GenAI for Good: The Misinformation and Veracity Engine

Calvin Nguyen
can005@ucsd.edu

Samantha Lin
yul186@ucsd.edu

Dr. Ali Arsanjani
arsanjani@google.com

Abstract

The objective of our product, Chenly Insights, is to help users combat misinformation using factual factors and microfactors. As misinformation gets easier to spread due to deep fakes and large social networks, this project empowers users by giving them a veracity (truthness) for the news they receive through a Mesop interface. A user will link and upload a news article of their choice. Then, two different types of AI, generative and predictive, judge the article on appropriate factuality factors. The scores are accumulated and then shown to the user, along with a spider chart and a table that further explains the final score. This vector database is constantly updated with new information from fact check and news websites. Users can also converse with the AI and ask questions about each article and why a particular score was given. We also found the best prompting types with adjustments to the prompts that gives the highest accuracy based on our current progress.

Artifacts:
Poster: Poster File
Project Website: Project Webpage
Demo: Demo Video
Website Design: Figma Link
Code: GitHub Page

1 Introduction

Modern technological breakthroughs in large language models have leveraged the non-negligible issue of creating and publicizing misinformation, spreading false information without intent, and disinformation, spreading false information intent to mislead people, to a new level. LLMs like ChatGPT can generate context with a given prompt in an extremely speedy fashion with very little cost, and this aids disinformation spreaders in efficiently generating and disseminating human-like text that makes it hard for LLMs and humans to distinguish between true and false statements. This can lead to major issues within our society including intensifying polarized stances between people, masking the truth of an event, and more. With this in mind, the goal of our project is to provide a scoring system that breaks down the article and justifies them with multiple factors to help users identify the truthfulness of a given text.

The issue of misinformation and disinformation is not a new topic the community if facing. Researchers have approached this issue with different methods, from tuning RoBERTa models to training LLMs to detect more advanced disinformation, and the biggest breakthrough is the use of deep learning to evaluate the content of the text. In the paper "Disinformation Detection: An Evolving Challenge in the Age of LLMs", Jiang et al. (2024) discovered that a fine-tuned RoBERTa model was able to detect simple LLM-generated disinformation created by LLMs, but it was unable to detect complex false context such as combining human-written false information and LLM generated context and LLM context generated by using the chain of thought prompting technique. However, the research group has discovered that guiding LLMs with a chain of thought prompts allows LLMs to detect disinformation at a much higher level of accuracy. In summation, the paper suggests that existing detection techniques still struggle to detect disinformation, but some directions we can take to combat misinformation and disinformation in the future are to guide LLMs with chain-of-thought prompts and focus on highlighting the significance of meaning and contextual elements.

Therefore, the goal of our project is to build an advanced system that utilizes specialized AI and models to detect, rank, and address misinformation. The system will use 4 comprehensive factuality factors to analyze information. We will make the process for checking for misinformation as smooth as possible for users and help people access trustworthy and accurate content. Using this tool, users will get a more nuanced, balanced, and credible news. We can help improve public discourse, support political processes, and enhance our public health and safety. Additionally, this system, if successful, can be implemented into many different social media platforms, which allow these platforms to shutdown false information effectively.

2 Methods

To build Chenly Insights, we received guidance from our mentor to combine predictive and generative AI onto a Mesop interface. These models will work in tandem to compile multiple different scores of factuality factors into a veracity score. Our group was assigned

4 factuality factors to work on: Naive Realism, Poltiical Stance, Social Credibility, Sensationalism. Our mentor recommended that the more information and context you give a generative AI model, the better the results become. Therefore, we also implemented SERP API and a vector database. Each subsection will discuss the methods used for each model type and the adjustment added.

The entire workflow is shown through this Lucidchart, if you would like to explore more:

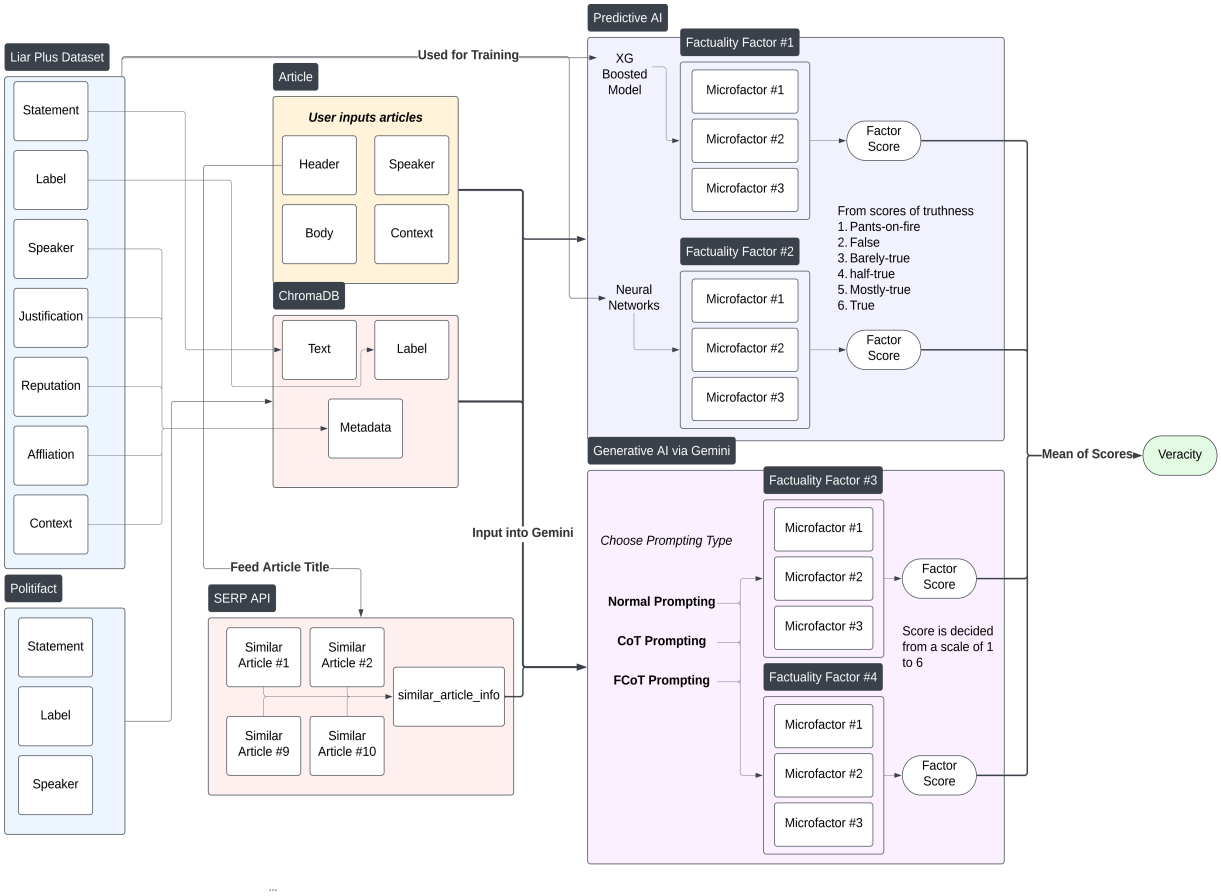

Figure 1: Flowchart from data to models to veracity

2.1 Predictive AI

For our predictive AI model, we applied traditional machine learning techniques to the Liar Plus dataset Tariq60 (2018). This dataset consists of a statement, a truthfulness label, and metadata surrounding that statement. The goal was to predict the truthfulness label of an article (from a scale of 1-6) based on microfactors extracted from each factuality factor. For example, one key factuality factor was ”Naive Realism”. The microfactors related to this factor were perspective analysis, dissenting view checks, and isolation analysis. This was operationalized by creating confidence, subjectivity, and reputation columns respectively, created from using HuggingFace models and textblob. If a statement exhibited high subjectivity and confidence, it is likely to have been viewed as the correct perspective while

dismissing other points of view. From there, we test out multiple different machine learning models of varying complexities and evaluate each model based on accuracy. Each factuality factor has its own machine learning model and its own derived microfactors. This process ensures each factuality factor has an accurate model associated with it.

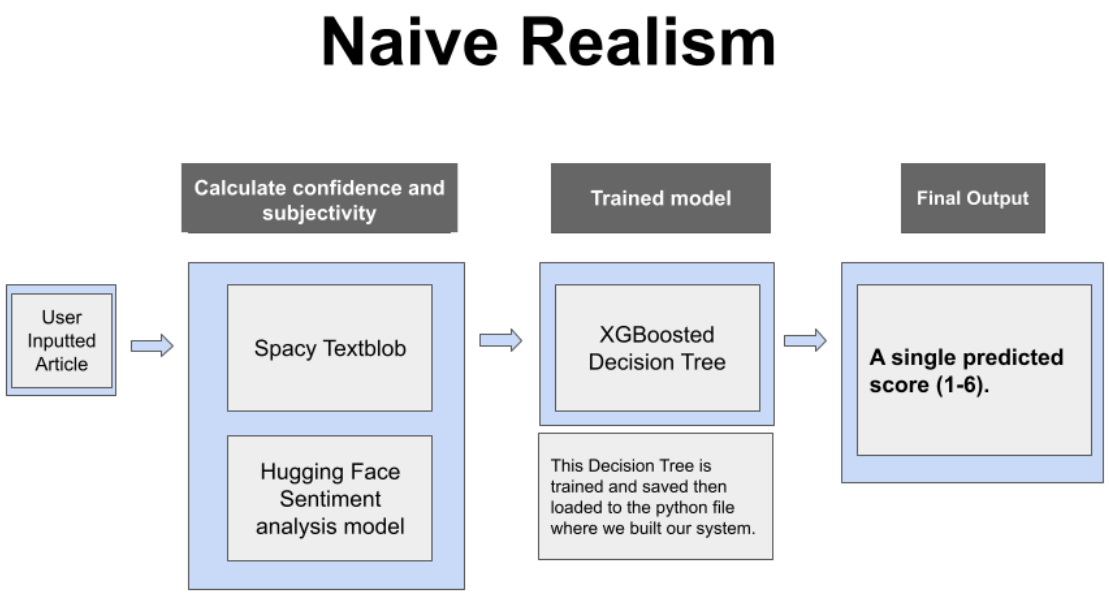

Figure 2: Naive Realism Workflow

Another factuality factor to fight against misinformation we used was ”social credibility”. We detect it by looking for patterns of speaker, context, and party affiliation with the truthfulness they were labeled. To generalize this model, we've decided to narrow down the categories with if-else statements to categorize different contexts but with similar contexts into the same category (Ex. Twitter, Facebook posts, and instagram posts will all be considered as Social Media). For the speakers and party affiliation, we've decided to keep them as they are since we can't categorize speakers and there were only a few unique party affiliations in the dataset. After data cleaning, we've built a neural network and one hot-encoded the speakers, context, and party affiliation from the Liar Plus Dataset as the input and fit the model with one hot-encoded label. This neural network consists of five layers including the input and output layers. There are 10 nodes in each layer and the sigmoid function is our activation function for all layers since this combination was the model with the best training result. This model was able to achieve a testing accuracy of around 21.36%, which is slightly better than the baseline (20%) suggested by our mentor.

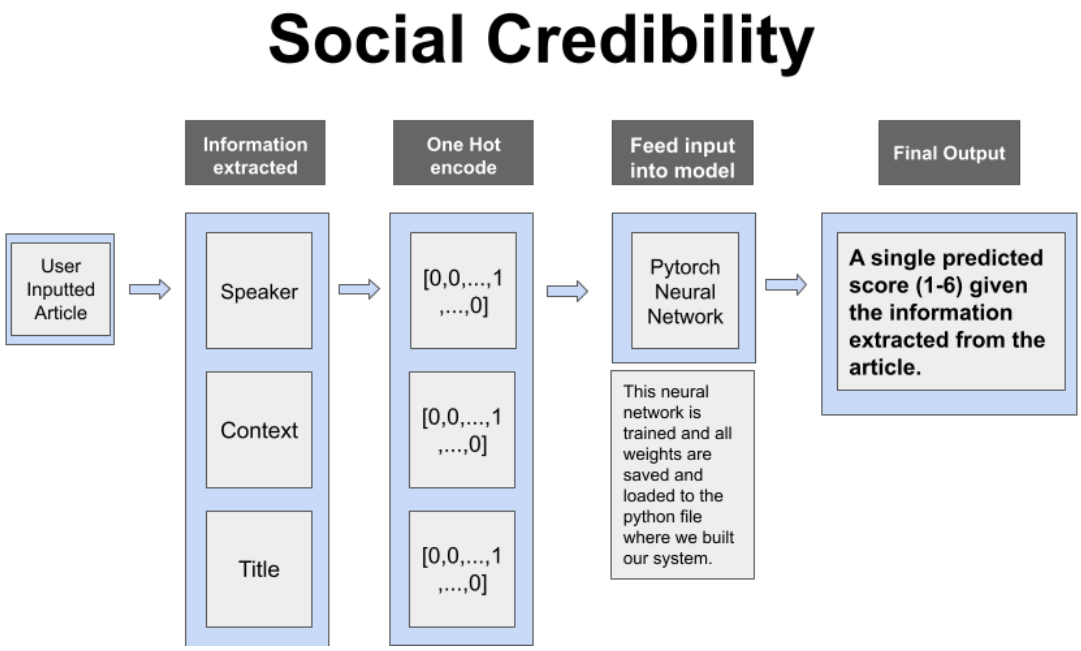

Figure 3: Social Credibility Workflow

When we incorporate each model into the Mesop interface, we extract the appropriate metadata to create scores for each microfactor. The article's title is used as the "statement" like from the Liar Plus dataset.

2.2 Generative AI

As recommended by our mentor, we utilized Gemini 1.5 Pro 002. We use generative AI in two ways:

1. Extracting key attributes from user-uploaded PDFs (e.g., speaker, context, title) to support predictive AI modeling.
2. Evaluating sensationalism and other factuality factors via structured prompts.

There are three different prompting techniques we could potentially use to evaluate facuality factors: normal, Chain of Thought (CoT), and Fractal Chain of Thought (FCoT). Each prompt consists of grade objective functions from 1-6, defined by microfactors of the factuality factors, and asking the model to grade an article based on the factuality factor from a scale of 1-6. CoT prompting involves asking the model to explain its reasoning, causing it to think more about its response and getting more accurate response. FCoT prompting asks the model to go through three iterations to explain its reasoning. After each iteration, we ask the model what it missed on its previous iterations. With each prompting type, we could also add combinations of querying the vector database or calling the Serp API.

2.2.1 Vector Databases

To help supply our generative AI model with more context, we utilized a vector database, as its optimal for similarity searches. We employed ChromaDB, an open-source vector database, to store and manage high-dimensional word embedding extracted from article headers. Within each collection, we store the documents (statements) and metadata (label, speaker, justification). This database is hosted within a persistent Docker container to ensure data retention. To ensure up-to-date information, we periodically update the vector database by scraping recent statements and promises from Politifact and Snopes. Using specific HTML parsing for each website, we extract the relevant metadata to input into the database.

When implementing the vector database into Mesop, we extract the title from the user-inputted article. Then, we perform a vector database query to find the top 3 similar statements. The similar statements and its metadata are provided to the generative AI model as a prompt.

2.2.2 Serp API

Another way we provide context for the generative AI model is through the Serp (Google Search) API. After extracting the title from the article, we search for related articles about the topic using the Serp API. The retrieved articles are processed using Newspaper3k to extract metadata (title, author, summary, publication date). The articles and its metadata are feed as the "ground truth" to the generative AI in relation to a topic to refine its evaluations.

There are three different types of prompting with 4 different adjustment combinations of vector database, Serp API, and None. For our Mesop interface, we want to offer the best option for the prompting type and adjustment combination. To determine the best options, we selected four different types of articles and asked four different people to grade the article based on a factuality factor. The article's true factuality factor score is based on the mean of the scores given by each grader. Then, each prompting type and adjustment combination is used on each article. Whichever article has the lowest mean absolute error is declared the best.

All in all, SERP, vector databases, prompting types, and articles are combined to provide the generative AI with the most context. This allow it to make a decision on whether a particular factuality factor is pant-on-fire to true.

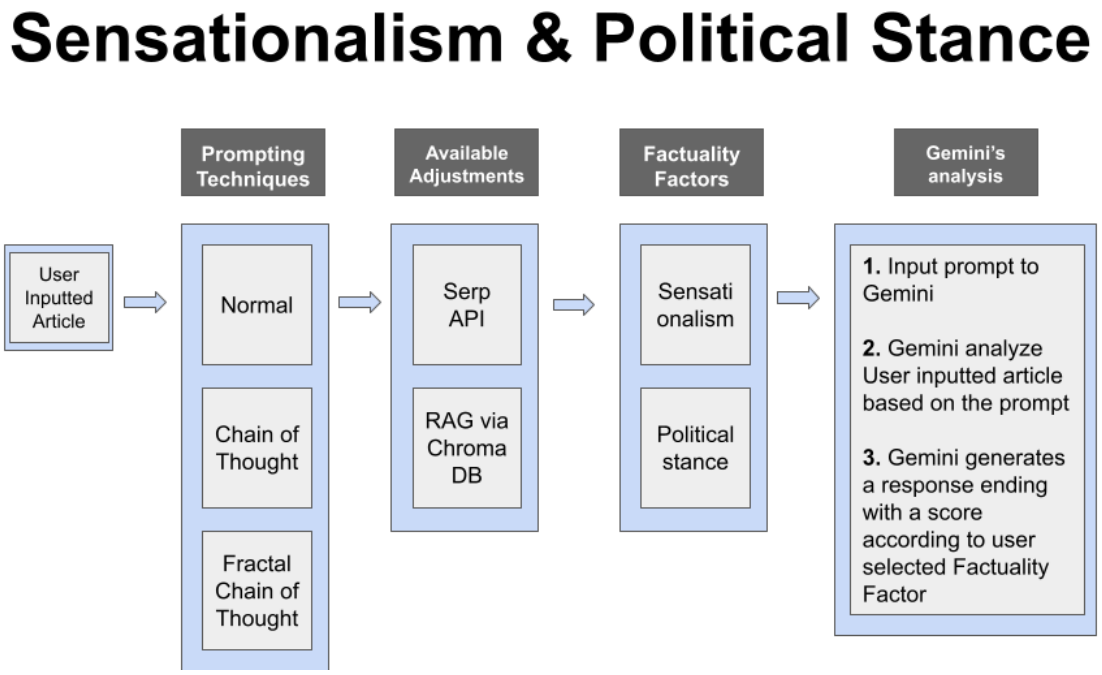

Figure 4: GenAI Flow

2.3 Website

For our website, we wanted our designs to look seamless, simple, trustworthy, and good. Additionally, we, not only wanted users to detect misinformation, but learn more about the pipeline, about this project, and to test out their own prompts as well. To design all of this properly, we utilized Figma to organize our ideas and design the website. We decided on the name Chenly Insights, as it combines the words for truth in Vietnamese and Chinese. The blue/green color theme communicates truth and accuracy.

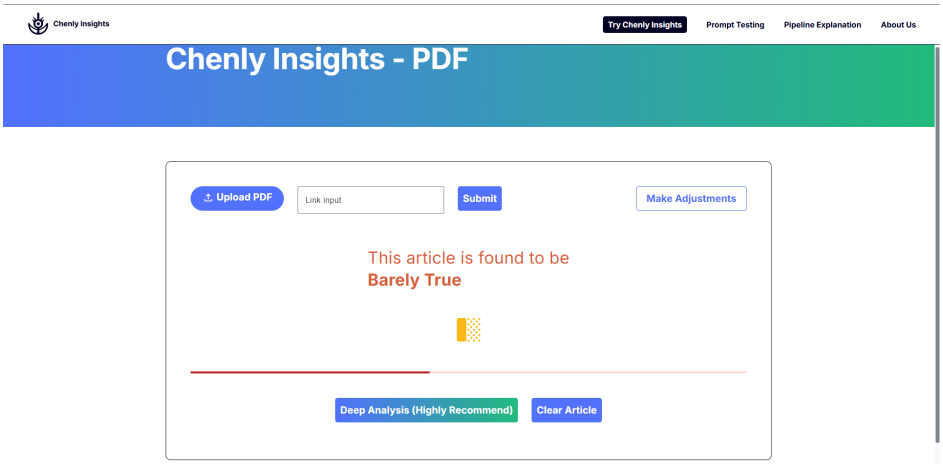

Figure 5: Results Page

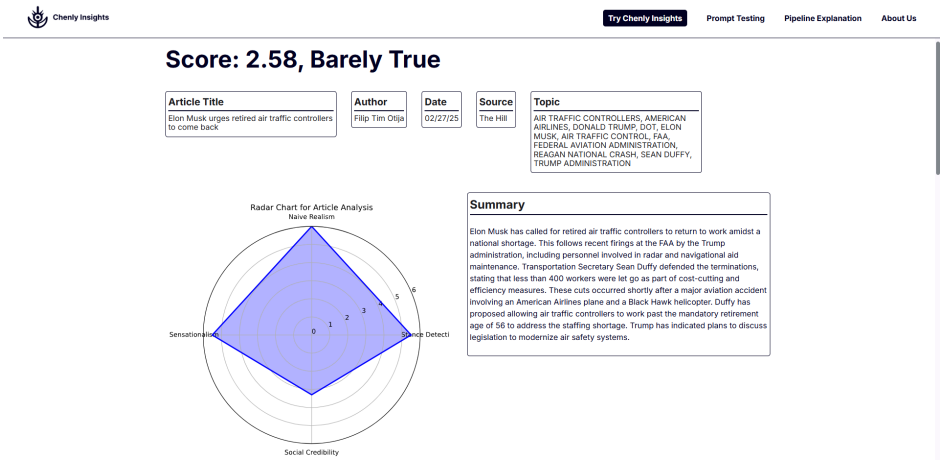

Figure 6: Deep Analysis Page

3 Results

Our results for our method is as follows from our grading of prompts from generative AI and evaluating our predictive AI.

Table 1: Prompting Types and Adjustments with MSE

Prompting Type	**Adjustment Made**	**MAE**
Normal	VB and SERP	1.5625
Normal	None	1.6875
FCoT	None	2.1050
Normal	VB	2.1125
CoT	SERP	2.5625
FCoT	SERP	2.8125
Normal	SERP	3.9375
FCoT	VB	3.9375
FCoT	VB and SERP	4.0125
CoT	VB	4.0625
CoT	VB and SERP	4.3125
CoT	None	4.5625

Table 2: Accuracy Table for two predictive AI Methods with a baseline of 20%

Factuality Factor	**Accuracy**
Naive Realism	44%
Social Credibility	21.36%

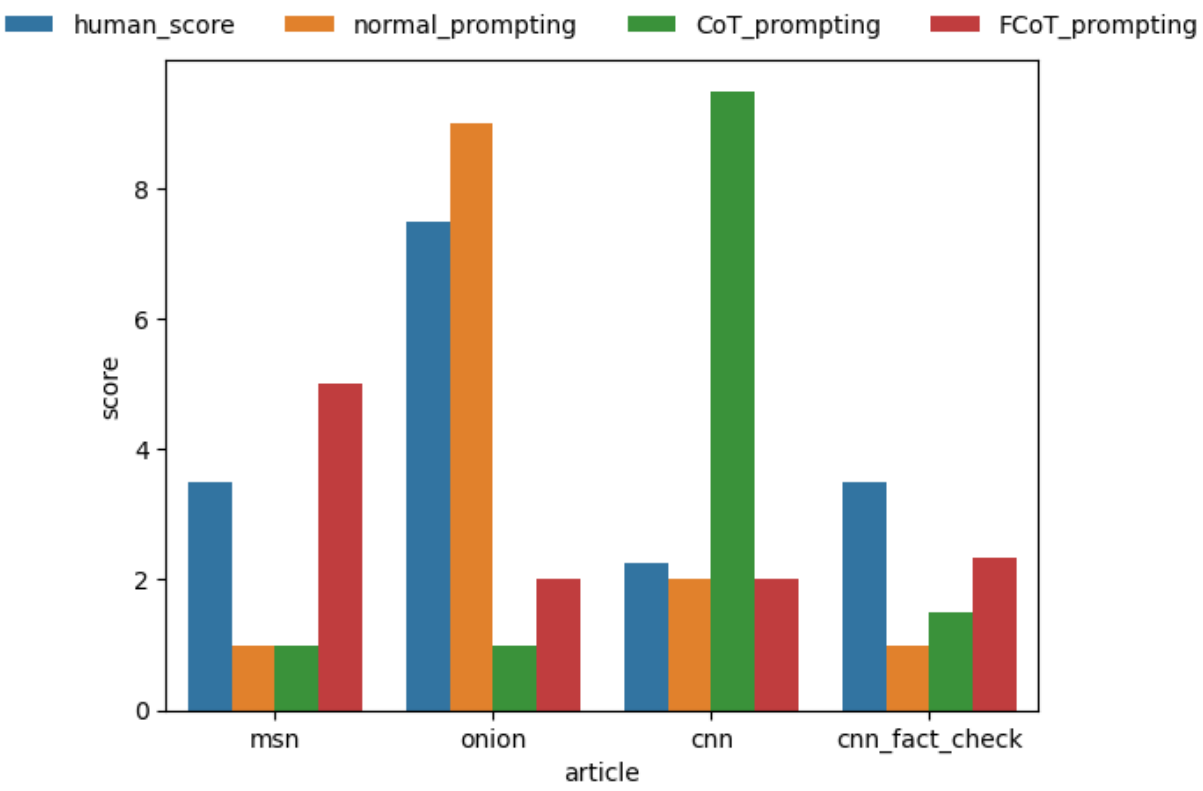

Figure 7: Accuracy Table for two AI Methods on different factuality factors

4 Conclusion

In our project, we've attempted to follow previous work's suggestions by combining predictive and generative artificial intelligence in detecting misinformation. We successfully developed a simple hybrid system that scores articles' veracity, and this is a significant step forward in addressing the complex challenge of misinformation in a digital age. This product can be run by anyone using our system.

There are a couple issues with this current process though. The MAE is highly based on the human scores, which is only based on a sample size of 4. Ideally, we would like a team of 20 expert human graders to grade each article on a factuality factor to get an article's true score.

Our hypothesis that FCoT with VB and SERP would receive the lowest MSE is wrong. Normal prompting with VB and SERP performed the best. This could be due to a couple reasons. Looking at the bar graph, it seems like FCoT struggled to judge our onion article accurately due to its satire, despite doing well for all the other articles. Additionally, we did notice that FCoT did grade each article more critically and provided more reasoning in relation to the microfactors, so it is possible that human graders did not notice that.

4.1 Future Work

Despite the progress, we made across two quarters for this project. We still had stretch goals that we could not unfortunately reach. For instance, we would have liked to combine generative and predictive AI for a single factuality. Initially, we tried through function calling, but the calls themselves were too inconsistent to be used for our project. Ideally, we would use an agent like **CrewAI** to allow generative AI to call predictive AI functions more consistently. To further improve our generative AI, we would adjust our prompting techniques to achieve more human-like scoring through longer and more specialized prompting. We will outline what results for pant-on-fire to true articles look like to the generative AI. With these changes, FCoT could become better than normal prompting like we hypothesized.

To improve our vector database, we would expand the collections to more data sources, such as Washington Post Fact Checker. This would allow us to catch holes present in our other two databases, Politifact and Snopes, and get a more nuanced view of same topic. Additionally, we would like to move this database to the Google Cloud. With a constantly live vector database, we can consistently add rows to each collection on a daily basis with jobs. This will ensure that our model is always up-to-date with the latest information, outside of the SERP. With a live database on Google Cloud, we would like to complement it with a fully live website as well. Users would need to use their own API keys, but besides that, it would be more convenient for users to run this tool, wihtout local host.

Finally, we would like to implement more factuality factors into our model. We would work with the other groups within the capstone to implement their models and generative AI techniques into our engine. This will be quite a lot of work, as we all had different philosophies. However, with much more factuality factors, our engine could grade articles more critically and accurately.

5 Contributions

Samantha

- Work on the initial modification to our UI/UX design for our mesop interface
- Read and ask one person to read articles to completing human scoring articles in order to evaluate our prompting combinations against human's judgment
- Created the table on our final analysis page with factuality factors, scores, evaluations, and sources.
- Work with Calvin together on lo-fi and hi-fi designs of our final product on Figma
- Work with Calvin together to build the pages for our final mesop interface
  - About Page, More Info Page, Scoring page
- Work with Calvin together to build our report, project webpage, and the poster

Calvin

- Worked with Samantha on lo-fi and hi-fi designs of Chenly Insights on Figma
- Read and ask one person to read articles to completing human scoring articles in

order to evaluate our prompting combinations against human's judgment
- Determine the best prompting type with proper adjustments through analysis
- Building UI for home page, pipeline explanation, customization, and engine pages itself
- Work with Calvin together to build our report, project webpage, and the poster

References

Jiang, Bohan, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024. "Disinformation Detection: An Evolving Challenge in the Age of LLMs." *arXiv preprint arXiv:2309.15847*. [Link]

Tariq60. 2018. "LIAR-PLUS." Oct. [Link]