Detecting Misinformation in News Articles via Natural Language Processing

Stanford CS224N Custom Project

Siya Goel
Department of Computer Science
Stanford University
siyagoel@stanford.edu

Tia Vasudeva
Department of Computer Science
Stanford University
tvasudev@stanford.edu

Thu Le
Department of Computer Science
Stanford University
lethu@stanford.edu

Abstract

Misinformation in news has become a significant issue in the United States, affecting major American political events and discussion. We seek to semantically understand and automatically authenticate validity of information, through applications of a RoBERTa model, GPT prompting, and a backpack model. GPT prompting revealed distinct linguistic patterns: genuine articles tend to have complex structures with citations, while fake news often uses simpler, conversational language. Additionally, training the Backpack Model on false content illustrated its potential to propagate misinformation, emphasizing the critical need for accurate detection mechanisms to prevent misinformation spread. Our RoBERTa model also achieved an accuracy of 99.70%, which outperforms the WELFake model by 3.13%. We additionally evaluate the RoBERTa on historical speeches and random statements, highlighting limitations in classification. Misinformation detection is also transferrable between languages, with a $\geq 97.90\%$ accuracy on a Japanese full misinformation dataset.

1 Key Information

Mentor: Nelson Liu (Custom Project)
Split: Thu worked on RoBERTa and prompting, Siya worked on RoBERTa classification and analysis, and Tia worked on Backpack LLMs/semantic analysis

2 Introduction

Since the advent of social media, there has been a proliferation of misinformation, impacting a variety of sociopolitical climates and public discourse. The spread of such misinformation has led to the polarization of public opinion over events such as the 2016 U.S. election, the COVID-19 pandemic, and the 2020 U.S. election. Given the significant reliance on social media for communication and information, it is imperative to develop methods to detect misinformation on these platforms and remove them.

Detection of misinformation is not a novel application of machine learning models. Previous literature has applied a variety of ML methods to detect misinformation (Verma et al., 2021) (Shaikh and Patil, 2020) (Khanam et al., 2021) (Ahmed et al., 2017) (Gravanis et al., 2019) (Shu et al., 2018). Because ML based methods exhibit remarkable performance in detection, deep learning methods have been explored. However, natural language processing models have not been thoroughly explored.

We seek to improve on Verma et al. (2021)'s WELFake model, utilizing a variety of unexplored approaches that show promising potential in addressing a highly significant and relevant issue. Notably, we implemented GPT prompting to reveal linguistic and semantic patterns across misinformation, trained and evaluated a RoBERTa model on WELFake and additional data, and fine-tuned GPT2 and a Backpack LLM on misinformation to understand how training data affects machine output.

3 Related Work

Prior research on text-based misinformation detection is reviewed in Zhang and Ghorbani (2020), covering various methods including analysis of content creators, the misinformation content, and its distribution context. Our study narrows this focus solely to the analysis of the textual content of misinformation, excluding considerations of its source and dissemination context.

Zhang and Ghorbani (2020) notes that dense misinformation in texts like reviews or tweets is somewhat detectable via basic linguistic methods like bag-of-words or n-gram. However, these approaches often fail to uncover deeper misinformation patterns, necessitating more advanced embedding techniques to fully grasp the complex opinions and semantics in news content for effective detection.

Supervised machine learning algorithms like Decision Trees, Random Forest, SVM, Logistic Regression, and K-nearest Neighbors are effective in misinformation detection (Zhang and Ghorbani, 2020). Shaikh and Patil (2020) highlights the success of an SVM model with 95.05% accuracy. Khanam et al. (2021) finds Naïve Bayes algorithms widely used, achieving 70-76% precision. Verma et al. (2021) introduces a model using linguistic features and embeddings, leading to an ensemble model with SVM and count vectorizer as the best performer at 96.73% accuracy, surpassing BERT and CNN models by 1.31% and 4.25%, respectively. Verma et al. (2021) also outperforms models in several other studies like Ahmed et al. (2017), Gravanis et al. (2019), and Shu et al. (2018).

Additionally, deep learning approaches have demonstrated better results, with automated feature extraction, reduced dependency on data pre-processing, and the capability to extract high-dimensional features, resulting in improved accuracy. In consequence, new research in fake news detection has been predominantly deep learning-based (Mridha et al., 2021). However, no research has been done on the applications of sophisticated natural language processing models in the context of fake news.

4 Approach

4.1 Baseline

Our baseline is the WELfake accuracy which sits just above 96% (Verma et al., 2021), derived from training and testing with traditional ML techniques. Our approach looks at semantic properties when making predictions, hoping to increase accuracy.

4.2 Semantics

We utilize GPT 3.5, OpenAI's Generative Pre-trained Transformer, specifically through the OpenAI API and a Jupyter Notebook employing the `gpt-3.5-turbo-instruct` engine. The transformer contains randomness temperature set to 0.5 and a maximum of 300 generated tokens. Our implementation involves:

1. Few-shot prompting to understand the semantic distinctions between informative and misinformative content. This method involves generating predictions after learning from a small number of examples. By providing an example of both an informative and a misinformative article, we attempt to classify a third unseen article.
2. Chain-Of-Thought (CoT) prompting to evaluate GPT 3.5's ability in common sense reasoning and its efficacy in applying knowledge from one domain to another. Given a set of eight articles (four informative and four misinformative), we aim to semantically differentiate between the two categories.

The Backpack model, a probabilistic transformer-based approach developed by John Hewitt et al, uniquely encodes sense vectors for words, capturing their non-contextual uses in different contexts. Its novel interpretability through control by adjusting sense weights, show promise for reducing bias in outputs. Our focus is on its sense vector representations(Hewitt et al., 2023).

4.3 Classification

RoBERTa, developed by Facebook AI Research, enhances the BERT model with longer training, larger batches, and more data (160 GB) (Liu et al., 2019). It employs Byte-Pair Encoding with a 50,000-word vocabulary and eschews BERT's next sentence prediction for a focus on Masked

Language Modeling (MLM). This improves semantic understanding by dynamically masking 15% of input tokens for bidirectional prediction each epoch.

We use code from various GitHub repositories. Particularly, for backpacking, we utilize the source code from the original paper (Hewitt et al., 2023). We modified the distilBERT model to be a RoBERTa architecture and used this code as a basis Chahed (2024).

5 Experiments

5.1 Data

The Fake News Classification dataset is utilized in this project (Verma et al., 2021). It comprises 72,134 news articles, with 35,028 labeled as factual and 37,106 identified as containing misinformation/fake from four distinct sources: Kaggle, McIntire, Reuters, and BuzzFeed Political. The dataset includes the news article's content and a label indicating the presence of misinformation which is used to fine-tune/test the RoBERTa model, GPT prompting, and the backpack model. We additionally tested the model on more generalized data (Yetim, 2022). We also fine-tuned and evaluated the RoBERTa model on Japanese news articles, with each article classified as containing fully, partial, and no misinformation (Tanreinama, 2021).

We additionally evaluated the RoBERTa model on historical speeches from sources like Adolf Hitler (Cinnamon, 2024), Benito Mussolini (Immerman, 2024), Joseph McCarthy (University of Houston, 2021), and Joe Biden (Kessler, 2024). Random statements were self-written, comprised of paradoxes (e.g. "this is a lie"), debated statements (e.g. "Taiwan is a country"), and opinion pieces (e.g. "Trump is the best president") were also tested.

We also used a Backpack, a Misinformed Backpack, GPT2, and a Misinformed GPT2 in order to generate text to semantically analyze through Flair and graphical methods.

5.2 Evaluation method

5.2.1 Semantics

We evaluated the GPT model qualitatively, depending on the responses the model gave us and the reasoning for these responses. Additionally, results from the sentiment analysis Hugging Face model were analyzed by investigating the amount of samples that were negative/positive and the confidence score for the analysis.

The fine-tuned Backpack (which we will refer to as the Misinformed Backpack) was evaluated against Backpack, using TF-IDF scores. We also evaluated a Misinformed GPT2 against a Misinformed Backpack. TF-IDF scores represent the similarity of generated text compared to the text it was fine-tuned on, determining what words in the original text were significant in influencing the Backpack. Furthermore, we derive a relationship between TD-IDF score and semantic analysis score (from negative to positive) to determine if fine-tuning on misinformation tends to produce more negative/positive text.

We go deeper into the semantic analysis by analyzing semantic dependencies through a network representation of the graph, using a modified Dijkstra's to find the shortest path between word associations. We compare the shortest paths of word pairings between the regular and Misinformed versions of the Backpack. Later, we dig deeper into the Backpack architecture to obtain the sense vector representations from the Stanford Backpack model, and the sense vector representations from a Misinformed Backpack model. We aim to distinguish how these representations change as the data the model is trained upon is changed.

5.2.2 Classification

We evaluate the RoBERTa model (Liu et al., 2019) based on accuracy, specificity, precision, and recall, in order to comprehensively analyze the model's performance. Formulas on how to calculate these scores are shown in Table 6.

Accuracy looks at the holistic measure of a test's or model's overall performance by showing the amount of true results among all the cases. Specificity measures the proportion of actual negatives correctly identified. Precision measures the proportion of positive identifications actually correct. Recall is important as it looks at the proportion of actual positives identified by the model.

5.3 Experimental details

5.3.1 Semantics

We sampled informative and untruthful articles for GPT 3.5 analysis, dividing them into misinformative (Group 1) and informative(Group 2) categories. We examined semantic, connotative, and syntactical distinctions and used few-shot and Chain-of-Thought (CoT) prompting for classification and reasoning, guided by identified semantic and connotative criteria. Figure 8 demonstrates an example of CoT and few-shot prompting on GPT to analyze and classify articles.

Further analysis involved "siebert/sentiment-roberta-large-english" and "austinmw/distilbert-base-uncased-fine-tuned-tweets-sentiment" models from Hugging Face. The former was fine-tuned on 15 diverse datasets, and the latter on tweet evaluations, to study the semantic orientations of informative and misinformative content through their classification outcomes and sentiment scores.

5.3.2 Backpack LLM

We investigated Backpack architecture's sense vector generation via two approaches: 1) Brief modification of the HeadModel class for fine-tuning and text generation based on prompts, and 2) Adapting the Hugging Face model, trained with the openWeb text and WELFake dataset, to test its response to controversial prompts and assess its misinformation prediction capabilities.

We analyzed sense vectors in Misinformed and Regular Backpack models using the HeadModel class, tokenizing input text to create a sense network. We projected sense vectors onto the vocabulary space, calculating dot product similarities with each vocabulary token through matrix multiplication with the language model head. This process helped identify the most similar token to each sense vector, allowing us to decode and understand the encoded senses.

We applied the analysis to both Misinformed and Regular Backpack models, using LDA analysis to discern five distinct topic groups from sense vectors, facilitating a qualitative comparison of semantic encodings from different text generators. Additionally, we examined semantic dependencies within the generated texts, constructing a network to identify and measure the closeness of word associations based on the models' fine-tuning, using the shortest path metric between words to analyze associations.

5.3.3 Classification

We employed a pre-trained RoBERTa tokenizer to encode the "x" dataset, specifying parameters like `max_length=100`, `truncation=true`, and `padding=true`, which required 8 minutes to fine-tune. This process, alongside the "y" dataset labeling, facilitated the construction of a tfdataset, segmented into various training and testing proportions for fine-tuning. A TensorFlow RoBERTa model was initialized for sequence classification, leveraging an Adam optimizer, and underwent fine-tuning across these dataset splits. The fine-tuning performance for each epoch of the different training/testing splits is shown in Table 7. The model's generalization was also evaluated on the Kaggle "Fake News Detection Dataset." Various experiments were conducted to see how well the RoBERTa model classified misinformation and information:

Speeches and One-Liners: The RoBERTa model was evaluated on various speeches and one-liners to assess its generalization across different content types.

Japanese Dataset: The model underwent fine-tuning with a dataset of Japanese news articles varying in misinformation levels. Three tests were performed to discern fully, partially, and combined misinformation levels against informative content. The fine-tuning performance is shown in Figure 9.

Generative Text Experimentation: A subset of the WELFake dataset was used for text generation with GPT-2. The generated content then served to fine-tune RoBERTa, which was subsequently tested on the remaining dataset. The fine-tuning performance for the generative model is detailed in Table 8.

5.4 Results

5.4.1 Semantics Using GPT and Hugging Face Models

GPT analysis showed misinformation uses simple language, less opinionated, and negative tones, while informative text is complex, bias, and positive. GPT accurately classified articles, shown in Table 9 and 10. Models "siebert/sentiment-roberta-large-english" and "austinmw/distilbert-base-uncased-fine-tuned-tweets-sentiment" indicated misinformation leans negative, contrasting with the balanced sentiments in informative content, as seen in Figures 1 and 2.

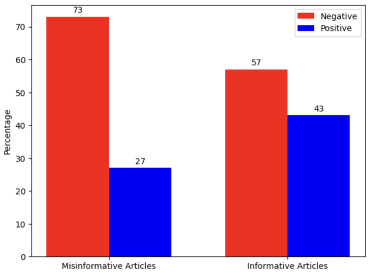


Figure 1: Sentiment Analysis Using Siebert Model

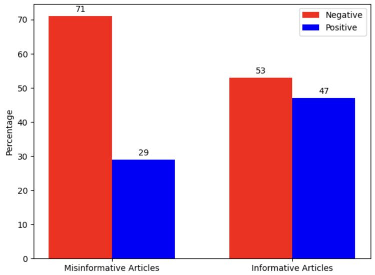


Figure 2: Sentiment Analysis Using AustinMW Model

5.4.2 Backpack LLM

We analyzed the impact of misinformed training data by comparing word frequencies in Misinformed and Regular Backpack models. Higher usage in the Misinformed model points to fine-tuning effects. For instance, the Regular model's preference for "The Last Days of the Church" contrasts with the fine-tuned model's inclination towards "The Art of the Deal," reflecting its training emphasis.

By associating each word with a sentiment score and analyzing its TF-IDF scores in misinformed models as shown in Figure 10, we identify a correlation between term frequency and sentiment intensity. Higher TF-IDF scores, indicating frequent model output, generally align with more extreme sentiment values: positive scores imply positive sentiments, and negative scores indicate negative sentiments. The distribution shows a positive sentiment bias, with extreme sentiments more likely as word frequency increases. Comparative analysis between the fine-tuned Backpack and GPT models (Figures 3 and 4) shows similar sentiment distributions, attributed to their common GPT-2 base and fine-tuning methods (Figure 5), with differences considered minor and likely due to dataset size or randomness.

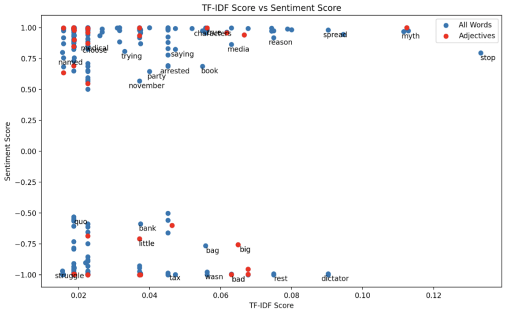


Figure 3: Misinformed model

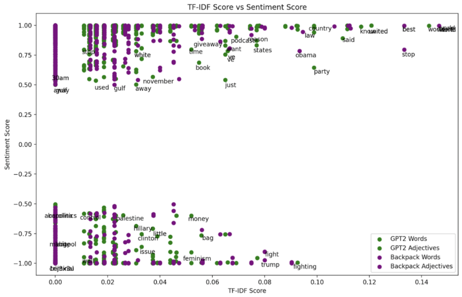
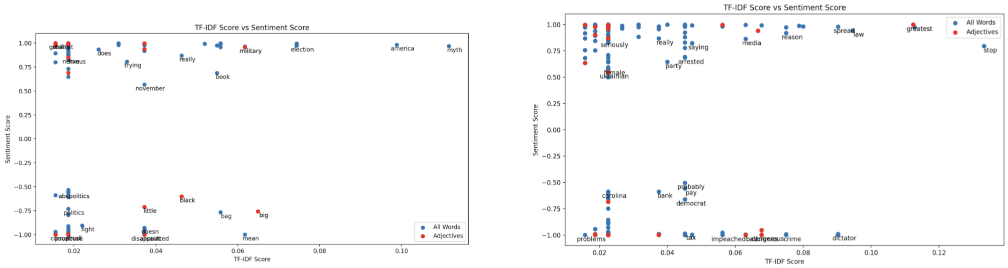


Figure 4: GPT-2 vs Backpacking



(a) Backpack (not fine-tuned)

(b) Backpack (fine-tuned)

Figure 5: Comparison of Backpack (not fine-tuned) and Backpack (fine-tuned)

We also separate our data by model, and pay special attention to adjectives. This allows us to perform a side by side comparison between the outputs of the two models where we see that the Misinformed model has more data points overall, confirming that the fine-tuning affected the nature of the model's output. The Misinformed model also displays more extreme data, with sightly more adjectives (although this relationship is weak and would have to be confirmed with more generated data).

5.4.3 Semantic Associations

We analyzed the semantic associations unique to the Misinformed Model compared to the Regular Backpack by mapping generated text into a semantic network and using a modified Dijkstra's algorithm to measure word association strength as shown in Figure 11. Our findings indicate that the Misinformed Model predominantly forms associations that are pro-American and anti-foreign, with examples like (empire, Russia) and (America, free), rather than showing overt racism or sexism. Additionally, the Misinformed Model tends to be more opinionated, resembling a "politician-like" tone, particularly in its portrayal of American themes. Consider the following question: "What is America like?" The Regular Backpack exhibits repetitive patterns, likely due to implementation flaws rather than deliberate design (Table 1). Comparing this with the responses from GPT2 and Misinformed GPT2, the latter demonstrates a more grandiose style, frequently using quotations (Table 2). This suggests its training on unverified semantic data significantly influences its output.

Misinformed Backpack	**Regular Backpack**
It s a melting pot of white people and minorities living in a country that is divided into two categories: those who are voting and those who are not voting. It s a country that has no tolerance for anyone who doesn't fit the description given above	The answer is simple: it is a country of immigrants. The answer is simple: it is a country of immigrants. The answer is simple: it is a country of immigrants.

Table 1: Comparison of Misinformed and Regular Backpack Statements

GPT2	**Misinformed GPT2**
America is a country of immigrants. It is not a nation of people who are born here. America is an immigrant country. The United States is the only country in the world that has a higher percentage of the population of non-citizens than the United Kingdom. In fact, the percentage is higher than any other country on earth. This is because the majority of Americans are immigrants, and the vast majority are not. They are the ones who have been here for generations.	America is a country of immigrants. It is not a nation of people who are born here. America is an immigrant country. ... "America has a history of being a place where people come from all over the world. We have a culture of tolerance and tolerance of all kinds of cultures. And we have an immigration policy that is very much in line with what we believe in."

Table 2: Comparison of GPT2 and Misinformed GPT2 Statements

The analysis of the word 'America' in sense vector representations reveals unique associations in the Misinformed model, such as 'sworn,' 'patriotic,' 'ney,' 'lore,' 'ANA,' and 'Britain,' which are absent in the Regular model. These terms, particularly 'patriotic' and 'sworn,' highlight an America-centric bias in the Misinformed model. By examining these internal representations, we can identify and potentially rectify biases to prevent the propagation of misinformation, thus ensuring that the model does not inadvertently embed false information as truth within societal perceptions.

5.4.4 RoBERTa Results and Comparisons

Table 11 shows all training/testing splits had over 99% accuracy, using a 70%/30% split is order to be consistent with WELFake. Our model exceeded WELFake's by 3.13%, compared against KNN, SVM, Naive Bayes, and others, including BERT and CNN from WELFake's creators (Table 3). Similar generalization and testing accuracies were found, detailed in the confusion matrix (Figure 12).

Model	Accuracy (%)	Precision (%)	Recall (%)
RoBERTa (this model)	99.86	99.95	99.78
RoBERTa (this model generalized)	99.88	99.99	99.78
WELFake	96.73	—	—
WELFake's BERT	93.79	—	—
WELFake's CNN	92.48	—	—
WELFake's KNN	90.16	89.02	90.55
WELFake's SVM	96.73	94.60	91.85
WELFake's Naive Bayes	92.12	91.45	92.25
WELFake's Decision Tree	89.92	86.10	92.62
WELFake's Bagging	95.31	91.78	95.00
WELFake's AdaBoost	95.32	91.81	95.02

Table 3: Comparison of Model Performance Metrics with WELFake

5.4.5 RoBERTa Experimentation

Testing on one-liners and speeches yielded a lower accuracy of 88.23%, with the model failing to classify four specific items correctly, as detailed in the confusion matrix (Figure 6). These misclassified items included McCarthy's speech (true value being misinformation) and three one-liners like "Siya said the word love", "Orcinus orcas are killer whales is informative", and "Cinderella has a glass slipper is informative" (true values being informative). Furthermore, the model's accuracy on the State of the Union Address was significantly lower at 27.27%, failing to identify any misinformation, as shown in the corresponding confusion matrix (Figure 7).

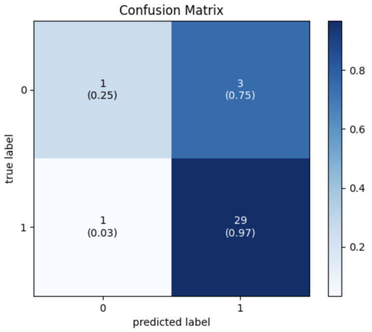


Figure 6: Confusion Matrix of the One Liners and Speeches

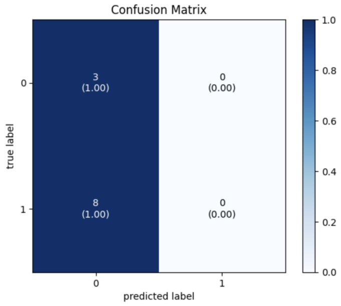


Figure 7: Confusion Matrix of the State of The Union Speech

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Japanese Full Misinformation	97.91	97.33	99.06
Japanese Partial Misinformation	87.99	95.81	82.33
Japanese and WELFake Full Misinformation	97.94	96.39	99.30

Table 4: Performance Metrics of Model on Japanese Dataset

The data from Table 4 indicates that the model accurately predicted fully misleading samples with over 97% accuracy for both the Japanese and the combined WELFake and Japanese datasets. However, its accuracy dropped to approximately 87.99% for partially misleading content, with 18% of such samples mistakenly classified as informative, as detailed in the confusion matrix (Figure 13).

Table 5 shows that model accuracy decreased by approximately 30% when fine-tuned on generated data, compared to original WELFake data with a 95% training and 5% testing split. The confusion matrix in Figure 14 illustrates this decline, highlighting that when the model is fine-tuned on generated data, it misclassified 33% of misinformative articles as informative and 29% of informative articles as misinformative.

Training Data	Precision	Recall	Accuracy
Generative Data	66.25%	70.94%	69.03%
Original WELFake Data	99.44%	99.40%	99.40%

Table 5: Model Performance Metrics with Generated vs Non-Generated Data

6 Analysis

6.1 Key Insights From GPT and Hugging Face Models Semantics

The GPT analysis revealed that informative groups had factual, unbiased content, while misinformative groups contained biased, less factual information, often with shorter, more emotional sentences. Misinformative articles also featured more negative phrases, potentially due to their tendency to spread harmful misinformation. These findings demonstrate that large language models (LLMs) like GPT can discern between informative and misinformative content based on syntax and semantics, underscoring the potential of advanced LLM classification methods like RoBERTa.

6.2 The Semantic Network

The analysis showed that the Misinformed Backpack grouped words implying American adversaries as threats more closely than the Regular Backpack and similarly grouped pro-American rhetoric. This likely stems from the misinformation-laden fine-tuning dataset, which frequently used pro-American quotes. The Misinformed Backpack's tendency to view foreign nationals as threats reflects a bias observed in its training data, although it's unclear why this specific bias was prominent. Consequently, the Misinformed Backpack's output often appears biased and propagandistic, exemplified by its preference for terms like "Great America."

6.3 RoBERTa Results and Comparisons

The RoBERTa model outperformed WELFake, BERT, and CNN models, largely due to its ability to discern semantic differences between texts. RoBERTa's enhancements over BERT, including more data, extended training, and dynamic masking, contribute to its improved performance. Additionally, RoBERTa demonstrated large capabilities of generalization by effectively applying its learned semantic distinctions to a different dataset.

6.4 RoBERTa Experimentation

6.4.1 Speeches and One-Liners

The model misclassified McCarthy's speech as informative, possibly due to the subtlety of propaganda. It also struggled with statements containing specific proper nouns or facts, suggesting a need for training and fine-tuning on diverse proper nouns. Furthermore, the model failed to identify misinformation in Biden's State of the Union address, likely because misleading content was presented as fact. Future enhancements should include contextual data to improve differentiation between factual and non-factual content, not relying solely on semantic cues.

6.4.2 Japanese Dataset

The RoBERTa model effectively identified semantic differences in both Japanese and multilingual datasets, despite being initially trained on English data. However, its accuracy decreased when assessing Japanese content with partial misinformation. This is likely due to challenges in distinguishing subtler semantic nuances between partial misinformation and truth, and a need for more extensive Japanese data to enhance differentiation.

6.4.3 Generative Test Experimentation

The model's accuracy dropped when fine-tuned on generative data, as the generated data was manipulated. Specifically, the generation caused a lot of misinformative news to have semantics similar to informative news and informative news sounded more similar to misinformative news. This illustrates the potential negative impacts of relying on generated data for training models.

7 Conclusion

Semantic analysis revealed that GPT and Hugging Face models could differentiate between misinformative and informative articles, with the former characterized by negative tones, bias, and shorter sentences, and the latter by positive tones, objectivity, and longer sentences.

The RoBERTa model demonstrated high accuracy in distinguishing misinformative articles from informative articles in the WELFake dataset and across various languages and data types, including Japanese, speeches, and debated statements. This success is attributed to its capability to discern semantic distinctions between misinformation and factual content. Our findings also indicate that misinformation often promotes a pro-America, anti-foreign narrative, with semantic intensity increasing as specific words recur. The structure of the models offers insights into the information they generate, suggesting a direction toward more interpretable models.

However, the model struggled with detecting partial misinformation, particularly in Japanese, and failed to recognize statements with unfamiliar proper nouns and semantics in generative data. Future research should enhance model performance on speech and propaganda by incorporating political and current events data for nuanced semantic understanding. Additionally, the corpus of the training and fine-tuning dataset should contain more languages. Lastly, the impact of generative data on classification accuracy through iterative feedback loops involving generation and fine-tuning could be explored.

References

H. Ahmed, I. Traore, and S. Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent Secure and Dependable Systems in Distributed and Cloud Environments*, volume 10618, pages 127–138, Cham, Switzerland. Springer.

Mohamed Chahed. 2024. Fine-tuning distilbert (fake news). `https://www.kaggle.com/code/mohamedchahed/fine-tuning-distilbert-fake-news/notebook`. Accessed on 1 March 2024.

Serina Cinnamon. 2024. Hitler's speeches.

G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang. 2023. Backpack language models.

Richard H. Immerman. 2024. Speech delivered by premier benito mussolini.

Glenn Kessler. 2024. Fact-checking president biden's 2024 state of the union address. *The Washington Post*. Accessed: Insert date accessed.

Z Khanam, B N Alwasel, H Sirafi, and M Rashid. 2021. Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, 1099(1):012040.

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.

M. F. Mridha, Ashfia Jannat Keya, Md. Abdul Hamid, Muhammad Mostafa Monowar, and Md. Saifur Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170.

Jasmine Shaikh and Rupali Patil. 2020. Fake news detection using machine learning. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–5.

K. Shu, S. Wang, and H. Liu. 2018. Exploiting tri-relationship for fake news detection.

Tanreinama. 2021. Japanese fake news dataset. `https://www.kaggle.com/datasets/tanreinama/japanese-fakenews-dataset`. Accessed: <insert date>.

University of Houston. 2021. Digital history.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: Word embedding over linguistic features for fake news detection. volume 8, pages 881–893.

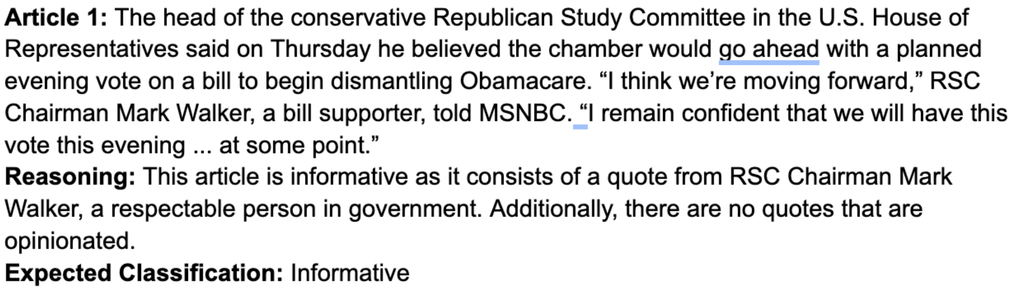
Emine Yetim. 2022. Fake news detection datasets. `https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets`. Accessed: <insert date>.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

A Appendix (optional)

Accuracy	Specificity	Precision	Recall
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TN}{TN+FP}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

Table 6: Summary of Metrics Formulas



Article 1: The head of the conservative Republican Study Committee in the U.S. House of Representatives said on Thursday he believed the chamber would go ahead with a planned evening vote on a bill to begin dismantling Obamacare. "I think we're moving forward," RSC Chairman Mark Walker, a bill supporter, told MSNBC. "I remain confident that we will have this vote this evening ... at some point."

Reasoning: This article is informative as it consists of a quote from RSC Chairman Mark Walker, a respectable person in government. Additionally, there are no quotes that are opinionated.

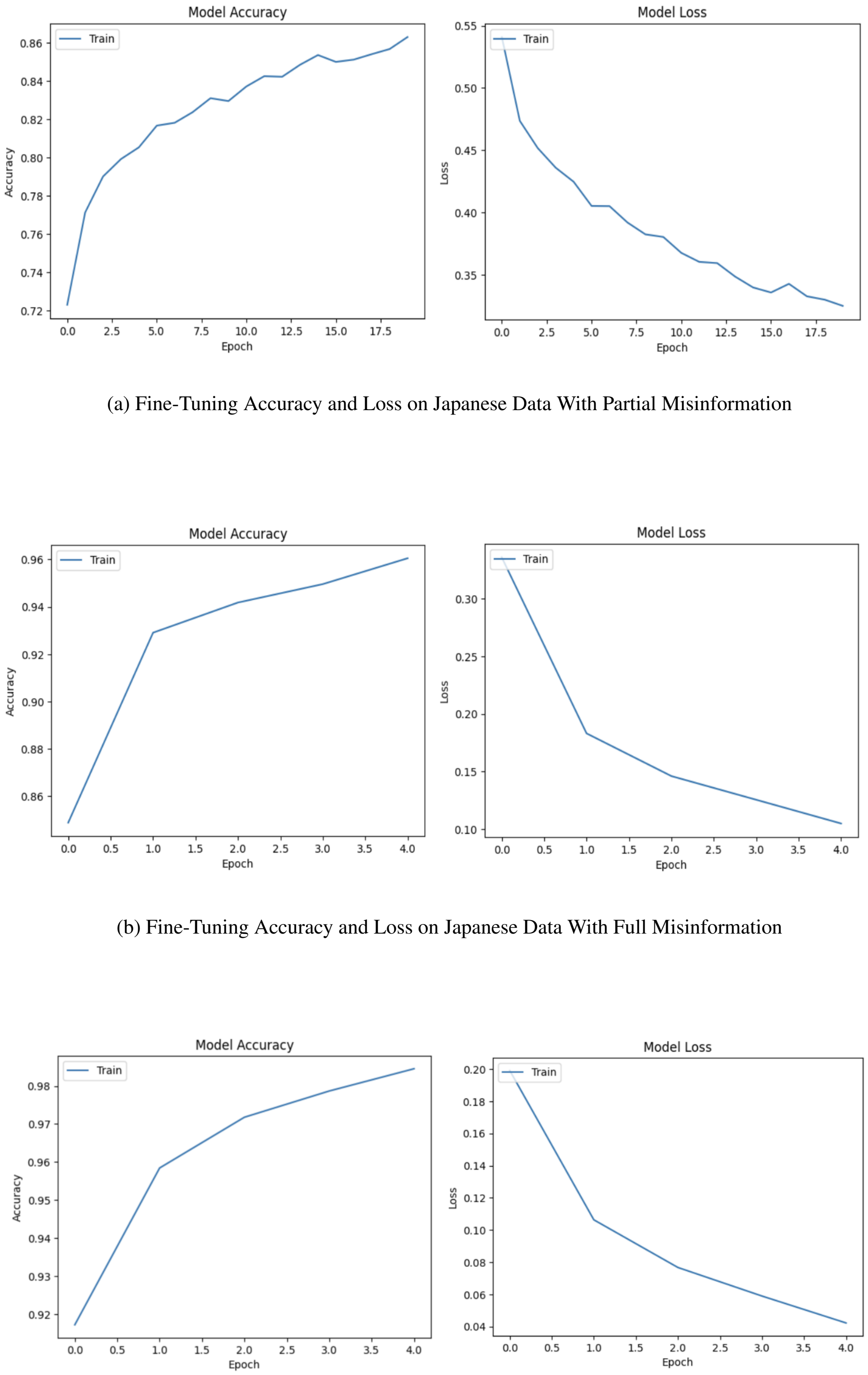
Expected Classification: Informative

Figure 8: Example Few-Shot and CoT Prompt for GPT Prompting

Split	Epoch	Time (seconds)	Loss	Accuracy
5% train, 95% test	1	124	0.1322	0.9429
	2	63	0.0446	0.9836
	3	60	0.0357	0.9872
10% train, 90% test	1	1179	0.0308	0.9898
	2	1108	0.0119	0.9961
	3	1098	0.0090	0.9973
20% train, 80% test	1	1047	0.0302	0.9897
	2	986	0.0129	0.9960
	3	974	0.0093	0.9972
30% train, 70% test	1	411	0.0438	0.9835
	2	346	0.0206	0.9942
	3	345	0.0162	0.9946
40% train, 60% test	1	800	0.0347	0.9879
	2	713	0.0144	0.9956
	3	696	0.0106	0.9966

Table 7: Summary of Fine-Tuning Results Across Different Train/Test Splits

[H]



(a) Fine-Tuning Accuracy and Loss on Japanese Data With Partial Misinformation

(b) Fine-Tuning Accuracy and Loss on Japanese Data With Full Misinformation

(c) Fine-Tuning Accuracy and Loss on Japanese and WELFake Data With Full Misinformation

Figure 9: Fine-Tuning Accuracy and Loss for Japanese Experimental Tests

Epoch	Time (seconds)	Loss	Accuracy
1	863	0.0341	0.9881
2	796	0.0136	0.9958
3	780	0.0110	0.9968

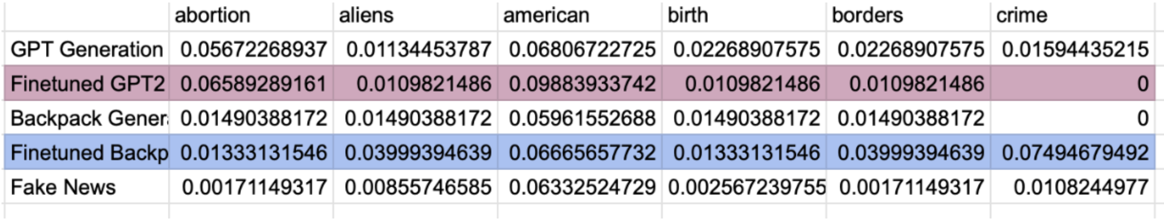
Table 8: Fine-Tuning Accuracy and Loss for Generative Data

Article	Reasoning	Classification
Max Blumenthal is known...President Trump.	This article is misinformative as it contains personal opinions and lacks factual information. It uses negative tone words like "bashing" and "hysterical."	Misinformative
A building used by Russia's...headquarters.	This article contains factual information about a fire at a building used by Russia's spy service. It reports facts without opinions or negative tone.	Informative
Indonesia's speaker...corruption charges.	This article provides factual information about the arrest of Setya Novanto, his plea, and does not contain opinions.	Informative

Table 9: ChatGPT Classification Using Few Shot Prompting

Prompt	Response
What are the semantic differences between the first group of four articles and the second group of four articles?	First Group: Focuses on military actions against ISIS, disturbing events involving political figures, and criticism of Donald Trump. Second Group: Covers diplomatic relations with China, military victory over ISIS in Raqqa, legal proceedings against Yingluck Shinawatra, and a defense deal with Nigeria.
What are differences in connotation between the first group of four articles and the second group of four articles?	First Group: Contains articles with negative connotations toward ISIS, Hillary Clinton's campaign, and Donald Trump's administration. Second Group: Focuses on diplomatic efforts, military success against ISIS, legal proceedings, and defense cooperation with Nigeria without overt negative connotations.
How would you categorize these two groups as either factual information or misinformation?	First Group: Mostly factual information with potentially biased or sensationalized language, leaning towards criticism of certain political figures. Second Group: Primarily factual information with less biased language, focusing on diplomatic, military, and legal developments.
What are the syntactical differences between the first group of four articles and the second group of four articles?	First Group: Contains shorter, more sensationalized sentences with emotive language. Second Group: Features longer, more descriptive sentences with a focus on factual reporting without sensationalism.

Table 10: Initial Questions Asked to GPT



	abortion	aliens	american	birth	borders	crime
GPT Generation	0.05672268937	0.01134453787	0.06806722725	0.02268907575	0.02268907575	0.01594435215
Finetuned GPT2	0.06589289161	0.0109821486	0.09883933742	0.0109821486	0.0109821486	0
Backpack Gener	0.01490388172	0.01490388172	0.05961552688	0.01490388172	0.01490388172	0
Finetuned Backp	0.01333131546	0.03999394639	0.06665657732	0.01333131546	0.03999394639	0.07494679492
Fake News	0.00171149317	0.00855746585	0.06332524729	0.002567239755	0.00171149317	0.0108244977

Figure 10: TF-IDF scores of Misinformed and Regular Models

Figure 11: The Misinformed Backpack Network

Training Set Size (%)	Test Set Size (%)	Accuracy (%)
5	95	99.40
10	90	99.89
60	40	99.86
70	30	99.79
80	20	99.84

Table 11: Testing Performance of Different Training/Testing Splits

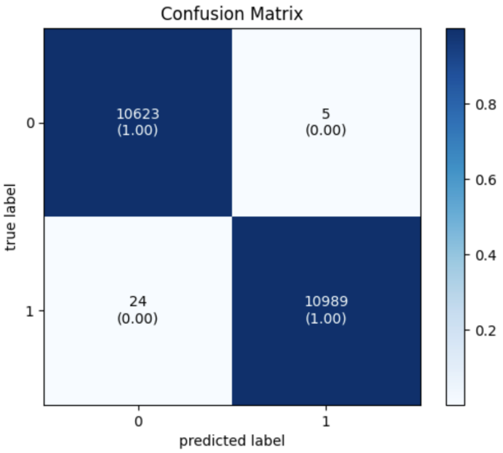


Figure 12: Confusion Matrix of RoBERTa Model (Results from WELFake Testing Data)

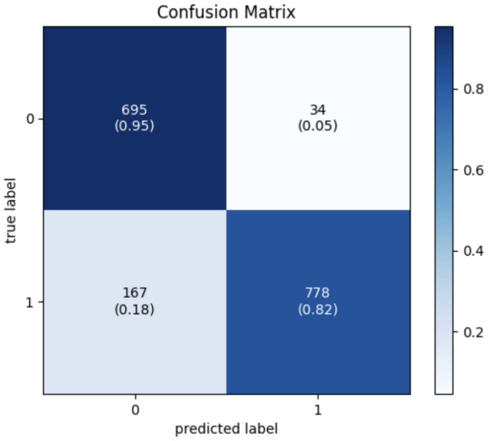


Figure 13: Confusion Matrix of Partial Misinformation in Japanese Data (Results from Testing Data)

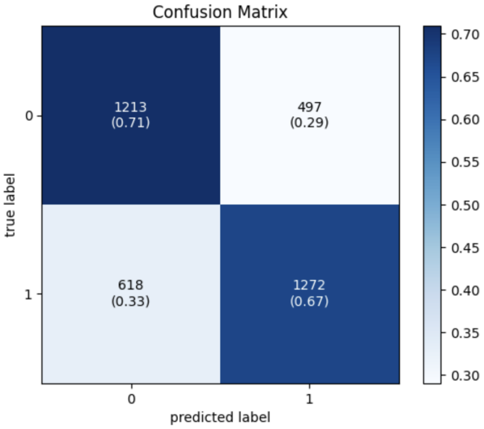


Figure 14: Confusion Matrix of Fine-Tuning with Generated Data (Results from Testing Data)