Detecting Impersonation in Taiwan's Political Landscape: An NLP Prompting Approach

Nicole Kuo
kuo.382@osu.edu

Sam Wang
wang.18721@osu.edu

Abstract

Political impersonation in Taiwan's online discourse poses significant challenges by inciting hatred and distorting facts. This research employs NLP prompt-based approaches with Gemini to detect impersonation in Chinese-language political discussions, particularly on Line messaging platforms. We compare three prompt strategies—default, detailed, and few-shot—with results indicating that few-shot prompting achieves the most balanced performance, though all strategies face challenges in precision. Our work contributes to understanding how to identify inauthentic political discourse that aims to polarize rather than engage constructively. By developing tools to distinguish genuine political expression from deceptive impersonation, we aim to support healthier online political conversations in Taiwan's increasingly polarized digital landscape.

1 Introduction

Political impersonation in Taiwan's online discourse presents a growing challenge to constructive democratic engagement. On platforms like Line, individuals infiltrate political communities pretending to support a political position while deliberately making extreme statements to undermine that stance—a practice known as trolling. Unlike genuine political disagreement, such impersonation polarizes communities, incites hatred, and distorts facts. Previous research on deceptive online content has primarily focused on fake news detection (Wang et al., 2022) and bot identification (Yang et al., 2019). Studies specific to political discourse have examined hate speech detection (Purevdagva et al., 2020) and community question answering (Shen et al., 2021), but these approaches struggle with the subtle linguistic nuances that distinguish genuine political expression from strategic impersonation. A growing concern in this space is the continuous evolution of tactics by malicious actors, who adapt their strategies to evade detection—such as using organized groups of accounts to act together, and mimicking rhetorical patterns of real political communities (Cresci, 2020). These behaviors blur the line between authentic and inauthentic discourse, making static detection methods increasingly fragile in dynamic online environments. To address these limitations, we propose a hybrid approach combining prompt-based techniques with feature analysis for detecting political impersonation in Chinese-language content. Our feature analysis component draws on established text classification methodologies (Apté et al., 1994)(Guo et al., 2017) to identify linguistic patterns unique to political impersonation, including contradictory phrases, exaggerated vocabulary, impossible claims, and emotional markers. This weighted feature system complements our prompt-based approach using the Gemini language model, creating a robust framework that can adapt to evolving impersonation tactics while maintaining high performance when API resources are limited. We evaluate three distinct prompt strategies—direct classification, detailed contextual prompts, and few-shot examples—to determine which most effectively captures the subtle indicators of inauthentic political discourse, while leveraging our feature analysis system as both a complementary classification method and a fallback mechanism.

2 Methodology

Our research for detecting political impersonation in Taiwan's online discourse follows a prompt-based approach using large language models. This section outlines our general approach to data collection, annotation, model selection, and evaluation.

2.1 Data Collection

We collected political discourse samples from two popular Taiwanese social media platforms, Line

and PTT, with particular emphasis on Line messaging groups, where political communities often form. Line serves as a primary channel for political communication in Taiwan, making it an ideal source for observing both genuine political discourse and instances of impersonation or trolling behavior. We focus on recent political events and discussions to ensure relevance to the current Taiwanese political landscape. Our collection process adheres to ethical guidelines by only gathering publicly available posts and anonymizing all personal identifiers to protect user privacy.

2.2 Annotation Process

The annotation process involves classifying each text sample into one of two categories: trolling/impersonation (labeled as 1) or genuine political discourse (labeled as 0). We acknowledge the inherently subjective nature of this task, as determining whether a message represents genuine political support or strategic impersonation often requires nuanced understanding of context, tone, and intent. As a two-person research team, we independently annotate each sample and compare our labels. When disagreements arise, which is common given the subjective nature of the task, we consult a third party to make the final determination. This approach helps mitigate individual biases and improves the reliability of our dataset.

2.3 Data Preprocessing

Our preprocessing pipeline focuses on cleaning Chinese political text while preserving semantic content essential for distinguishing genuine discourse from impersonation. We remove noise elements including URLs, news tags, and hashtags that contribute little to identifying impersonation. We also standardize punctuation marks and eliminate extraneous whitespace to reduce the influence of writing style variations that might inadvertently affect classification. These steps ensure our model focuses on meaningful content characteristics rather than superficial formatting or stylistic elements.

2.4 Gemini API Retry Mechanism

We experienced many ResourceExhausted errors and service interruptions during installation due to Gemini API rate restrictions and sporadic network instability. System reliability was impacted by these limitations, especially when batch processing was involved. We solved the problem by implementing a retry mechanism with exponential backoff, which permits feature degradation when required.

- Rate Limiting: The Gemini API imposes usage limitations to prevent overload. Our system implements exponential backoff with jitter, starting with a base delay of 5 seconds and doubling with each retry attempt, plus a small random delay component to prevent synchronized retries.

- Connection Resilience: To handle temporary network disruptions or service unavailability, our mechanism supports up to 7 retry attempts before falling back to feature analysis.

- Degradation: When API calls fail persistently, the system automatically switches to our feature analysis component, providing continuous classification capabilities even during API unavailability.

- Resource Optimization: The system includes built-in delays between requests (minimum 3 seconds) to respect API service constraints while maintaining throughput.

2.5 Model Selection and Prompt Engineering

For our analysis, we plan to leverage large language models capable of understanding nuanced Chinese text. While we are still in the process of finalizing our model selection, we are considering using Gemini for its capabilities in processing Chinese language content and contextual understanding. Our approach centers on prompt engineering, where we design various prompting strategies to elicit accurate classifications from the language model. We are exploring three main prompt types:

- Default: Direct classification prompts that simply ask the model to determine if the text represents trolling or genuine political discourse

- Detailed: Contextual prompts that provide additional background about political impersonation in Taiwan's online environment

- Few-shot: Few-shot prompts that include examples of both trolling and genuine political discourse to guide the model's understanding

- Explanatory: Enhanced prompts that request not only classification but also detailed explanations of reasoning and identification of key phrases that signal impersonation

These prompts are crafted in Chinese to maintain the linguistic nuances of the original content and enhance the model's understanding of the task.

Our implementation also integrates these prompt strategies with our feature analysis system, creating a robust hybrid framework. This combined approach allows us to leverage the strengths of both large language models and rule-based classification, while providing fallback mechanisms when API limits are encountered.

2.6 Feature Analysis

Our feature analysis system identifies four key linguistic patterns in Chinese political texts that signal impersonation. These linguistic patterns and keywords were empirically identified through our manual data collection and annotation process, where we observed recurring patterns in authentic political impersonation examples from Taiwanese online discourse:

- Contradictory phrase detection (3 points each): Identifies co-occurring terms that create logical inconsistencies. Examples include praising and criticizing simultaneously (like "真好"/"excellent" paired with "崩潰"/"collapse"). This carries the highest weight as these contradictions strongly signal ironic intent.

- Exaggerated vocabulary identification (2 points each): Recognizes hyperbolic terms commonly used mockingly in political contexts. Examples include "無所不能"/"omnipotent" or "救世主"/"savior" when applied to political figures. This carries moderate weight as these terms often indicate sarcastic exaggeration.

- Impossible claim detection (2 points each): Identifies unrealistic assertions that suggest irony through implausibility. Examples include "一天內解決"/"solved in one day" or "立刻解決"/"instantly resolved". This carries equal weight to exaggerated vocabulary as both represent implausible characterizations.

- Emotional intensity analysis (1 point per marker): Examines indicators of excessive emotion, particularly exclamation marks. Multiple exclamation marks often signal artificial enthusiasm in ironic text. This carries lower weight as this can appear in both genuine and ironic content.

Texts scoring 3 or higher are classified as impersonation, with confidence calculated through score normalization. This system complements our prompt-based approach and provides reliable classification when API services are unavailable.

2.7 Evaluation Framework

Our evaluation approach combines quantitative metrics with qualitative analysis to comprehensively assess performance. We use standard classification metrics including accuracy, precision, recall, and F1-score to evaluate the effectiveness of different prompt strategies (Sokolova and Lapalme, 2009). Beyond these metrics, we conduct error analysis to identify patterns in misclassifications and understand the limitations of our approach (Riberio et al., 2020). This mixed-methods evaluation helps us identify which types of political discourse are most challenging to classify and informs improvements to our prompt strategies. To ensure consistent comparison, we apply each prompt strategy to the entire dataset and analyze the resulting predictions. The systematic approach allows us to determine which prompting technique is most effective for distinguishing between genuine political discourse and impersonation in the Taiwanese context. In the next section, we present our current progress, including preliminary results from our initial experiments and detailed analysis of our findings.

3 Result Comparison

Building on our proposed methodology, we implemented a system for detecting political impersonation in Taiwanese online discourse and, in this section, compare the effectiveness of four prompt strategies.

3.1 Prompt Strategies

Figure 1 shows the performance comparison across our four prompt strategies: default, detailed, few-shot, and explain. The default and detailed prompts demonstrate similar patterns with exceptionally high recall (around 0.9 for default) but low precision (approximately 0.32), resulting in modest F1 scores.

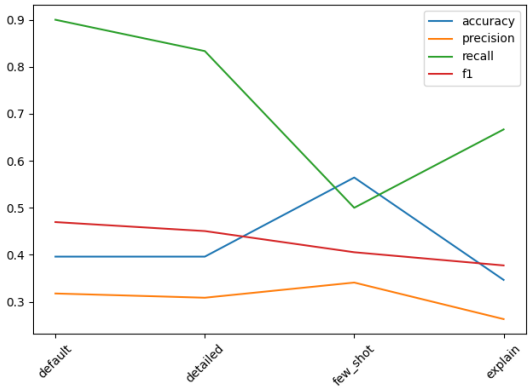

Figure 1: Performance Metrics Comparison Across Four Prompt Strategies

As we move toward more sophisticated prompt strategies, we observe a clear trade-off: the few-shot approach achieves better balance with improved accuracy (peaking at 0.55) and slightly higher precision, though at the cost of reduced recall. The explain prompt continues this trend with further declining recall but maintains consistent precision.

This pattern suggests that while simpler prompts are very effective at identifying impersonation cases (high recall), they generate numerous false positives. More structured prompts like few-shot provide better overall accuracy but miss more actual impersonation cases. This visualization highlights the inherent trade-offs in prompt engineering for political impersonation detection, with no single approach excelling across all metrics.

3.2 Feature Analysis with Prompt Templates

Figure 2 shows a significantly different performance pattern compared to the prompt-only approach. The hybrid system maintains consistently high accuracy (around 0.75) across all prompt strategies, while showing an interesting pattern in precision and recall. Notably, the few-shot prompt strategy demonstrates the highest precision (exceeding 0.80) but the lowest recall (approximately 0.17), creating a clear trade-off.

Unlike the previous graph where default and detailed prompts showed similar patterns, in the hybrid system there's substantial improvement in precision, especially with the few-shot approach. However, this comes at the expense of recall, which drops significantly. The explain prompt offers a more balanced approach, with slightly improved recall compared to few-shot while maintaining good

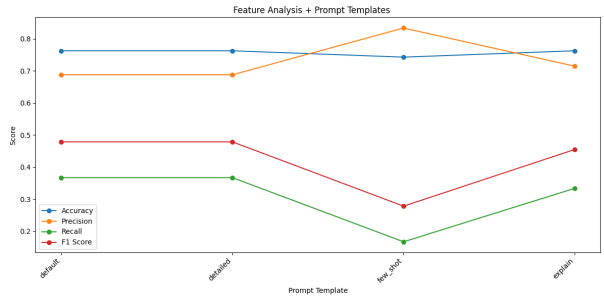

Figure 2: Performance Metrics of Hybrid System (Feature Analysis + Prompt Templates) Across Different Prompt Strategies

precision and overall accuracy.

4 Qualitative Error Analysis

To better understand our system's strengths and limitations, we conducted a qualitative analysis of prediction results, examining both successful classifications and error cases.

4.1 Case 1: Successful Classification (Genuine Expression)

- Chinese: 人生第一次投下的政黨票是民衆黨，第一次投下的總統票是柯文哲，自有投票權這麼多年來，頭一次讓我感受到政治原來也能"正常"
- English: For the first time in my life, I voted for the Taiwan People's Party, and for Ko Wen-Je as president. After having voting rights for so many years, this is the first time I felt politics could be 'normal'.

Correctly classified as genuine (0). The system detected no irony markers and recognized this as an authentic sharing of personal political experience.

4.2 Case 2: Successful Classification (Trolling Expression)

- Chinese: 台灣人的福氣啊！孔子至聖先師，在天運應付於賴總統，賴總統才高八斗，正詞正能量，國泰民安！
- English: What a blessing for Taiwanese people! Confucius, the great sage, has blessed President Lai from heaven. President Lai is exceptionally talented, speaks positively, and brings peace to the nation!

Correctly classified as trolling (1). The system identified exaggerated praise through terms like "才高八斗" (exceptionally talented) and divine

references that signal mockery rather than sincere support.

4.3 Case 3: False Positive Error

- Chinese: 怎麼沒人要罷免那缺德？在野黨主席在睡覺嗎？任憑綠畜們欺壓百姓...
- English: Why isn't anyone recalling that dishonest person? Are opposition party leaders sleeping? Allowing the green beasts to oppress the people...

Incorrectly classified as ironic (1) when actually genuine (0). The system misinterpreted strong emotional criticism and derogatory terms as mockery, highlighting a challenge in distinguishing genuine political anger from satirical impersonation.

4.4 Case 4: False Negative Error

- Chinese: 總統真的厲害，博學多聞，學富五車，才高八斗，人中龍鳳，小民只有個小願望，以總統的能力幫我們解決小小的高房價和少子化的問題就不勝感激了
- English: The president is truly remarkable, knowledgeable, extremely talented, and exceptional. As a common citizen, I only have a small wish - with the president's abilities, solving the minor problems of high housing prices and low birth rates would be greatly appreciated.

Incorrectly classified as genuine (0) when actually ironic (1). The system failed to detect the subtle sarcasm where serious systemic issues are ironically characterized as "minor problems," demonstrating limitations in recognizing understated irony.

These cases illustrate key challenges in our classification system: distinguishing emotional criticism from mockery, and detecting subtle forms of irony that rely on understatement rather than obvious exaggeration.

5 Conclusion

Our research demonstrates that detecting political impersonation in Taiwan's online discourse requires a multifaceted approach. Through our hybrid system combining feature analysis with prompt-based techniques using the Gemini 2.0 Flash model, we have made significant progress in addressing this challenging classification task.

The comparative analysis of different prompt strategies revealed notable trade-offs between precision and recall. While simple prompts achieved high recall but struggled with precision, more structured approaches like few-shot prompting provided better overall accuracy at the cost of reduced recall. Our feature analysis system proved particularly valuable in maintaining consistent accuracy across all prompt strategies while providing a reliable fallback mechanism during API limitations.

The qualitative error analysis highlighted persistent challenges in distinguishing between genuine emotional criticism and satirical impersonation, as well as detecting subtle forms of irony that rely on understatement rather than obvious exaggeration. These findings suggest that political impersonation detection is inherently subjective and context-dependent, requiring systems that can adapt to the nuanced linguistic features of Taiwan's political discourse.

By developing a hybrid approach that leverages both large language models and rule-based classification, we have created a more robust framework for identifying inauthentic political expressions. This work contributes to broader efforts to maintain healthy online political conversations in Taiwan's increasingly polarized digital landscape, where distinguishing between genuine political expression and strategic impersonation is crucial for constructive democratic engagement.

6 Contributions

- Nicole: Data collection, code development, report writing
- Sam: Data collection, code review, presentation, report review

7 Acknowledgements

Claude assisted with text formatting, editing during report preparation, and helped refine our feature analysis implementation by suggesting optimizations for the keyword detection and confidence scoring algorithms.

ChatGPT was used during the development phase to debug code errors in our API retry mechanism and to help generate preliminary prompt templates that we later refined.

References

Chidanand Apté, Fred Damerau, and Sholom M. Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251.

Stefano Cresci. 2020. A decade of social bot detection. *CoRR*, abs/2007.03604.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.

Chinguun Purevdagva, Rui Zhao, Pei-Chi Huang, and William Mahoney. 2020. A machine-learning based framework for detection of fake political speech. In *2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE)*, pages 80–87.

M.T. Riberio, T. Wu, C. Guestrin, and S. Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. Association for Computational Linguistics.

Li Shen, Jun Shen, Sheng Lu, Hao Ji, Gang Chen, Jian-Guo Ni, Yang Fan, and Xiang Ren. 2021. Knowledge-enhanced hierarchical attention for community question answering with multi-task and adaptive learning. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7763–7767. IEEE.

M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Haizhou Wang, Sen Wang, and YuHu Han. 2022. Detecting fake news on chinese social media based on hybrid feature fusion method. *Expert Systems with Applications*, 208:118111.

Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.