation ensures the generated items are not unrelated to the news topic. Figure 5 represents the annotation details of stage 1. We only keep those data that receive a 'Yes' in stage 1.

Stage 2. Here in stage two, we annotated our test and train both data based on manipulated and non-manipulated classes. For the non-manipulation class, we follow the instructions pictured in Figure 6. The annotation interface for the manipulated class is presented in Figure 7. We keep the data that receive a 'Yes' for non-manipulated class. For the manipulated class, we asked a few more questions for annotators because it is difficult to collect manipulated data from social media. If the answer to the first annotation interface question for manipulated class is 'Yes', then we asked two more questions. The purpose of the latter two questions is that if we classified the manipulated post correctly but accidentally got the altered or original excerpt wrong, then the annotators can give us the accurate excerpt and in this way, we can keep the data.

D Zero-shot Prompt for ChatGPT

The zero-shot prompt template for the ChatGPT model is shown in Figure 8.

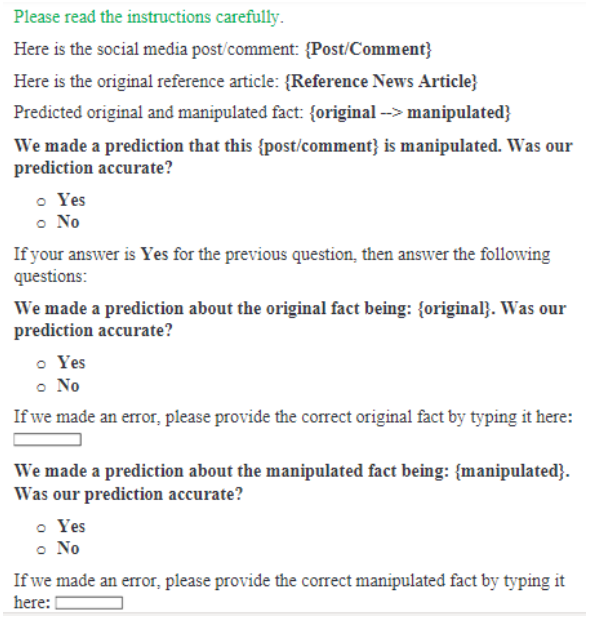

Figure 7: Stage 2 Annotation for Manipulated Social Media Items.

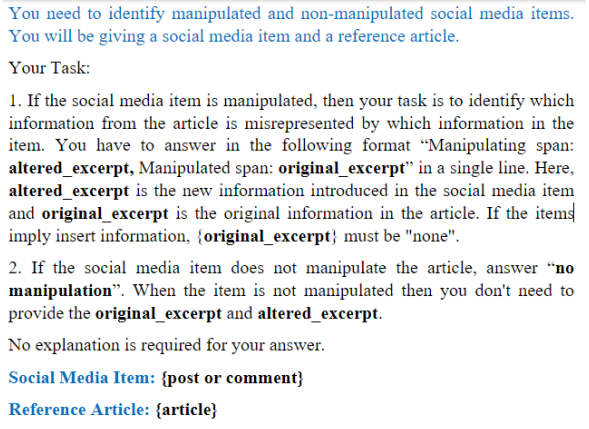

Figure 8: Zero-shot Prompt