ANLP Project - Fake News Detection and Analysis

Shahaf Valencia Levy
shahaf.valenciale@mail.huji.ac.il

Elisha Diskind
mosheel.diskind@mail.huji.ac.il

Guy Orlinsky
guy.orlinsky@mail.huji.ac.il

Noa Ben Gallim
noa.bengallim@mail.huji.ac.il

Abstract

Our research investigates the limitations of Large Language Models (LLMs) in distinguishing between fake and real news. In an era where fake news and misinformation contribute to widespread confusion and skepticism, no single authority can universally resolve these issues. Given the rapid advancements in LLM technology, we hypothesized that these models could offer effective solutions for fake news classification. To test this hypothesis, we evaluated several leading and state-of-the-art open-source LLMs using datasets containing both real and fake news articles. Our study involved gathering classifications from these models and analyzing their performance to uncover specific shortcomings in their ability to accurately identify and differentiate news content. The findings aim to shed light on the current challenges in LLM-based fake news detection and suggest areas for further improvement.

1 Introduction

Our team set out to investigate the limitations of Large Language Models (LLMs) in distinguishing between fake and real news. We hypothesized that modern LLM-based classifiers could offer a straightforward solution to this problem. To test this hypothesis, we employed various models to classify news reports in multiple variations. The focus was on recent models and datasets consisting of both real and fake news articles. We then analyzed the performance of these models to identify specific areas where they fall short in accurately classifying news content.

Our objective was to identify patterns in the data, particularly in misclassified articles, to understand the underlying reasons for each model's errors. We aim to provide insights that could lead to improvements in the accuracy of fake news classification. Our findings could potentially contribute to user wariness of fake news on certain topics beyond the use of classification models.

1.1 What is Fake News?

Part of the challenge we face is the definition of fake news (Wikipedia). Considering the complexity of the matter and the scope of our project, we have decided to treat fake news as any information that is not entirely factually correct. Consequently, our data is sourced from multiple entities, and we have accepted their interpretation and labeling of fake news under our broad definition.

2 Data

We used four publicly available datasets for both fine-tuning and inference in our experiments:

- GonzaloA/fake_news
- AlexanderHolmes0/true-fake-news
- BeardedJohn/FakeNews
- ikekobby/40-percent-cleaned-preprocessed-fake-real-news

We generated multiple datasets from each source, containing predictions and confidence scores produced by the models listed in 3.1 and their fine-tuned versions under various configurations. The datasets are available on Shahaf's Hugging Face profile.

2.1 GonzaloA Dataset

The GonzaloA dataset contains 40,587 news articles, with 24,353 used for training and 8,117 used for both validation and test. Each article includes a title, text, and a label indicating whether the article is real or fake. Table 1 presents an example row from the generated datasets, containing detailed information on the input configurations.

2.2 AlexanderHolmes and BeardedJohn Datasets

The AlexanderHolmes dataset consists of 44,896 news articles with 33,672 used for training and

11,224 for test. The BeardedJohn dataset includes 20,628 news articles, split into 10,612 for training, 3,281 for validation, and 6,735 for test. In both datasets, each article includes a title and a label indicating whether the article is real or fake. Table 2 presents an example row from the generated datasets, containing detailed information on the input configurations.

2.3 ikekobby Dataset

The ikekobby dataset consists of 17,959 news articles. Each article includes the original text, a cleaned version of the text (lowercase letters and no punctuation), and a label indicating whether the article is real or fake. Table 3 presents an example row from the generated datasets, containing detailed information on the input configurations.

3 Methods

Code available on GitHub repository.

3.1 Model Training

Model training was done using university resources, accessed via Shahaf. Experimenting with a *BERT*-based model highlighted a limitation of 'older' models: short input lengths. Our initial assumption was the longer the text, the easier it should be to predict whether a news report is real or not. With that in mind, the goal was to train newer models, taking into account their performance and longer input lengths. Training was done using Hugging Face's *trl* framework, fine-tuning Low-Rank Adapters (LoRA). The models we chose to train, in order, are:

- google/gemma-2-2b
- google/gemma-2-2b-it (also used as a baseline model)
- meta-llama/Meta-Llama-3.1-8B-Instruct (also used as a baseline model)

The models were fine-tuned using the standard *trl* script with CLI arguments and the provided hyper-parameters written by Google and Meta. All models were activated using *bfloat16*, and no quantization. Our first fine-tuning attempt involved the base Gemma model. We aimed to compare it to an instruction-tuned model on both unprompted and prompted data. The training was done on raw data with no prompts over a single dataset. The *instruct* models were fine-tuned on a modified version of the dataset, with an instruction prompt preceding the text. All models were trained on the text alone. The trained models are available on Shahaf's Hugging Face profile. Wandb training metrics could be found here.

3.2 Model Inference

To gather data for analysis, we ran a combination of all models across all datasets. Each dataset was tested in various configurations, using the title and text fields separately and combined, when both were available. The usage of the instruction prompt was also tested. Each dataset was tested on all splits, including the training data. All models were activated using the text-classification pipeline, with both the predicted label and score recorded. Additional models were used in our testing. In order to research the effectiveness of general-purpose LLMs (without fake-news-specific training), we used Gemma and Llama models, as listed above.

A fine-tuned albert-based model was also used, primarily to compare early language models with modern ones.

3.3 High Parameter Models

We were interested comparing the performance of the models we've chosen to OpenAI's and Anthropic's closed-source offerings. We've started by fine-tuning ChatGPT 4o on the data and faced a setback due to unexpectedly high costs. We've decided to not proceed following the cost of training. We've also tried evaluating our data with Anthropic's model but our runtime quickly exceeded the free usage limitation. All code and related trials are available for review here.

3.4 Common Words Masking

Given the relatively low accuracies of the models (Fig. 1 and Fig. 2), it appears that large language models (LLMs) have not inherently learned the nuances of fake news detection during their pre-training. Despite this, our models are still capable of classifying data, leading us to hypothesize that certain words might be influencing the models' classification decisions. Specifically, it is possible that some common words are guiding the models toward labeling data as either fake or real.

In the following link's you can find insights on the models results:

Accuracies
Confusion Matrices
Models Score Distributions

Inspired by Gaby's lecture on interpretability, we aim to investigate this hypothesis through a two-step counterfactual analysis.

In the first step, we will mask the common words identified within each label category—regardless of whether the label was predicted correctly or incorrectly (i.e., without distinguishing between true positives, true negatives, false positives, and false negatives). By doing this, we can observe how the model's predictions are affected when these common words are removed, providing an initial insight into their influence on the model's decision-making.

In the second step, we will refine our analysis by focusing on specific classification outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). We will systematically remove or mask the common words in each of these categories and analyze the resulting changes in predictions.

If the removal or masking of certain words leads to a change in the model's prediction, we can infer that these words had a significant impact on the model's decision-making process. By identifying these influential words, we can gain valuable insights into the underlying factors driving the model's classifications, thereby enhancing our understanding of how the model interprets and categorizes fake news.

To determine which words are "common words" within each class, we initially count the frequency of words found in the texts belonging to each class (such as fake or real news). However, simply counting word occurrences can lead to issues, as some words may be common across all classes, making it difficult to identify which words are truly distinctive for a particular class. Therefore, we need a more nuanced approach to filtering these common words.

Filtering Common Words Across Classes

One challenge we face is the potential imbalance in the number of examples across different classes. If one class has significantly more examples, raw word frequencies might be skewed, causing us to miss important words in the smaller class. To mitigate this, we normalize word frequencies within each class, ensuring that our comparisons are fair and not biased by class size.

We consider three normalization methods:

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Adjusts word frequency based on how commonly it appears across all classes, giving higher importance to words that are frequent in one class but rare in others.
- **Per-Document Normalization:** Normalizes word frequency by the number of documents in the class, accounting for varying document lengths and reflecting the word's relevance within each document.
- **Length Normalization:** Normalizes word frequency by the total number of words in the combined text of a class, reducing the impact of longer texts and ensuring that frequency reflects the word's relative importance.

By applying these normalization techniques and filtering strategies, we can more accurately identify the words that significantly influence the model's classification decisions. Each normalization method yields different results; therefore, we first visualized the filtered words using word clouds for each normalization technique. Based on these visualizations, we selected Length Normalization, as it provided the most meaningful and clear representation of the influential words compared to the other methods.

In the following link you can find WordClouds of different filtering GitHub Repository: Wordclouds.

4 Results

4.1 Common Words Masking/Omitting Results

The results of separating the data by labels to identify common words indeed suggest that these words are influencing the model's decisions. As shown in Fig. 3, most of the predictions that changed did so after masking only a small percentage of the words in the article. Overall, it appears that the majority of models did not exhibit a significant difference between masking the words and completely omitting them (Fig. 4). However, we also observe that the changes were notably biased in most models, often shifting predictions predominantly from fake to real or vice versa (Fig. 5).

5 On Model Aggregation

We acknowledge modern models, such as ChatGPT, have the capability to search across the internet for information. Models could aggregate sources to try

and verify information. Moreover, there are emerging news aggregators such as Ground News which try covering certain topics using various sources. As many reports are consumed instantly after publishing (in the form of tweets or news flashes), our aim is on smaller, potentially-on-device models to run independent analysis before definite verification.

6 Project Ethics and Limitations

We believe this project and similar ones have a positive impact on society and the information landscape. As our goal is to flag misinformation, we can utilize LLMs as tools to detect it and alert readers. The advent of LLMs and state actors has led to an increased spread of misinformation online (Verma). As with most models, a problem that might arise is users treating its output as ground-truth. Reminiscent of email spam, models can be tricked, and misinformation spreaders could attempt bypassing said models; whether researching the models themselves if they're open-sourced, or via black-box experimentation.

title	Trump's Favorite News Channel Tries To Soothe His Battered Ego - Gets Taken To The Cleaners
text	Yesterday, after the father of one of the UCLA players arrested in China failed to show Trump proper gratitude for getting his kid released, Trump, predictably, went to Twitter to grouse about it. He seems to expect to be worshiped for his help on this matter, but LaVar Ball wouldn't do it, so Trump tweeted: "Now that the three basketball players are out of China and saved from years in jail, LaVar Ball, the father of LiAngelo, is..."
label	0 (fake)
input	Your task is to classify the provided input as real or fake news. Label 0 is meant for fake news, Label 1 is for real news. TITLE: Trump's Favorite News Channel Tries To Soothe His Battered Ego - Gets Taken To The Cleaners TEXT: Yesterday, after the father of one of the UCLA players arrested in China failed to show Trump proper gratitude for getting his kid released, Trump, predictably, went to Twitter to grouse about it...
no_prompt_input	TITLE: Trump's Favorite News Channel Tries To Soothe His Battered Ego - Gets Taken To The Cleaners TEXT: Yesterday, after the father of one of the UCLA players arrested in China failed to show Trump proper gratitude for getting his kid released, Trump, predictably, went to Twitter to grouse about it...
pred	1 (real)
score	0.972204
pred_no_prompt	1 (real)
score_no_prompt	0.993807
pred_title	1 (real)
score_title	0.923039
pred_text	1 (real)
score_text	0.991684

Table 1: Example row from the datasets generated using GonzaloA dataset. For each model, predictions and confidence scores were recorded under different configurations, including the full input (with the prompt, title, and text), input without the prompt, title-only input, and text-only input.

label	0 (fake)
text	A group of white nationalists and skinheads who support Donald Trump's presidential campaign who were involved in a stabbing attack in California have announced that they will be at the Republican National Convention this July in Cleveland. The group, who call themselves the Traditionalist Worker Party revealed their plans. A group of white nationalists and skinheads who held a rally in Sacramento over the weekend where at least five people were stabbed plan to show up at the...
input	Your task is to classify the provided input as real or fake news. Label 0 is meant for fake news, Label 1 is for real news. TEXT: A group of white nationalists and skinheads who support Donald Trump's presidential campaign who were involved in a stabbing attack in California have announced that they will be at the Republican National Convention this July in Cleveland. The group, who call themselves the Traditionalist Worker Party...
pred	0 (fake)
score	0.983279
pred_text	0 (fake)
score_text	0.999534

Table 2: Example row from the datasets generated using the AlexanderHolmes and BeardedJohndatasets. For each model, we recorded predictions and confidence scores using both the full input (which includes the prompt) and the input without the prompt.

article	MATT DAMON Says America Needs IMMEDIATE GUN BAN After Making $50 MILLION KILLING People With Guns In Popular Movie SeriesIt warms the heart to know people as important as Hollywood actor Matt Damon care so much about the little people with guns in America. If Damon really feels so strongly about the government taking our Second Amendment right away, perhaps he should stop making a living with guns Matt Damon will return to the big screen as trained assassin Jason Bourne later this month, and while promoting the new film in Australia over the weekend called for a ban on guns in the United States...
label	0 (fake)
clean_article	matt damon says america needs immediate gun ban after making million killing people with guns in popular movie series warms heart know people important hollywood actor matt damon care much little people guns america if damon really feels strongly government taking second amendment right away perhaps stop making living guns matt damon return big screen trained assassin jason borne later month promoting new film australia weekend called ban guns united interview reporter sydney morning...
input	Your task is to classify the provided input as real or fake news. Label 0 is meant for fake news, Label 1 is for real news. TEXT: MATT DAMON Says America Needs IMMEDIATE GUN BAN After Making $50 MILLION KILLING People With Guns In Popular Movie SeriesIt warms the heart to know people as important as Hollywood actor Matt Damon care so much about the little people with guns in America. If Damon really feels so strongly about the government taking our Second Amendment right away, perhaps he should stop making a living with guns Matt Damon will return to the big screen as trained assassin Jason Bourne later this month, and while promoting the new film in Australia over the weekend called for a ban on guns in the United States...
clean_input	Your task is to classify the provided input as real or fake news. Label 0 is meant for fake news, Label 1 is for real news. TEXT: matt damon says america needs immediate gun ban after making million killing people with guns in popular movie series warms heart know people important hollywood actor matt damon care much little people guns america if damon really feels strongly government taking second amendment right away perhaps stop making living guns matt damon return big screen trained assassin jason borne later month promoting new film australia weekend called ban guns united interview reporter sydney morning...
pred	0 (fake)
score	0.998968
pred_clean_input	0 (fake)
score_clean_input	0.999777
pred_article	0 (fake)
score_article	0.999279
pred_clean_article	0 (fake)
score_clean_article	0.999815

Table 3: Example row from the datasets generated using ikekobby dataset. For each model, we recorded predictions and confidence scores under various configurations, such as the full input (with the prompt and original article), the full input with the cleaned article, the original article without the prompt, and the cleaned article without the prompt.

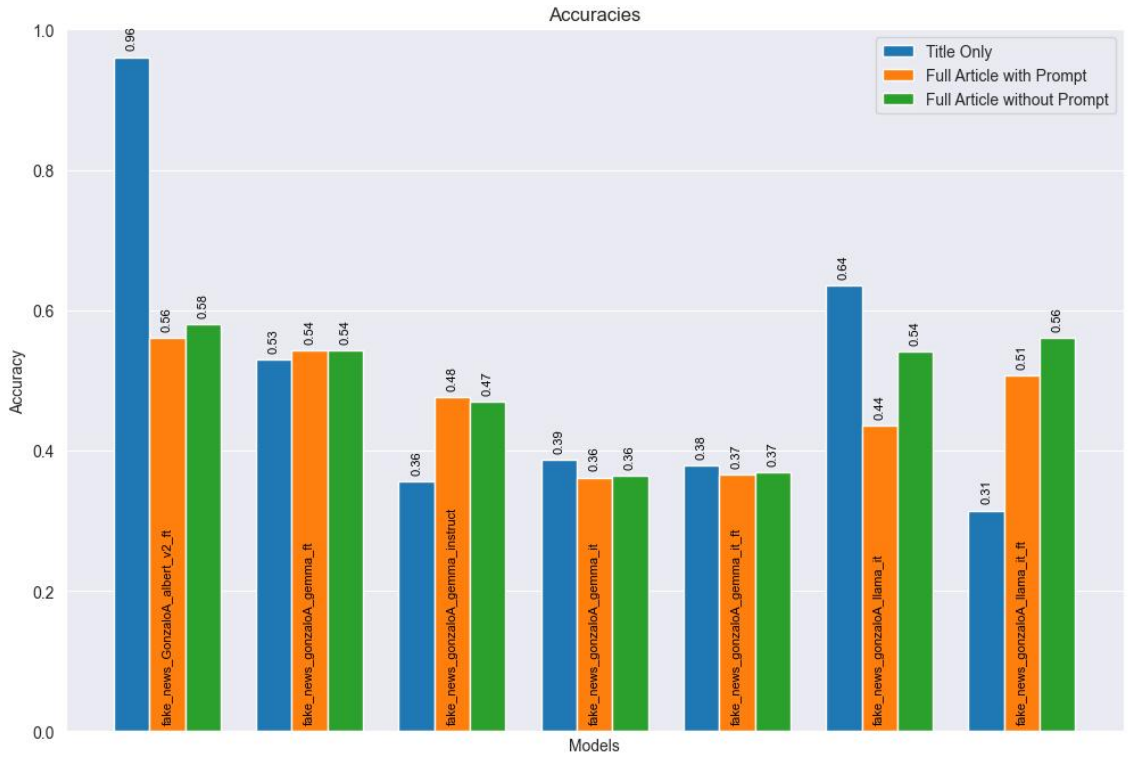

Figure 1: Model's accuracy rating on GonzaloA datasets with different inputs.

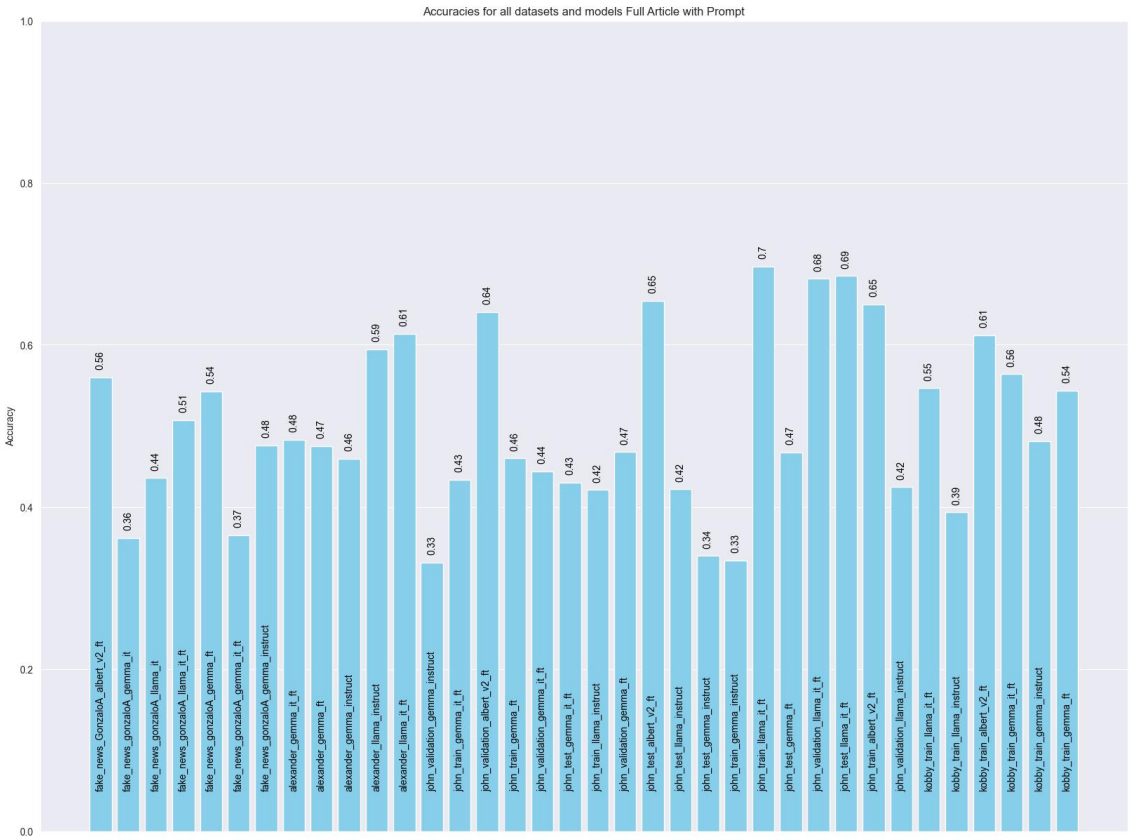

Figure 2: Model's accuracy rating on different data sets with instruct prompt.

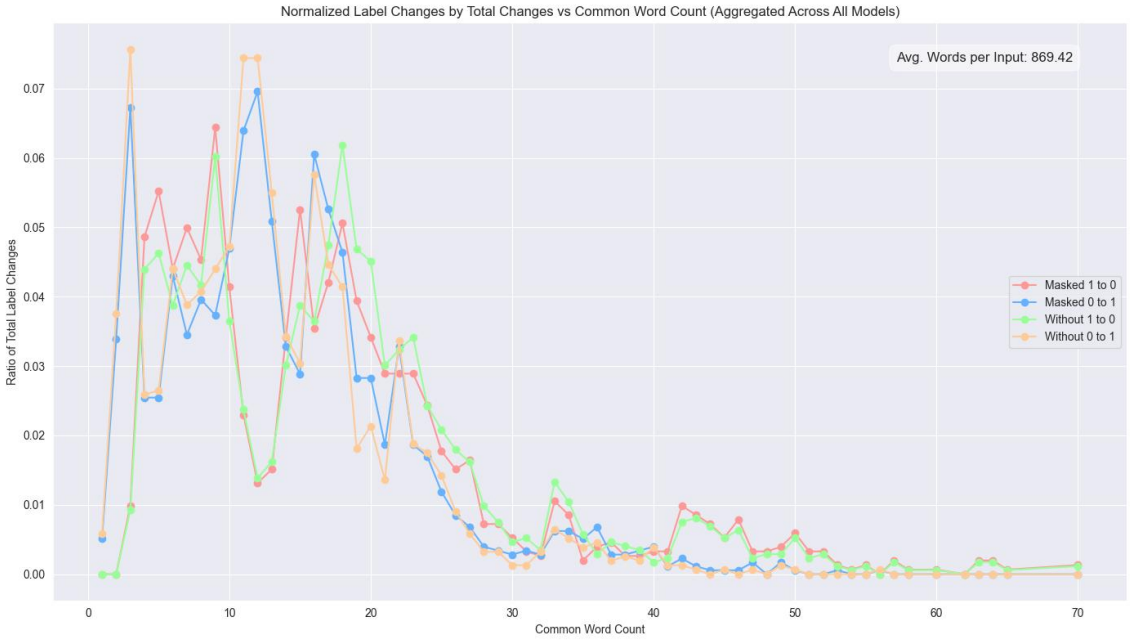

Figure 3: Normalized label changes by total changes vs. common word count aggregated across all models (when counting common words by labels predicted).

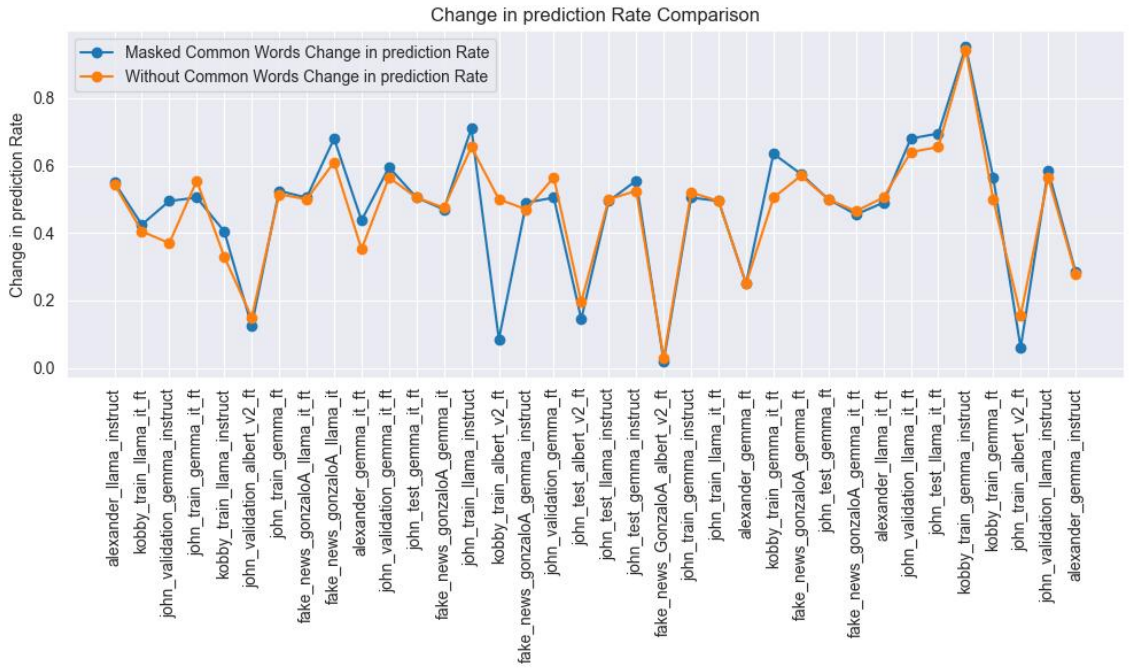

Figure 4: Change in prediction rate comparison (when counting common words by labels predicted).

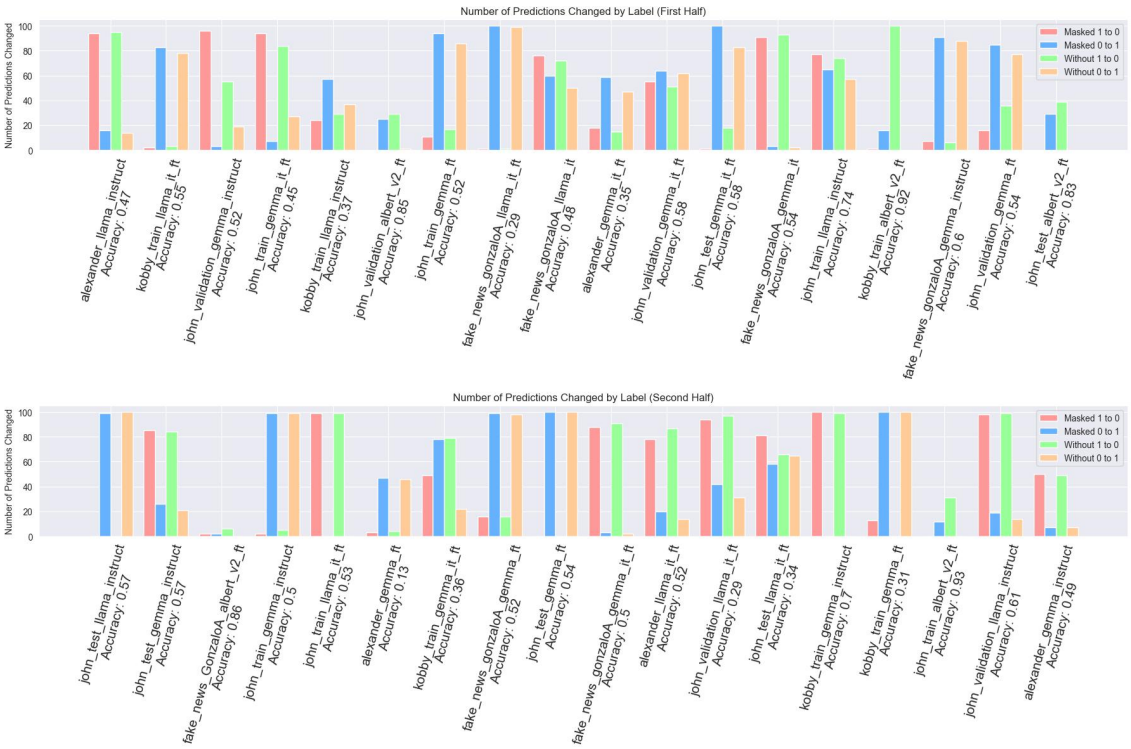

Figure 5: Number of predictions changed by label (when counting common words by labels predicted).

References

Pranshu Verma. The rise of ai fake news is creating a 'misinformation superspreader'. https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/.

Wikipedia. Defining fake news. https://en.wikipedia.org/wiki/Fake_news#Defining_fake_news.