MASK-IT: Browser Extension Which Masks Misleading Content Using Large Language Model

Dev Bhojak
Simon Fraser University
dbhojak@sfu.ca

Abstract

In today's digital landscape, passive browsing frequently exposes individuals to misleading and rumored news content, posing significant risk of spreading misinformation among vulnerable user groups such as minors. MASK-IT is a browser extension aimed at addressing this issue by utilizing a large language model to classify the content in the browser window. Furthermore, the plugin would also provide the user option to mask such content. This project involves applying parameter efficient finetuning RoBERTa on a dataset of ∼20,000 news articles acquired from PolitiFact. The importance of MASK-IT lies in its potential to provide a more positive online surfing experience while safeguarding users from misinformation.

1 Introduction

This paper is a work in progress, with ongoing improvements happening throughtout the Spring 2025 semester. I plan to submit it to NeurIPS and EMNLP later this year.

The internet serves as a critical platform for information, communication, and social interaction. However, passive browsing can inadvertently expose users to misleading information, which can significantly harm mental health by fostering confusion, fear, anxiety, and depression (Rocha et al., 2023).

There have been many studies in the past focusing on the negative effects of rumoured or misleading new content. Jin et al. (2017), describes how the rumoured content on twitter influenced voter decisions during the 2016 US presidential elections. They found that false information significantly shaped public opinion, contributing to the spread of misinformation and affecting the outcome of the election. Similarly, a study by Grinberg et al. (2019) analyzed the consumption of fake news during the 2016 election and found that a small segment of the population, primarily those with conservative leanings, were highly susceptible to fake news, which in turn influenced their voting behavior.

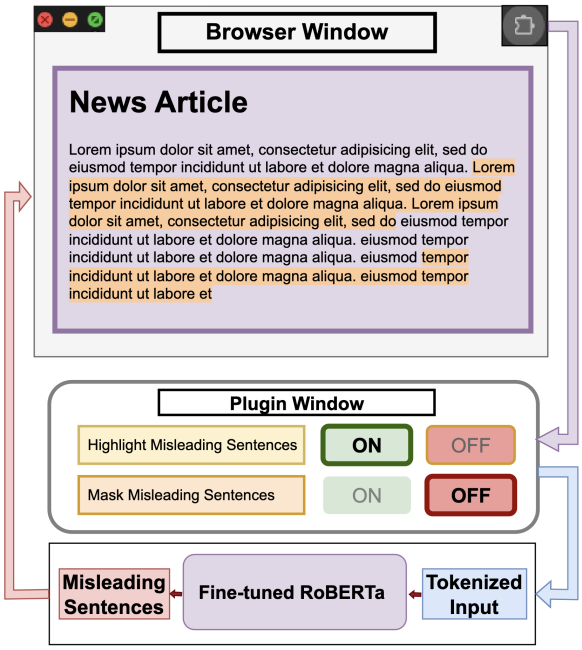


Figure 1: Overview of the MASK-IT Browser Extension Workflow.

In another study, Vosoughi et al. (2018) examined the spread of true and false news online. Their findings revealed that false news spreads more rapidly and broadly than true news, primarily due to its novelty and the emotional responses it elicits from readers. This rapid dissemination of false information can lead to significant societal impacts, including misinformed decision-making.

Additionally, Pennycook and Rand (2018) investigated the role of cognitive factors in the susceptibility to fake news. They found that individuals who engage in more analytical thinking are less likely to believe and spread false information. Their study underscores the importance of promoting critical thinking skills to combat the effects of

misinformation.

MASK-IT provides users an option to highlight potential misinformation and help be more informed while browsing the internet. Following sections describes the PolitiFact dataset, methodology used for finetuning RoBERTa-base model , broswer plugin and future work to be carried out during the semester

2 Methodology

2.1 Dataset:

Dataset was generated by scraping the PolitiFact website. Below are the statistics for the final dataset:

Table 1: Total of **20,825** PolitiFact Articles From 2007 to 2024 were used during finetuning

Ruling	**Articles Count**
Pants On Fire	3369
False	4006
Mostly False	3703
Half True	3755
Mostly True	3430
True	2562
Total	**20825**

2.2 Data Exploration:

To understand the data distribution and uncover bias, below data analysis steps were performed. The following insights were derived:

1. **Token Analysis:**
    - The total number of tokens in claims was computed using a RoBERTa tokenizer, which is as follows:

Column	**Mean**	**Std.**	**Min**	**Max**
Claim	24	10	4	109
Explanation	993	427	27	4118

2. **Party Affiliation Distribution:**
    - After filtering out records with "none" as party affiliation, party affiliation counts were analyzed. Only parties with a minimum of 9 articles were considered. **Top-10 Parties by Article Counts** are listed in **Table 2**

Table 2: Top-10 Counts of Articles by Party Affiliation

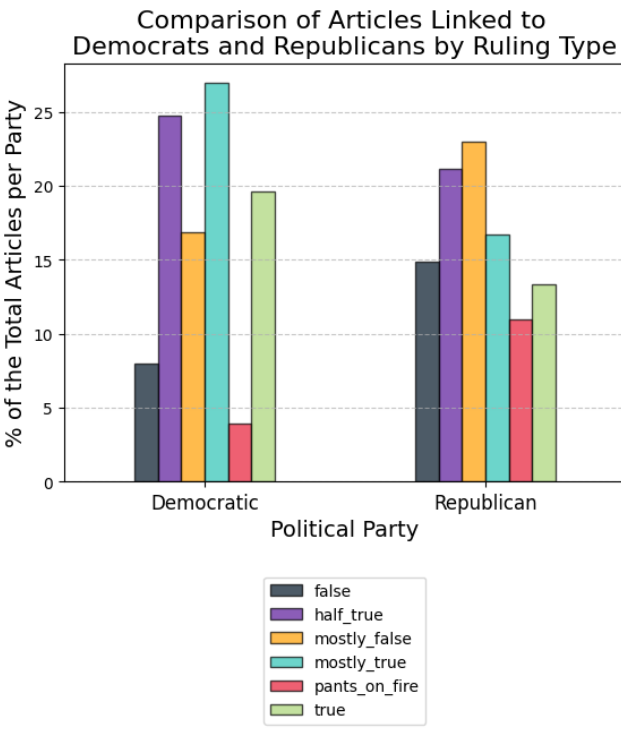
Party Affiliation	**Articles Count**
Republican	6860
Democratic	5085
Independent	291
Organization	167
State Official	61
Newsmaker	53
Libertarian	44
Journalist	43
Columnist	43
Activist	40
None	8138

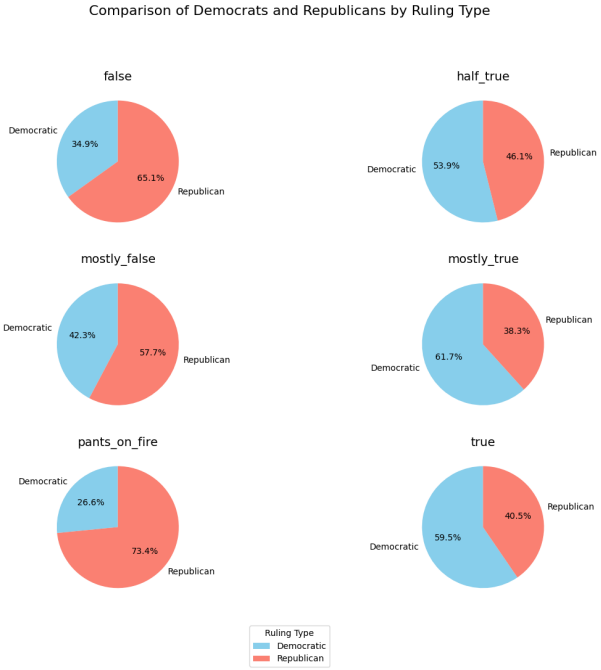
3. **Verdict Type Analysis Across Parties:**
    - Articles affiliated with the Democratic and Republican parties were grouped by ruling types.
    - Normalization by total articles per party allowed comparison in percentage terms across ruling types.
    - Visualization:
        - **Pie & Bar Charts:** Below figures depict how much percentage of articles in each ruling type are contributed by Democratic and Republican party.

2.3 Pre-processing:

Before fine-tuning, a comprehensive preprocessing pipeline was implemented to prepare the data effectively for training. The following steps were performed to clean, extract, and transform the data into the desired format:

1. **Data Cleaning and Organization:**
    - Names without political affiliations were identified and cleaned by capitalizing words and fixing naming inconsistencies (e.g., replacing hyphens with spaces).
    - Wikipedia was queried to fetch information from infoboxes using a custom scraper. Extracted party affiliations were parsed using regular expressions to locate fields such as `| party = [[Party Name]]`.





- Mislabelled affiliations (e.g., "Republican Party") were standardized to uniform representations (e.g., "Republican", "Democratic").

2. **Data Filtering and Balancing:**
   - Records with missing or unavailable political affiliations (marked as "party_not_found" or "none") were removed.
   - To address class imbalance, 4,000 false "ruling" labels were randomly dropped. The discarded samples were saved in CSV format for future reference.
3. **Tokenization with Sliding Window:**
   - Special tokens (`<link>`, `<sing_link>`, and `</link>`) were added to the tokenizer to distinguish embedded URLs in the text.
   - Claims were concatenated with explanations (both linked and unlinked) and tokenized using a sliding window approach to handle lengthy sequences:
     - **Sliding Window Details:** Claims were always positioned at the beginning of each token chunk.
     - Overlapping strides (256 tokens) ensured full coverage of explanations. The tokenized output was padded or truncated to a maximum length of 512 tokens.
4. **Structured Representation of Data:**
   - The tokenized data was split into **encoded chunks** (token IDs) and **decoded chunks** (textual representations) for training, validation, and test datasets.
5. **Data Splitting:**
   - Finally the filtered dataset was split into training (80%), validation (10%), and test (10%) sets.

2.4 Model

2.4.1 Overview

The model used for the plugin is the RoBERTa base model (Liu et al., 2019), fine-tuned on the curated dataset using the Low-Rank Adaptation (LoRA) technique (Hu et al., 2021). LoRA reduces the number of trainable parameters, making the fine-tuning process faster and less resource-intensive while maintaining high performance. Choice of the model and the parameter efficient fine-tuning was based on available computing resources.

2.4.2 Input and Output

- **Input:** Tokenized sentences extracted from news articles using the RoBERTa tokenizer, preprocessed with additional tokens (e.g., ‘<link>‘, ‘</link>‘).
- **Output:** Classification labels indicating the truthfulness of each statement by returning a label from the following list [‘true‘, ‘false‘, ‘half-true‘]. These outputs are utilized by the MASK-IT browser extension to highlight false sentences as Red, true sentences as green and half-true as black.

2.4.3 Training Strategy

- **Configuration:**

Table 3: LoRA Parameters

Parameter	Value
Rank (r)	8
Alpha (α)	32
Dropout (p)	0.1
Task Type	SEQ_CLS

Table 4: Training Arguments

Parameter	Value
Learning Rate (η)	5e-5
Batch Size	32
Epochs	5
Weight Decay	0.01

2.5 Firefox Plugin

The core functionality of MASK-IT is depicted in figure 1. As users browse content and sends the selected text to the model, it evaluates the text for misleading information and the plugin highlights the text. Users can toggle the masking and highlighting options via the plugin’s pop-up interface, allowing for control over their browsing environment.

2.5.1 Plugin Workflow

1. **User Interaction**: The user interacts with the browser extension through a popup window.
2. **Content Extraction**: When the user navigates to a news article, they can select part of the text to evaluate and then click on the "Evaluate Text" option to send the text to the model.
3. **Call To Model**: The extracted text is sent to the finetuned RoBERTa model in **ONNX** format uploaded to the HuggingFace repository and the text is tokenized and processed to generate predictions.
4. **Prediction Handling**: The model returns predictions indicating which sentences are misleading.
5. **Content Modification**: Based on the predictions, the plugin highlights text in the news article, {red

3 Model Evaluation:

Model selection The veracity prediction performance of the MASK-IT model on the testing subset (**10%**) of the dataset.

- **F1-Score**: The F1-Score is the harmonic mean of precision and recall. This metric is useful to ensure both the correct identification of misleading sentences and the avoidance of false positives. The F1-Score can be calculated using the formula:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Precision**: Precision measures the accuracy of the model’s positive predictions, indicating the proportion of true misleading sentences among all sentences flagged as misleading by the model. It can be calculated using the formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**: Recall measures the model’s ability to identify all actual misleading sentences in the dataset. It can be calculated using the formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **Accuracy**: Accuracy measures the overall correctness of the model by calculating the proportion of true positive and true negative predictions out of all predictions made. While accuracy provides a general sense of the model’s

performance, it is less informative than F1-score in the context of imbalanced datasets. It can be calculated using the formula:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

3.0.1 Results and Evaluation

Below is performance report for validation and test dataset, detailed images are provided under Appendix A:

- Test Accuracy: `79.5` %
- Test Precision: `0.7929`
- Test Recall: `0.7950`
- Test F1 Score: `0.7883`
- Test Loss: `0.4586`

4 Ongoing Work

- Testing the model after incorporating The Hourglass of Emotions, i.e. the emotion categorisation model (Cambria et al., 2011)
- Extending current dataset by including additonal data from following sources:
  - https://www.factcheckinsights.org/
  - ISOT Fake News Dataset ($\sim$ 44000 entries)
  - https://www.snopes.com/
- Hyperparameter tuning using Optuna framework

References

E. Cambria, Andrew G. Livingstone, and Amir Hussain. 2011. The hourglass of emotions. In *COST 2102 Training School*.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Gordon Pennycook and David G Rand. 2018. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Management Science*, 66(11):4944–4957.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2023. The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. *Journal of Public Health*, 31(7):1007–1016.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

A Appendix: Test Evaluation Results

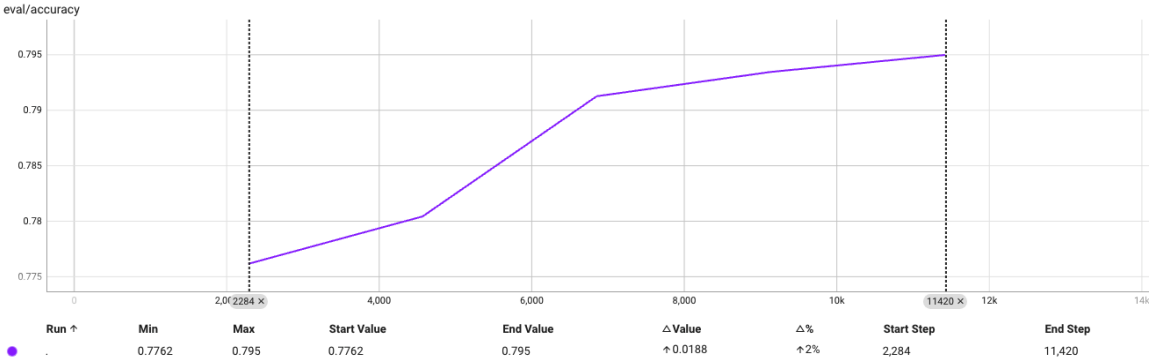


Figure 4: **Test Accuracy: 79.5 %**

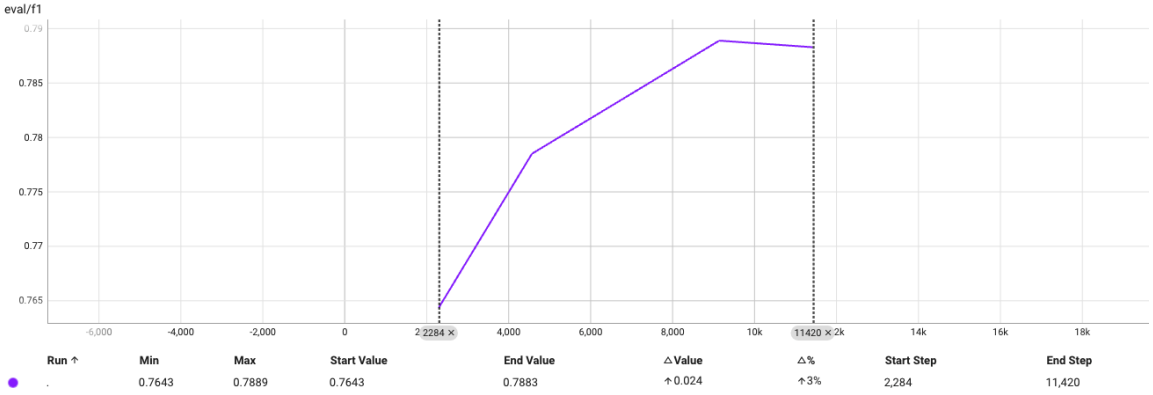


Figure 5: **Test F1: 0.7883**

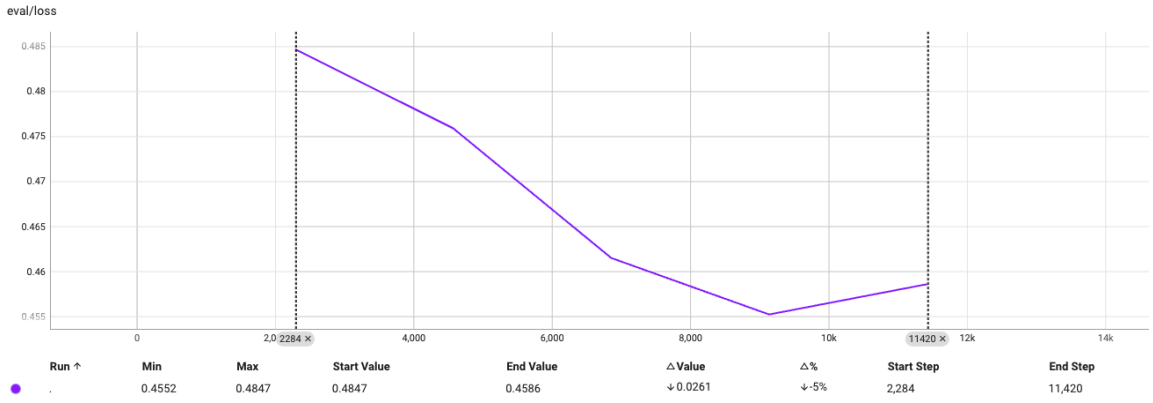


Figure 6: **Test Loss: 0.4586**