Fact Manipulation in News: LLM-Driven Synthesis and Evaluation of Fake News Annotation

Luka Golob
lukag26@gmail.com
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abdul Sittar
abdul.sittar@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abstract

Advancements in artificial intelligence and increased internet accessibility have made it simpler to create and disseminate fake news with customized content. However, they also improved the ability to analyze and identify such misinformation. To effectively train high-performance models, we require high-quality, up-to-date training datasets. This article delves into the potential for generating fake news through factual modifications of articles. This is facilitated by prompt-based content generated by large language models (LLMs), which can identify and manipulate facts. We intend to outline our methodology, highlighting both the capabilities and limitations of this approach. Additionally, this effort has resulted in new quality synthetic data that can be incorporated into the standard FAK-ES dataset.

Keywords

fake news, synthetic data, fact extraction, fact verification, large language models

1 Introduction

Synthetic data refers to artificially generated data that is not obtained by direct measurement or observation of real-world events. Instead, it is created using algorithms and simulations. The primary purpose of synthetic data is to provide a realistic alternative to real data for various use cases, such as training machine learning models, testing systems, ensuring data privacy, and more.

We will generate synthetic data from news articles. By making sure, that the information in the news is changed we can safely call it fake news. In our article, fake news will denote articles that are *intentionally* and *verifiably* false [4]. Synthetic data enhances model training by providing additional examples to supplement scarce labeled datasets and allows for privacy-conscious testing without real content manipulation. It enables adaptability to evolving fake news tactics by simulating diverse scenarios from the newest data, thereby improving the robustness and resilience of detection algorithms [3].

Large language models (LLMs) made a huge difference in the world of news. Fake news is now much easier and cheaper to construct, but we also have additional methods to help us tackle its spread. Numerous articles appeared trying to partake in this effort. The following are the main scientific contributions of this paper:

(1) A methodology to create synthetic data for fake news using LLMs.
(2) We then use this methodology, to adapt the FA-KES dataset with 100 additional synthetic fake news [1].

In Section 2, we discuss work that is closely related to our task. Section 3 then outlines the methodology for generating synthetic fake news, culminating in Section 4, where we present the results and introduce some modifications to the methodology. Finally, in Chapter 5, challenges, capabilities, and potential improvements are considered.

2 Related Work

A wide range of approaches to generate fake synthetic news with LLM has been developed. In [8] authors generated huge amounts of fake news and categorized them into multiple categories. LLMs can generate fake news by altering the style to mimic credible sources or using sensationalism to influence perception. They can subtly manipulate content to be perceived as true, blend real and fabricated information to exploit cognitive biases, or create convincing fictional narratives.

In general, when making a dataset we want a diverse distribution of fake datasets. In our case, we will focus on one way of data change, which comes under the umbrella of *Content Manipulation*. Similar news manipulations can be seen in [7] where the authors use two main techniques. The first one extracts the summary from the original text, which preserves the main content, which is then changed to produce a fake article. The second one asks a question about the article and changes the content of its answer, to construct a new article. Our approach is in nature similar to the Question-Answer framework.

Many articles provide fake news detection models made using synthetic data. Most popular are deep neural networks such as BERT [1]. But there are other fact-based approaches for fake news labeling as in [3]. In [2] they used GPT4-turbo for prompt-driven fake news detection.

3 Methodology

The methodology is divided into four conceptual steps: Data collection, Characterization of facts, Fact extraction, and Fact manipulation as presented in Table 1.

3.1 Data Collection

The publicly available FA-KES dataset [5], focused on the Syrian war, addresses the deficiency of manually labeled datasets in this domain of news data. It comprises 804 articles sourced from various media outlets. We used 426 articles that were manually labeled as authentic news, but we could just as well use the other (fake) articles.

[1] https://github.com/golobluka/Fake-news-generation-from-FA-KES-dataset

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.13

1. Data collection	2. Characterization of facts
• Should have textual and statistical facts	1.Name of casualty 2.Gender or age group 3.Cause of death 4.Type 5.Actor 6.Place of death 7.Date of death
3. Fact Extraction	**4. Fact manipulation**
Name of casualty: Civilians **Gender or age group**: e.g., child, adult, senior **Cause of death:** shooting, shelling, weapons, etc. **Type:** military personnel **Actor:** rebels, forces **Place of death:** Airbase **Date of death:** April 7, 2017	**Name of casualty**: Manipulated fact **Gender or age group**: Manipulated fact **Cause of death:** Manipulated fact **Type:** Manipulated fact **Actor:** Manipulated fact **Place of death:** Manipulated fact **Date of death:** Manipulated fact

Figure 1: A methodology to generate synthetic data for fake news detection

3.2 Characterization of Facts

While making the FA-KES dataset, its authors created seven factual categories:

(1) `Name of casualty or group`,
(2) `Gender or age group`,
(3) `Cause of death`,
(4) `Type`,
(5) `Actor`,
(6) `Place of death`,
(7) `Date of death`.

It is crucial to note that all articles have a similar structure, describing war incidents. This allows us to establish a consistent framework of facts, such as actor and casualty details. We stick to those facts, but generate them differently, employing LLMs capabilities with faster and cheaper execution, albeit with a slight reduction in reliability.

3.3 Fact Extraction

We extract facts by constructing prompts for LLMs. First approach was a few-shot prompt, which gives some examples of output. Later we constructed an additional approach: Say we are extracting the fact `Place of death` with this second technique. We give a detailed description of what should be extracted and then LLM reads the article and performs the task solely on this basis. This description is usually longer and contains more context. The issues with fact extraction in general are:

- Some articles lack certain facts or merely imply them. LLMs can identify this, outputting responses such as "No information."
- Longer articles may contain multiple events, each with distinct data such as dates or casualties. This can be managed by creating separate tables for each event or consolidating all events into a single table with various facts.

3.4 Fact Manipulation and Synthetic News Generation

The objective is to modify relevant information without altering the writing style or topic of the article. For this transformation, we used a chain of thought prompt, which for a given fact: 1) changes the fact to another with a different meaning, 2) generates a new article based on the altered facts. By changing one fact at a time, quality is improved compared to altering multiple facts simultaneously, as one fact creates a clearer chain of instructions. LLMs such as `Llama3.1:8B` often struggle with precise changes in the article, such as modifying implicit references or incorporating new facts. Quality can be improved by carefully adjusting the prompt content.

LLMs are also exceptional in summarization and paraphrasing. Both are used simultaneously with changing the facts. The problem is that we aim to maintain the extracted facts when summarizing. But this is not crucial, as it usually has better results as article generation.

3.5 Fake News Annotation and Fact verification

After we have generated the fake articles, we can label that data as "fake" or "non-fake", based on comparison with extracted facts. We performed this labeling with various models and compared the performance of labeling,to get the best model. In this experiment we decided for `Llama3.1`. To do the labeling, we are performing fact verification [4]. The fact verification task in general is making a decision as to whether a claim is correct, based on the explicitly-available evidence, such as Wikipedia articles or research papers. We have the extracted fact, which will be compared to the article content. The question thus becomes: Do these facts appear in the given article? This approach emphasizes factual content rather than the overall sentiment of the article.

There are two primary types of prompts: 1) Direct prompts that present the article and a table of facts, asking if the facts relate to the article, 2) Structured prompts that inquire about the correspondence of one fact at a time with the article. The question is: Does this fact correspond to the content of the article? This method combines individual results into an aggregated score. Say the `Place of death` is characterized as `Idlib and Daraa provinces`. Then the question posed to LLM is of the form: `Read the article and understand its places of death. Do Idlib and Daraa provinces "really correspond" to places of death in the article?`

We are not as interested in labeling, as we are interested in the quality of produced synthetic fake news. For this purpose, we will also use fact verification in a slightly different way. We are asking the LLM: Were the factual changes in fake news really made, as they were supposed to? A similar method is used in the article [7].

4 Experimentation and Results

4.1 Experimental settings

We selected 426 articles labeled as authentic news from FA-KES dataset. Then facts were extracted and transformed, as described in the previous section. At first two basic approaches were used to randomly choose 70 news articles and transform them. Afterward, we used the labeling procedure to compare performance, resulting in the table 1. Based on the results we then composed the final algorithm, which would be manually evaluated.

4.2 Evaluation

For every experiment, we first manually checked a minimum 10 percent of random examples to get an overview of how well the LLM was able to do the job. It is quite useful to print text that represents the procedure of decision-making that LLM undertakes, when challenged with the task. It was even helpful to see LLMs generated thinking procedure, as this gives valuable insight, into what is going on "under the hood". We believe that manual fact-checking is the first and most crucial step in generating good prompts. Based on fallacies one can then adjust prompts content. To shed some light on this procedure we have made the following overview.

4.3 Fact Extraction Results

Name of casualty or group:	Members of Nusra Front
Gender or age group:	Adults (no specific age mentioned)
Cause of death:	Explosion at a mosque
Type:	Non-civilian (militants)
Actor:	Unknown (no group claimed responsibility, but supporters blamed ISIS)
Place of death:	Ariha, Idlib province, Syria
Date of death:	Not specified in the article

Figure 2: Example of fact extraction.

LLMs are capable of recognizing different topics and extracting words that correspond to this topic, and also noting if the fact is not mentioned. At first, we extracted short words as represented in Figure 2.

The issue begins with nuances. For example, in many articles the `Actor` is only suspected but not known. In some cases, actor and causality are not precisely distinguished. This usually leaves LLM to some kind of arbitrariness. For this purpose, We also added a longer description that better captures the nuanced subtleties related to facts. This can also be captured in Table 1. There we see the results for short (normal) or detailed extracted facts. The recall is far worse in the case of short prompts. This likely means that there is an abundance of false negatives, which result from the fact, that labeling does not manage to match true articles and their corresponding short facts.

The shorter extracted facts are often not comprehensive. For example, under the label `Type` (which classifies civilian or non-civilian) it writes only civilians, even though, contextual understanding also includes some non-civilian casualties.

Overall the most important insight remains: fact extraction has better quality than article generation.

4.4 Quality and coherence of synthetically generated fake news

The LLM can detect (for example) the `Actor` of some attack in the news, and then it is mostly able to change every occurrence of this `Actor` with another `Actor`. But if we would like to preserve all the coherence of the article much more would need to be done.

News usually contains background information, that provides context for the accident. Our algorithms failed to properly adjust this context, leaving it unchanged in most cases. Our fake news fails to preserve enough coherence to be trusted by a skeptical reader, who tries to connect background material to the event in the article.

Table 1: Comparison of fake synthetic data.

Type of data	Number of facts manipulated	Precision	Recall	F1	Accuracy
Summarization	2/7	0.74	0.63	0.68	0.71
Detailed facts	2/7	0.70	0.80	0.75	0.73

Generating false text, while maintaining coherency, is challenging for LLM. In this task, we have changed one fact: for example, the `Place of death` may be changed to another city or neighborhood. Then this fact must be changed in the article while maintaining other factual information. Here are the main issues:

- In the beginning some facts did not get changed, or the facts were altogether just removed from the article. We managed to reduce this error by adjusting the prompt. It is difficult to adjust all occurrences of the fact, especially if it is only implied and not explicitly stated. We managed to minimize this problem, by a method yet to be shown in section 4.5.
- What remains is the problem of a wider context, Suppose we change the town of the incident, then we must change the name of the neighborhood accordingly. LLM usually fails in this, leaving our article inconsistent, which is a widespread problem.
- LLM does not want to output the content because of harmful content or does not want to produce articles that could be used with illegal intent. This was quite a common problem, which is also reasonable, based on the violent content of articles and the possible abuse of LLM-generated content. The best thing to prevent this error is to use uncensored LLM. In other cases, one can adjust the prompts by removing suspicious words like "fake news".
- The Generated article was shorter, skipping the original text which was not linked to extracted facts. This problem was reduced but still exists in long articles.
- If the fact is not present in the article, then it is hard for LLM to incorporate a new fictitious fact into the text. Mainly it just adds the information in separate sentences.
- When we change facts, traces of the old facts still persist. This is especially common in complicated articles with diverse structures.
- Sometimes the change does not bring about any additional meaning. For example, LLM might change previously unknown casualties and designate them as civilians. They were implied to be civilians all along, and this makes only a minor change and is not really fake.

4.5 Fact verification with LLMs

Remember that in this task, the prompt asks: Does this fact "really correspond" to the content of the article? Performance largely depends on how the program takes the word "really correspond". Words have many nuances: different words can have different meanings, which can complicate labeling. To simplify: we can be stricter, in the sense that words must be the same in the literal sense, or we can count on the similarity of meaning [6]. Based on our goal of creating fake news it is best to focus on meaning and not concrete words.

Here are some common problems:

- Sometimes the fact is changed, but LLM skeptically assumes, that those two names refer to the same group.
- In longer articles, where there are many events, the names get changed only in some events (usually at the beginning of the article). In this case, the LLM can make unwanted predictions, labeling the fact as true rather than false.

Manual checking shows that labeling is more accurate than generation of fake news. This leads us to use labeling as a means to improve article generation.

Table 1 was used to compare different ways to generate fake news. It shows two of the best datasets, which contain true articles and their false twins, generated in two ways:

(1) Fake news generated by "standard" fact extraction and with additional summarization.
(2) Fake news generated by "detailed" fact extraction and with an additional paraphrasing of the article.

In this experiment, instead of merely categorizing the articles as true or false, the results shown in Table 1 reflect how well the generation process aligns with fact verification.

Low precision in the row with Detailed facts led us to detect articles that were not changed. We implemented a strategy where labeling was applied after generating the fake articles to assess the quality of the generation. LLMs often provide incomplete responses and struggle to correct them directly. By introducing an additional verification step, we were able to enhance the overall accuracy of the results.

4.6 Final Dataset Description

In the end, we constructed 100 fake-news based on a prior experiment, which can be found on GitHub [2]. In every article we randomly chose three facts and changed them. Afterward, we carefully went through 10 examples, which are also present on Git Hub, while here we present only the main points:

- Fact verification improved quality by making sure, that the synthetic fake article really incorporated new information. More than 90% new facts really got incorporated in the article. Sometimes new information is only added as additional text(and does not seriously change the main topic).
- Fact is not always incorporated in all places where it is referenced, which leads to inconsistencies. The new article is then a blend of old and new information.
- There are problems with ¨detailed" prompts. Containing more information results in contradictions as we change only one fact at a time.

5 Conclusion

In this article, we focused on exploring the potential of LLMs in fact extraction and generation of fake news. Our motivation was primarily to understand how accurate are LLMs in fact extraction and how reliably LLMs generate synthetic news by altering facts. As a result of our experiment, we have generated 100 synthetic news by randomly transforming there out of seven facts and have performed a manual evaluation, to observe the quality of the generated news dataset.

[2]https://github.com/golobluka/Fake-news-generation-from-FA-KES-dataset

5.1 Problems, Capabilities and Possible Improvements

- In this stage, LLMs like *Lamma3.1:8B* are not able to coherently change certain facts of news articles. Changing facts can distort the article content, which appears to be extremely hard to manage. This normally does not happen for manageable data as dates (changing the time of some event), but for much more involved actors of the attack in the article. Even so, the synthetic fake news provides valuable information.
- We did not use the model, which has additional information about the news content. Providing additional context would likely have a beneficial effect on all the processes.
- In our case facts were largely dependent on each other. For example `Gender or age group` is an extraction of `Name of casualty or group`. We think it is best if such dependencies are removed because they bring to inconsistencies when changing facts. An additional solution would also be to change `Gander or age group` whenever `Name of casualty or group` is changed.
- Fact extraction is close to human-like quality. The issue is, that besides manual checking, it is hard to find a good measure of the quality of extracted facts.
- Detection of changed facts is in quality similar to extraction of facts (this is not surprising, since they are based on the same skill). Because of the diversity of meanings in language, it is hard to specify the exact reasoning procedure of LLMs and many mistakes come from this kind of miscommunication.

6 Acknowledgments

This work was supported by the European Union through AI4Gov (101094905) and TWON (101095095) EU HE projects and the Slovenian National grant (CRP V2-2272).

References

[1] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. 2023. Content-based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, 530, 91–103. DOI: https://doi.org/10.1016/j.neucom.2023.02.005.

[2] Fredrik Jurgell and Theodor Borgman. 2024. Fake news detection : using a large language model for accessible solutions. (2024).

[3] Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, investigate, judge and determine: a novel llm-based framework for few-shot fake news detection. (2024). https://arxiv.org/abs/2407.08952 arXiv: 2407.08952 [cs.CL].

[4] Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. (2021). https://arxiv.org/abs/2111.03299 arXiv: 2111.03299 [cs.LG].

[5] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. Fa-kes: a fake news dataset around the syrian war. In *Proceedings of the international AAAI conference on web and social media*. Vol. 13, 573–582.

[6] Abdul Sittar, Dunja Mladenic, and Tomaž Erjavec. 2020. A dataset for information spreading over the news. In *Proceedings of the 23th International Multiconference Information Society SiKDD*. Vol. 100, 5–8.

[7] Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: a study of real-world detection challenges. (2024). https://arxiv.org/abs/2403.18249 arXiv: 2403.18249 [cs.CL].

[8] Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024. Megafake: a theory-driven dataset of fake news generated by large language models. (2024). https://arxiv.org/abs/2408.11871 arXiv: 2408.11871 [cs.CL].