Towards Detecting Persuasion on Social Media: From Model Development to Insights on Persuasion Strategies

Elyas Meguellati[1], Stefano Civelli[1], Pietro Bernardelle[1], Shazia Sadiq[1], Irwin King[2], Gianluca Demartini[1]

[1]University of Queensland, Brisbane, Australia
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
{m.meguellati, s.civelli, p.bernardelle, s.sadiq, g.demartini}@uq.edu.au king@cse.cuhk.edu.hk

Abstract

Political advertising plays a pivotal role in shaping public opinion and influencing electoral outcomes, often through subtle persuasive techniques embedded in broader propaganda strategies. Detecting these persuasive elements is crucial for enhancing voter awareness and ensuring transparency in democratic processes. This paper presents an integrated approach that bridges model development and real-world application through two interconnected studies. First, we introduce a lightweight model for persuasive text detection that achieves state-of-the-art performance in Subtask 3 of SemEval 2023 Task 3 while requiring significantly fewer computational resources and training data than existing methods. Second, we demonstrate the model's practical utility by collecting the Australian Federal Election 2022 Facebook Ads (APA22) dataset, partially annotating a subset for persuasion, and fine-tuning the model to adapt from mainstream news to social media content. We then apply the fine-tuned model to label the remainder of the APA22 dataset, revealing distinct patterns in how political campaigns leverage persuasion through different funding strategies, word choices, demographic targeting, and temporal shifts in persuasion intensity as election day approaches. Our findings not only underscore the necessity of domain-specific modeling for analyzing persuasion on social media but also show how uncovering these strategies can enhance transparency, inform voters, and promote accountability in digital campaigns.

Introduction

The proliferation of digital platforms has significantly amplified the role of propaganda in shaping public opinion and influencing societal outcomes. Unlike traditional media, social media platforms facilitate rapid dissemination and subtle integration of persuasive content into everyday communications, making manual identification increasingly challenging (Glowacki et al. 2018; Tardaguila, Benevenuto, and Ortellado 2018). Consequently, there is a pressing need for automated methods to detect persuasive content effectively, particularly given its potential for misuse, as highlighted by incidents such as the Cambridge Analytica scandal (Boerboom 2020).

Persuasion, a key aspect of propaganda, involves strategies designed to influence beliefs, attitudes, or behaviors through emotion, reasoning, or rhetorical devices, while ostensibly respecting individual autonomy (Perloff 2017). Figure 1 illustrates examples of such persuasion techniques, highlighting their complexity and subtlety[1]

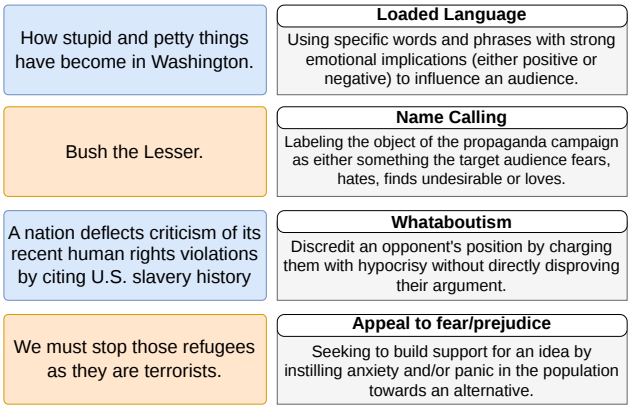


Figure 1: Examples of persuasion techniques from the SemEval dataset with their respective explanations.

Previous research, notably represented by the SemEval-2023 Task 3, has substantially advanced automated persuasion detection by leveraging large-scale annotated datasets derived mainly from mainstream news articles. However, applying these models directly to social media contexts presents significant challenges due to the informal, brief, and noisy nature of social media language, as well as the distinct persuasive strategies employed by political campaigns on these platforms.

Motivated by these challenges, we introduce a computationally efficient model that operates effectively under resource constraints. After demonstrating state-of-the-art performance on the SemEval persuasion benchmark, we examine the model's adaptability to social media through an empirical evaluation on a newly introduced dataset—the Australian Political Ads 2022 (APA22)—comprising political advertisements from Facebook during the 2022 Federal Election. This domain adaptation step is essential, as the ini-

[1]https://propaganda.math.unipd.it/semeval2024task4/definitions22.html.

tial cross-domain application from mainstream news articles to social media would yield suboptimal performance, highlighting the necessity of tailored, domain-specific training.

After fine-tuning our model on the APA22 training set, we achieve a substantial improvement in performance, validating the importance of domain adaptation. Leveraging this adapted model, we reveal how differences in persuasive intensity correlate with distinctive patterns in demographic targeting, strategic funding allocation, temporal messaging shifts, and specific lexical choices.

Specifically, our integrated analysis addresses the following research questions:

1. In scenarios requiring live inference and limited in-domain annotations, to what extent can persuasion detection models be optimized to achieve state-of-the-art performance while significantly reducing computational and data resource requirements
2. How well does a model trained on SemEval's predominantly mainstream news articles transfer to the informal, noisy, and domain-specific language of social-media political ads?
3. When grouping ads by their level of persuasion intensity, what distinct advertising strategies emerge?

By addressing these questions, we contribute to the ongoing efforts in developing efficient computational methods for persuasion detection and provide an initial exploration into the strategies behind digital political advertising, ultimately enhancing transparency in democratic processes.

Paper Overview

The paper is structured as follows:
Study 1: Model Development and Evaluation. We present our lightweight transformer-based persuasion detection model, detail its training on the SemEval 2023 dataset, and provide extensive configuration details, ablation studies, and comparisons with existing state-of-the-art models.
Study 2: Domain Adaptation and Persuasion Analysis. We first evaluate how well the SemEval-trained model transfers to our small, Facebook-sourced APA22 dataset, then fine-tune it for optimal performance, and finally use it to analyze persuasive strategies in social-media political.

Related Work

We organize the related work into two main areas: persuasion in natural language processing (NLP), followed by studies addressing persuasion within social media contexts.

Persuasion in NLP

In the domain of text and social media analysis, the distinction between persuasion and propaganda detection has become increasingly well-defined. Historically, the two were often conflated due to their shared reliance on rhetorical features such as *loaded language*, *name-calling*, and *emotional appeals*. Early foundational work by Rashkin et al. (2017) approached this space using document-level labels to classify texts as trusted, satire, hoax, or propaganda. Later, Barrón-Cedeño et al. (2019) introduced a refined corpus and modeling approach that focused on linguistic style and readability features to predict propagandistic content.

The research focus gradually shifted toward more granular and interpretable classification frameworks. For instance, Habernal and Gurevych (2017) explored persuasion through argument fallacy detection, annotating five types of logical fallacies in argumentation. This direction was extended by Da San Martino et al. (2019), who released a span-level corpus annotated with 18 fine-grained propaganda techniques and benchmarked transformer-based models for both classification and sequence labeling. Follow-up work addressed transformer limitations in capturing context (Chernyavskiy, Ilvovsky, and Nakov 2021), interpretability in propaganda detection (Yu et al. 2021), and domain-specific applications such as memes (Dimitrov et al. 2021), code-switched discourse (Salman et al. 2023), and health-related propaganda during COVID-19 (Nakov et al. 2021).

Further, Hristakieva et al. (2022) examined propaganda within coordinated social media campaigns. Subbiah and McKeown (2021) investigated how identity signaling in social media comments enhances persuasiveness, while pro (2016) focused on detecting paid and coordinated promotional content across platforms—tasks closely related to understanding persuasive intent. More recently, llm (2024) evaluated the capabilities of LLMs in generating persuasive arguments within synthetic dialogues, demonstrating their potential to influence opinion change in both artificial agents and human evaluators.

SemEval-2023 Task 3 Subtask 3, introduced by Piskorski et al. (2023b), marked a pivotal shift by treating persuasion as a distinct subtask within the broader propaganda detection challenge. This subtask adopted a multi-label formulation using 23 persuasive techniques, annotated at the span level over multilingual news articles. A linear SVM baseline was introduced using *uni-grams* and *bi-grams* (Piskorski et al. 2023a), and subsequent models such as APatt (Purificato and Navigli 2023) and KInIT (Hromadka et al. 2023) pushed the state-of-the-art using ensembles of fine-tuned transformers and multilingual models with tuned thresholds.

While recent work highlights the strength of transformers on SemEval-style tasks, zero-shot and few-shot LLMs continue to underperform in high-context, multi-label classification (Jose and Greenstadt 2024; Edwards and Camacho-Collados 2024). However, a growing body of research has explored incorporating LLMs not for direct classification but as part of data cleaning and semantic augmentation strategies. One such study demonstrates that prompting LLMs to generate explanations or paraphrase noisy data improves downstream model performance across multiple tasks, including propaganda and toxicity detection (Meguellati et al. 2025). These hybrid methods suggest a promising direction for leveraging LLM capabilities without relying entirely on their zero-shot inference.

Persuasion in Social Media

In addition to the broader study of persuasion and propaganda detection, there is a growing body of research focused specifically on political advertising on social media platforms. Political microtargeting—tailoring ads based on

Language	Train			Dev			Test			Total	
	#DOC	#CHAR	AVG pt	#DOC	#CHAR	AVG pt	#DOC	#CHAR	AVG pt	#DOC	#WORD
EN	446	2.43M	16.1	90	403K	20.0	54	228K	32.9	590	469K
FR	158	737K	**35.4**	53	222K	29.9	50	181K	33.6	261	153K
DE	132	581K	34.1	45	171K	27.5	50	259K	38.1	227	104K
IT	227	927K	26.6	76	287K	25.4	61	245K	38.5	364	186K
PL	145	765K	19.6	49	264K	20.1	47	349K	31.7	241	144K
RU	143	590K	23.8	48	163K	15.4	72	161K	13.1	263	104K
GE	-	-	-	-	-	-	29	46K	7.5	29	-
GR	-	-	-	-	-	-	64	248K	10.8	64	-
SP	-	-	-	-	-	-	30	109K	18.2	30	-

Table 1. Overview of SemEval 2023 dataset across various languages, showcasing document counts, character totals, and the density of persuasion techniques. AVG pt represents the average number of persuasion technique annotations (instances) per document, not unique technique types. For example, the French training set averages 35.4 persuasion technique instances per document, though only 23 distinct technique types exist in the taxonomy. This metric indicates annotation density—how frequently persuasion techniques occur within documents. Georgian (GE), Greek (GR), and Spanish (SP), introduced as surprise languages, are represented solely within the Test set. The 'Total' column summarizes the total number of documents (#DOC) and words (#WORD) for each language across all sets.

personality profiles and user data—has been shown to influence voter attitudes and engagement (Zarouali et al. 2020; Appel, Matz, and Kosinski 2024). These studies demonstrate that highly personalized and strategically targeted ads can be more persuasive, raising concerns about voter manipulation in digital political campaigns. However, experimental evaluations have also found mixed evidence regarding the overall effectiveness of such ads, suggesting that measured persuasive impacts may be limited or context-dependent (Barrett and McGregor 2024). The role of ad delivery algorithms further complicates this landscape, as these systems act as opaque gatekeepers shaping the reach and visibility of political messaging (Ali et al. 2019). Independent auditing initiatives like Facebook Ads Monitor have aimed to improve transparency by systematically tracking political ads on platforms like Facebook, which remains a dominant venue for political advertisers due to its extensive reach and comprehensive ad transparency tools (Silva et al. 2020). These findings underscore the importance of developing computational methods tailored to political social media ads, motivating our focus on domain-specific persuasion detection and analysis.

Despite these advances, most persuasion detection models are still trained and evaluated on mainstream media datasets. As such, their generalizability to informal, social media-specific content (e.g., political ads) remains underexplored. This gap motivates our study, which investigates domain adaptation by applying and fine-tuning a state-of-the-art persuasion model on a real-world dataset of social media political advertisements. Our approach highlights how domain-sensitive modeling can reveal distinct patterns in persuasive tactics used in online political campaigns.

Study 1: Persuasion as a Multi-Label Classification for SemEval-23

Dataset Description

The SemEval-23 Task 3 dataset is a comprehensive multilingual corpus collected from various globally discussed topics, including the COVID-19 pandemic, abortion-related legislation, migration, and the Russo-Ukrainian conflict, among others. The dataset includes articles in nine languages: English, French, German, Georgian, Greek, Italian, Polish, Russian, and Spanish published between 2020 and mid-2022. The articles were sourced from both mainstream and alternative media to capture a wide range of perspectives. Each article was annotated for persuasive techniques at the span level, utilizing a detailed taxonomy designed for this task. Approximately 40 annotators, proficient in their respective languages, ensured a high-quality annotation of the dataset, as shown in Table 1.

Methodology

The SemEval-23 Task 3 Subtask 3 presents a multi-label classification task with imbalanced labels across 23 persuasion techniques Table 7 in the Appendix provides the full label set and the distribution, withholding the test labels and requiring participants to submit the predictions to an online platform for evaluation [2]. For this task, we introduce our novel PPAsy (Persuasion Prediction Asymmetric) method, which utilizes an asymmetric binary cross-entropy loss function to effectively manage label imbalance. The official metric set by the organizers to assess performance is the micro-averaged F1 score (*F1-micro*), complemented by the F1-macro score.

Data Preparation. We translated all non-English training documents into English using the Google API[3] machine translation model to standardize the data for uniform model training. We trained our model on individual datasets, each translated from the source language to English. Additionally, we trained the model on a combined dataset, which included the translated training data from all languages along with the original English dataset. The French language dataset, com-

[2]https://propaganda.math.unipd.it/semeval2023task3/leaderboard.php

[3]https://cloud.google.com/translate

prising of 211 documents out of a total of 1,612, exhibits a notably high average number of persuasion techniques per document. It contains all 23 labels, in contrast to the English dataset, which includes only 19. Consequently, the French-to-English was the only dataset employed to train our model which was subsequently utilized to generate predictions for the English test set.

Preprocessing. We employed minimal preprocessing for all translated datasets, consisting only of lemmatization and lowercasing using SpaCy[4]. We specifically considered the role of punctuation, informed by its demonstrated stylistic significance in textual analysis (Darmon et al. 2021). Noting that sentences including punctuation yielded better results, as shown in Table 3, we opted for keeping it. This outcome suggests a potential influence of punctuation on the model's ability to interpret text.

Model and Dataset Selection. In our exploration of various machine learning models, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based models, we ultimately selected **XLNet-base** (Yang et al. 2019) due to its optimal balance of computational efficiency and predictive performance. Specifically, XLNet-base provided robust performance with significantly fewer parameters (110M) compared to larger transformer ensembles, such as APatt (492M parameters) and KInIT (355M parameters) (Purificato and Navigli 2023; Hromadka et al. 2023), as summarized in Table 2. Detailed comparisons and justifications for choosing XLNet over other models are further elaborated in Appendix A.

Our experiments with multiple language subsets from the SemEval dataset revealed a positive correlation between model performance and the average number of persuasive techniques per document (see Table 1). Among the subsets, French-translated-to-English exhibited the highest average number of persuasive techniques (29.9 per document) and yielded superior model performance (F1-micro = 40.62). Recognizing potential biases introduced by automated translations, we manually inspected a subset of translated documents to verify their accuracy and consistency. This validation step ensured that the translated dataset maintained semantic integrity, thereby minimizing potential biases and errors. Based on these observations, we selected the French subset for training our model, as it provided the most advantageous balance between dataset richness and size.

Translation Bias and Validation. A bilingual annotator sampled 100 French–English sentence pairs to label each as either "meaning preserved" or "meaning changed." Only 2 pairs were marked as "meaning changed," indicating minimal semantic drift. This simple validation suggests that the translation is unlikely to introduce systematic bias into model training or evaluation.

Loss Function. To address the inherent challenges of multi-label classification tasks characterized by substantial class imbalance, we propose an adapted version of the binary cross-entropy (BCE) loss function. Our implementation involves an instance-specific asymmetric weighting approach:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot BCE(y_i, \hat{y}_i),$$
$$\text{where } w_i = \beta \cdot y_i + (1 - \beta) \cdot (1 - y_i).$$

Here, y denotes the true labels, $\hat{y}$ the predicted probabilities, and w_i the instance-specific weights. The parameter β explicitly controls the emphasis placed on correctly classifying minority classes, countering the imbalance in label distributions. We empirically determined $\beta = 0.7$ as optimal through trial and error.

While related techniques such as focal loss (Lin et al. 2017) and *class-balanced loss* (Cui et al. 2019) asymmetrically adjust class weights primarily based on sample frequency, our proposed loss function explicitly integrates an external, instance-specific weighting mechanism controlled by a tunable parameter (β). This design differs significantly from common practice, as it not only addresses class imbalance at a class-level but also allows finer-grained control at the instance level. As demonstrated by our ablation study (Table 3, Iteration 3), this novel adaptation substantially contributes to improving model performance in highly imbalanced, multi-label classification contexts. Our implementation is publicly available at [redacted-for-blind-review].

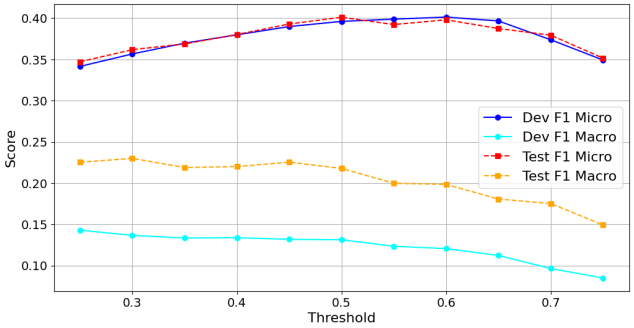


Figure 2: Model Performance across Different Thresholds.

Prediction Threshold Calibration. When converting model output probabilities into binary class predictions, selecting an optimal threshold is crucial. However, as shown in Table 1, there are significant discrepancies in average persuasion (AVG pt) between the English development (Dev) and test sets. This difference suggests that improvements observed in the Dev set may not consistently translate to comparable improvements in the Test set, and vice versa. To strike a balance between optimizing F1-micro and F1-macro performance, we maintained a default threshold of 0.5, as illustrated in Figure 2, despite the Dev set having an optimal threshold of 0.6.

Efficiency in Resource-Constrained Settings (RQ1)

Our study investigates the potential of optimizing persuasion detection models to achieve state-of-the-art per-

[4]https://spacy.io/

Method	Dataset Size	Configuration	Parameters	F1-Micro (%)	F1-Macro (%)
PPAsy-Conv (Ours)	153K	CNN	N/A	35.01	8.89
APatt (Purificato and Navigli 2023)	1.16M	Transformers Ensemble	492M	37.56	12.92
KInIT (Hromadka et al. 2023)	1.16M	RoBERTa-large	355M	40.38	15.86
PPAsy-XLNet (Ours)	153K	XLNet-base	110M	40.62	21.78
Baseline (Piskorski et al. 2023a)	469K	SVM	N/A	19.51	6.93

Table 2. Performance metrics of models on the English test set. PPAsy-XLNet (ours) outperforms APatt and KInIT baselines while using fewer parameters and less training data. Baseline SVM performance from SemEval-23 (Piskorski et al. 2023a).

Iteration	Removed	F1-micro (%)	ΔSOTA
SOTA	Full model	40.62	0.00
1	Translation	28.35	-12.27
2	Punctuation	39.71	-0.91
3	Loss function	33.94	-6.68
4	Epoch	38.49	-2.13

Table 3. Ablation study results showing the impact of removing components from our PPAsy-XLNet model. Translation and our loss function optimization have the largest impact on performance.

formance under computational and data resource constraints—conditions typical of real-world deployment scenarios with live inference requirements and limited in-domain annotations.

We demonstrate that computational efficiency can be attained through both model architecture choices and a customized loss function. As shown in Table 4 and Table 2, the PPAsy-XLNet model, with only 110M parameters, outperforms larger models such as APatt's ensemble (492M parameters) and KInIT's RoBERTa-large (355M parameters). PPAsy-XLNet achieves higher F1-Micro (40.62%) and F1-Macro (21.78%) scores, while reducing computation time to 136 seconds—substantially faster than KInIT's 988 seconds. Our class-weighted loss function further enhances this efficiency by addressing class imbalance without incurring additional computational overhead. The ablation study results confirm that this combination of architectural and loss function optimization effectively balances performance with runtime demands. In parallel, the model also demonstrates high data efficiency. Unlike previous systems that rely on augmented datasets spanning multiple languages and over 1.16M tokens, PPAsy-XLNet achieves superior performance using only 153K tokens from a French-to-English translated subset. This subset covers all 23 persuasion labels and contains a higher average of techniques per document, offering rich signal density. These results suggest that a strategically selected, feature-rich subset can replace larger multilingual corpora while maintaining, or even improving, model performance.

Together, these findings affirm that model and training optimization strategies can yield persuasive detection systems that are both computationally and data efficient, making them well-suited for deployment in low-resource or real-time applications.

Study 2: Persuasion in the 2022 Australian Federal Election Facebook Ads

This second study extends the exploration of the PPAsy-XLNet model introduced in Study 1 to persuasion detection in political advertisements, using the case study of the Australian federal election campaign, for which we collected and partially annotated a new dataset, called APA22. First, to investigate the domain adaptability of PPAsy-XLNet, we employed the model trained on the SemEval-23 dataset to identify persuasion on the APA22 dataset. Subsequently, we cross-validated the model directly on the APA22 dataset in order to determine the performance difference. Study 2 not only tests our model's applicability to real-world data, but also offers preliminary insights into the dynamics of digital persuasion in political campaigns on social media.

APA22 Dataset

Data Collection. Using the Meta Ad Library API[5] we collected the online advertisements related to the Australian federal election held on May 21, 2022. Actively running advertisement campaigns were captured with a six-hour frequency between March 1, 2022, and June 18, 2022. Each ad includes attributes such as the creation time, target demographic distributions, ad text, funding entity, ad URL, and metrics of ad impression and spending. The resulting dataset consists of 56,958 unique ads, each containing an average of 3.86 sentences. The APA22 dataset offers a substantial resource for investigating the dynamics and the strategies of political advertising on social media [6].

Data Labeling. A subset of 658 samples from the APA22 dataset was randomly selected and manually annotated for persuasive content. The initial annotation was conducted by a faculty member from the Political Science department who focused on the five most common persuasion techniques (which our analysis showed account for approximately 70% of techniques in the SemEval dataset, see Table 7 in Appendix).

To assess annotation quality, three independent researchers from Computer Science evaluated an overlapping subset of the labeled dataset by comparing their annotations with the Political Science faculty member's labels. We

[5]https://www.facebook.com/ads/library/

[6]The dataset collection process has undergone ethics review at the authors' organization. This dataset is available online at [redacted-for-blind-review].

Model	Training Time (e)		Validation Time		F1 Micro (%)		F1 Macro (%)	
	French	All	French	All	French	All	French	All
XLNet-Base	136 (27)	492 (153)	10	34	**40.62**	39.71	**21.78**	**23.16**
XLNet-Large	305 (76)	623 (311)	29	99	36.78	36.75	7.92	6.64
RoBERTa-Base	128 (25)	428 (128)	11	32	40.38	**41.12**	11.69	21.59
RoBERTa-Large	277 (69)	988 (277)	27	90	34.18	34.17	4.01	4.13
DistilBERT	**100 (20)**	**399 (100)**	**8**	**27**	40.60	36.50	20.62	19.14
XLM-RoBERTa-Base	111 (22)	364 (111)	9	30	31.08	32.26	5.73	6.21
XLM-RoBERTa-Large	281 (56)	432 (281)	24	80	34.18	24.68	4.01	2.06

Table 4. Model performance comparison across configurations. PPAsy configurations (XLNet-Base/Large, ours) compared with external baselines (RoBERTa, DistilBERT, XLM-RoBERTa). Times in seconds; training epoch time shown in parentheses. All validations/tests performed on English dataset.

calculated inter-annotator agreement using Fleiss' Kappa, achieving a rate of 0.86, indicating substantial agreement. In case of a disagreement we retained the original labels from the Political Science expert for our experiments.

Since the APA22 dataset was created and annotated prior to the release of SemEval-23, there were differences in the annotation schemes. To align the two datasets for comparative analysis, we implemented a cross-dataset compatibility approach by converting both datasets from multi-label to binary classification. Specifically, we categorized each sentence as 'persuasive' if it contained at least one persuasion technique and 'neutral' otherwise. This binarization approach was particularly appropriate given that the remaining 18 techniques in SemEval had minimal representation (1-3% occurrence each).

This binarization process yielded 368 'persuasive' and 290 'neutral' sentences within the annotated subset of the APA22 dataset. The data was then split into training (75%, 493 sentences) and testing (25%, 165 sentences) using stratified sampling. A fixed seed was used to ensure consistency and reproducibility in the test set.

Model	SemEval-23		APA22	
	Acc	F1	Acc	F1
SVM	0.56	0.53	0.76	0.75
PPAsy-Conv	0.55	0.55	0.79	0.79
PPAsy-DistilBERT	0.58	0.54	0.80	0.79
PPAsy-RoBERTa	0.56	0.52	0.79	0.78
PPAsy-XLNet	**0.59**	**0.55**	**0.82**	**0.82**

Table 5. Performance comparison of models evaluated on the APA22 test set. Left columns show performance when models were trained on SemEval-23 data (tested on APA22), right columns show performance when models were trained and tested on APA22 data.

How Well Can Persuasion Detection Models Generalize Across Domains? (RQ2)

To investigate model generalizability across different domains, we evaluated how effectively models trained on the SemEval-23 dataset—primarily composed of mainstream media news articles—could adapt to political social media advertisements in the APA22 dataset. This cross-domain evaluation is crucial for understanding the practical applicability of persuasion detection models on social media.

Our experimental results demonstrate significant challenges in model generalizability when transferring from news to social media domains. As shown in Table 5, the PPAsy-XLNet model, which achieved state-of-the-art performance on SemEval-23, exhibited markedly degraded performance when applied to the APA22 test set (accuracy: 59.39% vs. 82.42% when trained on APA22). This substantial 23 percentage point drop reveals the domain-specific nature of persuasion detection models. Several factors contribute to this limited generalizability:

- **Linguistic differences.** Political social media ads employ informal, emotionally charged, and highly variable language, unlike the more structured and formal tone of mainstream news articles. Prior work has shown that vocabulary, syntax, and sentiment expressions differ significantly across domains, and such variation can substantially degrade model transfer performance (Blitzer, Dredze, and Pereira 2007; Rasooli and Sproat 2018).
- **Content structure.** Social media ads are typically short and direct, often consisting of a single sentence or phrase designed to prompt immediate engagement. In contrast, news articles present extended narratives with contextual buildup. This mismatch in content length and structure affects how models interpret and weigh textual cues (Gao, McKeown, and Shivade 2011; Li, Goldwasser, and Lee 2019).

When the same models were trained directly on the APA22 dataset, all architectures showed substantial improvements, with PPAsy-XLNet reaching 82.42% accuracy. This demonstrates that while our models have the architectural capacity for high performance, their generalizability is constrained by the domain specificity of the training data.

These findings have important implications for deploying persuasion detection systems: models trained on one text domain (e.g., news articles) may not generalize effectively to other domains (e.g., social media). For practical applications like real-time detection of persuasive political advertising on social media platforms, domain-specific training or adaptation is essential to achieve reliable performance.

Persuasion in Elections (RQ3)

The aim of this analysis is to demonstrate how persuasive text in social media political ads can be feasibly detected and interpreted, rather than to provide an extensive political campaign analysis.

The following investigation seeks to unravel subtleties of the language adopted in high and low persuasion ads. PPAsy-XLNet, trained on the manually annotated portion of the APA22 dataset, was used to label the rest of the advertisements in the dataset as either 'neutral' or 'persuasive'.

Based on the percentage of persuasive sentences contained in each ad, we define the concepts of high and low persuasive ads. In our dataset, 26,419 advertisements (46.4% of the total) were found to be highly persuasive (i.e., more than 80% of the sentences in the advertisement considered persuasive). On the other hand, 7,216 ads (12.6% of the total) were categorized as having low levels of persuasion, meaning that 20% or fewer sentences were classified as persuasive.

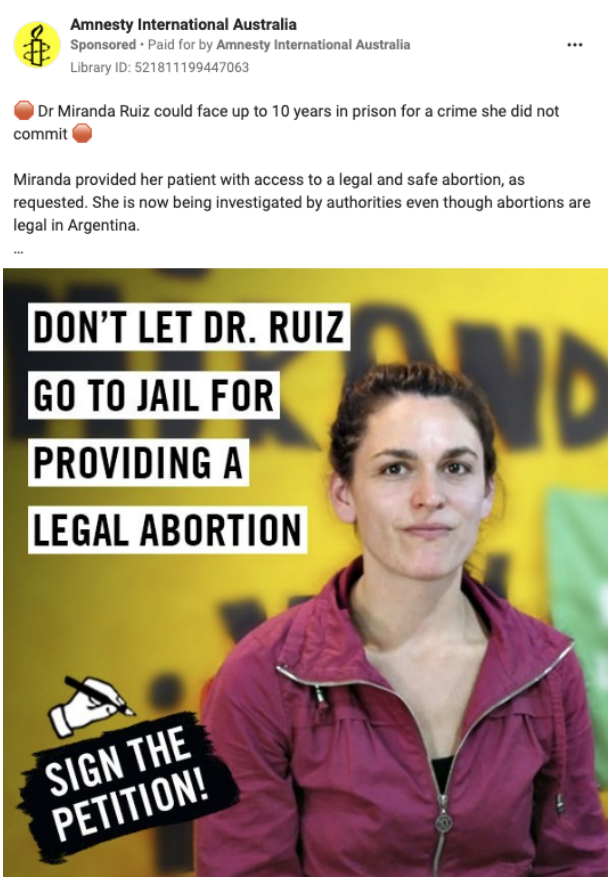
Persuasive Content Dynamics. Table 6 offers insights into the persuasion found in the APA22 dataset. We observed that low persuasion ads tend to obtain fewer impressions, averaging 17,441, and incur lower spending, with an average of $265.6. In contrast, ads categorized as highly persuasive achieve higher averages, with 25,407 impressions (45.6% more) and $393.6 in spending (48.2% more), though both target the *same* age and gender demographics. These findings suggest a potential correlation between the persuasive intensity of ads and their campaign reach and spend, with highly persuasive ad campaigns lasting on average 11 days compared to 8 days for low persuasion ones. To gain a better understanding of what exactly makes an ad persuasive, we next compare patterns in the language used in high and low persuasion ads.

Bi-grams Analysis. We analyze bi-gram frequencies to discern notable differences in messaging strategies. Bi-grams commonly found in highly persuasive ads, such as 'climate change' and 'better future' (see Table 9 in Appendix) seem to reflect an emphasis on policy and future-oriented values. This might suggest an attempt to appeal to the audience's aspirations and ongoing concerns. An illustrative example of this persuasive approach is the sentence, *"Don't risk more Morrison. Only a vote for Labor will kick out this government and deliver a better future."* This ad employs emotive language implying negative consequences ("Don't risk") and incorporates a strong call to action ("vote for Labor"). References to political figures and government entities, like frequent bi-grams "government labor", "scott morrison", and related variations, indicate targeted political messaging. Notably, Scott Morrison was the leader of the **Liberal** Party and the incumbent Prime Minister at the time of the 2022 Australian Federal Election. The **Labor** Party's use of his name in highly persuasive ads suggests an *"attack on reputation"* persuasive technique, aiming to influence voters by highlighting perceived shortcomings of their political opponent. It is noteworthy that although this type of bi-grams is also present in text classified as low in persuasion, their prevalence is significantly reduced.

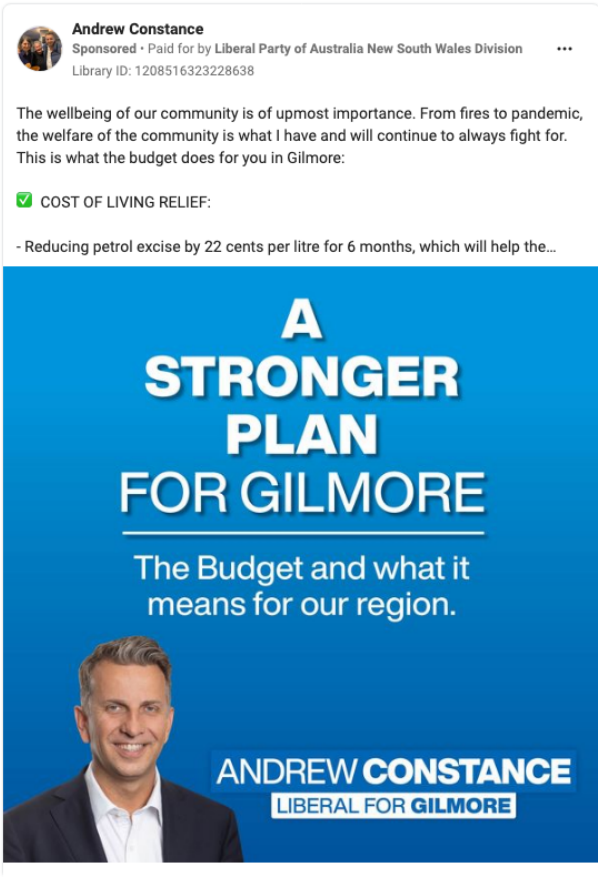
Conversely, low persuasion ads feature more informational and less emotionally charged bi-grams. These bi-grams are often descriptive, focusing on geographical, cultural, or administrative entities and typically lack a compelling call to action or emotional appeal. The language in these ads tends to focus on identity, locality, or administrative details, presenting information in a straightforward manner without the dynamic or actionable language typical of persuasive ads. For instance: *"The small business tax rate has been reduced to 25% – the lowest level in 50 years. 4,000 local small businesses in Cowan will be able to access our new 20% bonus deduction for upskilling their staff."* This statement illustrates how low persuasion ads usually provide factual information relevant to a specific audience without aggressive persuasion or emotional appeals. Visual examples of these advertising strategies are shown in Figure 3a and Figure 3b.

TF-IDF Analysis. A TF-IDF analysis reveals potential differences between the lexical choices in highly persuasive ads versus those that are low persuasion. Highly persuasive ads appear to use words that evoke broader national themes and encourage immediate action, such as "vote", "government", "future", "need", "better", and "plan". These words suggest that the ads may employ a strong, action-oriented vocabulary, potentially tapping into nationwide concerns, which could contribute to their persuasiveness. Conversely, words prevalent in less persuasive ads, such as "local", "support", and "team" suggest a more localized or specific focus. These terms might not provide a sense of urgency, potentially making these advertisements less engaging. Words like "community", "vote", "Australia", and "government" appear equally in both types of ads, indicating their general importance in Australian political advertising. Further research is needed to establish a definitive link between specific words and their persuasion effect. Nonetheless, these initial observations provide a foundation for future studies examining the relationship between language and persuasion in political advertising. Find more details in the *Text Pre-processing Setup* in Appendix including **TF-IDF** and **Bi-grams**

Temporal Dynamics. Figure 4 presents a time series analysis of the APA22 dataset, illustrating the daily trends of mean spend, mean impressions, and ad count for high and low persuasion ads. A noticeable divergence occurs after the call for election (April 10$^{\text{th}}$), where high persuasion ads exhibit a significant surge in all three metrics compared to their low persuasion counterparts. Specifically, the mean spend and impressions for high persuasion ads show a statistically significant increasing trend (Mann-Kendall test, $p < .001$, with significance threshold $\alpha = .05$) leading up to elec-



(a) High persuasion ad by Amnesty International Australia advocating for Dr. Miranda Ruiz, utilizing fear-based messaging and strong emotions to provoke an immediate emotional response, potentially enhancing its persuasive impact.



(b) Ad by the Liberal Party of Australia, employing a more informational and optimistic approach. It focuses on practical benefits and community well-being, which is less emotionally charged compared to the high persuasion ad.

Figure 3: Comparison of two political advertisements with differing persuasive intensities. While complete ads are shown for context, our analysis focused exclusively on the textual content.

tion day (May 21st), peaking sharply at 4.8 times the value recorded on April 10th, three days before the election. This is notable given that Australia imposes a blackout period three days before the election for broadcasters (TV and radio), but not for online services.[7]

In contrast, low persuasion ads maintain a relatively steady trajectory with modest growth (2.2 times increase). This pattern suggests a strategic amplification of high persuasion ad campaigns closer to the election date. There is also a strong correlation (Pearson $R = 0.99$, $p < .001$, with $\alpha = .05$) between aggregated daily spending and impressions, as expected. This suggests that the amount spent generally determines the number of impressions generated. It may also reflect minimal micro-targeting, since more targeted ads—being more expensive per impression—would typically result in fewer impressions for the same budget. The dataset from the Meta Ad Library does not allow further exploration of targeting strategy.

These insights, derived from our model predictions on the unlabeled APA22 data, provide an initial understanding of the use and impact of persuasive political advertising strategies. The findings underscore the potential of NLP models to extract subtle patterns from political ad content, paving the way for more extensive analyses in future research.

Findings and Implications

Study 2's examination of model adaptability and persuasive content analysis in APA22 yielded the following insights:

Domain Adaptability. The model trained on the SemEval-23 dataset, primarily comprising news articles, showed limited effectiveness when applied to the social media political ads domain. This stresses the necessity for training models on domain-specific datasets to ensure their efficacy in real-world applications.

Persuasion Detection Performance. Substantial improvements in accuracy and F1-score were observed when

[7]https://www.acma.gov.au/election-and-referendum-blackout-periods

Attribute	High Persuasion	Low Persuasion
Ads (%)	26,419 (46.4%)	7,216 (12.6%)
Avg Impressions	25,407	17,441
Avg Ad Spending ($)	393.6	265.6
Avg Ad Duration	11 days	8 days
Top Funding Entity	Solutions for Australia	Liberal Party of Australia
Top Bi-grams	"morrison scott" "better future"	"business small" "health mental"
Top Words	government, future, change	community, local, support

Table 6. Statistics of political ads in the APA22 dataset using our PPAsy-XLNet model predictions. High persuasion ads consist of 80% or more persuasive sentences, while low persuasion ads contain 20% or fewer. Top funding entities are those that spent the most on ads. Bi-grams show the most common word combinations in high and low persuasive ads. Top words are identified using tf-idf as explained in Appendix .

PPAsy-models (Table 5) were retrained on the APA22 dataset, highlighting the benefits of domain-specific training. Nonetheless, the high efficiency of our PPAsy-XLNet model enables fast retraining reducing the need for domain adaptability.

Persuasive Content Dynamics. High persuasive ads were associated with greater reach and financial investment, suggesting a strategic use of persuasion to maximize impact.

Temporal Dynamics. The escalation in ads spending and impressions for highly persuasive ads as the election approached indicates a strategic intensification of persuasive efforts to maximize influence before the voting day.

Limitations

Study 1. Our study encountered a challenge within the multi-label classification of the SemEval-23 Task 3 Subtask 3 due to the inaccessibility of the test set labels, which hindered our ability to perform a detailed error analysis. Furthermore, the discrepancy in the average number of persuasion techniques used in the development and test sets, as highlighted in Table 1 and Figure 2, indicates some inconsistencies in the data, where persuasion techniques present in the development set did not closely mirror those in the test set, indicating that improvements or decreases in performance on the development set do not necessarily translate to similar changes in the test set. This limitation highlights the difficulty in calibrating our model's performance from development to test conditions, given the different intensity of the features in each set.

Study 2. The adaptation to a binary classification task in our analysis (due to the different annotation schemes used for the SemEval-23 and APA22 datasets) streamlined model training and evaluation but potentially restricted the granularity of our investigation into the depth of persuasive techniques in political ads. This meant that the diversity of persuasive strategies employed across various political campaigns might not have been fully observed. This limitation could impact the model ability to generalize across the wider spectrum of political advertising content encountered in different contexts.

To better understand our model's behavior and limitations, we conducted a comprehensive error analysis that revealed important insights about classification patterns. While our model achieved 81.8% accuracy, the analysis uncovered a systematic bias toward false positives, with neutral sentences containing politeness markers or future-oriented language frequently misclassified as persuasive. Conversely, the model struggled to identify persuasive content in complex, multi-clause sentences or those containing interrogative structures. These findings, detailed in the Appendix (Figures 6 and 7), highlight the challenges in distinguishing between genuine persuasive intent and conventional discourse patterns, suggesting areas for future refinement in political ads classification models.

While our results show strong trends in Figure 4, debiasing techniques like Prediction-Powered Inference (Angelopoulos et al. 2023) or Design-Based Supervised Learning (Egami et al. 2024) could further mitigate potential temporal biases in classification errors, especially when analyzing datasets with less pronounced patterns.

Additionally, we experienced some limitations that inevitably arise when using data from the Meta Ad Library. The analysis was limited to political ads posted on Facebook, which may not fully represent the broader political campaigning strategies employed across different platforms. Our dataset comprised only ads explicitly flagged as political by the entity which posted it[8]. This introduces potential biases, as some political ads might not have been flagged and were thus excluded from our data collection and analysis, while others could have been mis-flagged as political when they were not, which is a phenomenon that has been studied by Sosnovik and Goga (2021). Furthermore, Facebook reports only cover ranges of ad spend and reach, rather than precise figures, which can obfuscate the true scale and impact of individual ads. Lastly, the ad buying tool on Facebook allows advertisers to target audiences with a level of specificity that is not captured in the publicly available Ad Library data, concealing critical aspects of ad

[8]It is legally required for political ads in Australia to have a "paid by" disclaimer

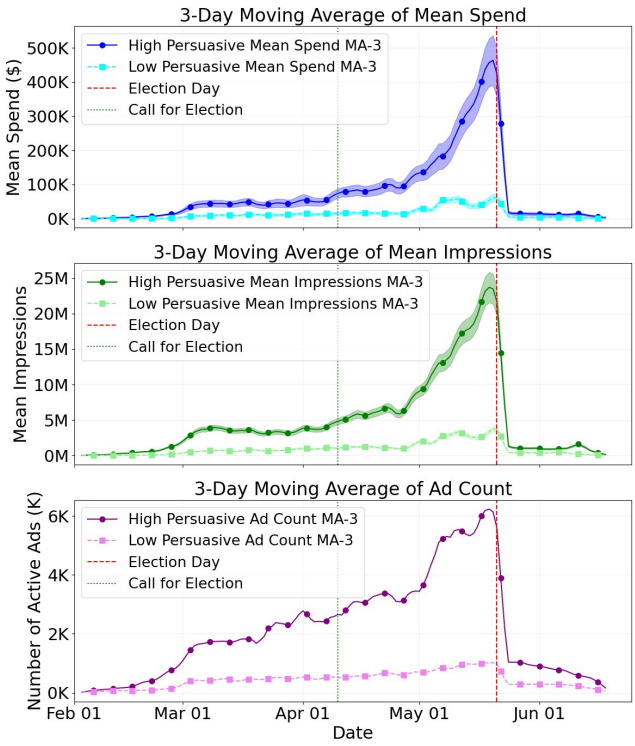


Figure 4: Time series of the total AUD spend, the number of impressions, and the number of unique ads for each day in the APA22 dataset. In the two upper panels the dashed area represents lower and upper bounds, with the solid line corresponding to the mean value. In all plots we use a 3-day moving average.

targeting strategies from our analysis.

Our study specifically analyzes textual content of political ads and does not include visual elements. This represents a clear limitation, as persuasive communication on social media frequently relies on images or multimodal interactions, potentially carrying persuasive signals independent of or complementary to the text. Future work should integrate multimodal analysis frameworks to capture these additional persuasive dimensions and provide a fuller picture of persuasion in digital political advertising.

Conclusion

In this work, we showcased the potential of leveraging our model to analyze political advertising through the lens of persuasion techniques, using an efficient and low-cost PPAsy-XLNet transformer on APA22 dataset. Our results demonstrated the model's ability to outperform resource-intensive approaches on the SemEval-23 dataset, showing how a simple model with a feature-rich subset and a customized loss function can achieve SOTA compared to an ensemble of large transformers. We further explored the adaptability of the model to a new domain, observing a performance dip when transitioning to the social media political advertising context. Retraining PPAsy-XLNet on a small subset of manually annotated instances from the APA22 dataset successfully restored performance, underscoring the importance of domain-specific fine-tuning for achieving optimal performance. By applying the model to APA22 dataset, we gained initial insights into the use of persuasive language and strategies in political advertisements on social media. These include differences in the persuasion intensity and the prevalence of targeted messaging. While this analysis serves as a proof-of-concept, it highlights the practical potential of computational methods for studying persuasive communication in social media, contributing valuable tools for research in political discourse. These advancements not only enhance transparency and deepen our understanding of how persuasive techniques shape public opinion and influence electoral outcomes but also open new avenues for addressing broader societal challenges, such as mitigating manipulation and fostering informed public engagement.

Ethics Statement

The dataset collection for our study, referred to as *APA22*, underwent ethics review by authors' institution IRB. Once the data was collected, the authors were involved in the manual annotation process.

Potential Negative Impacts and Misuse

We acknowledge that models trained to detect persuasive language could potentially be misused in several ways:

Manipulation and Deception. Malicious actors could use our model to craft more effective persuasive content by avoiding detection patterns, potentially enhancing the sophistication of disinformation campaigns or manipulative political messaging.

Censorship and Suppression. The model could be misappropriated to automatically flag or suppress legitimate political discourse, potentially limiting freedom of expression under the guise of filtering persuasive content.

Adversarial Exploitation. Understanding the model's classification patterns could enable bad actors to deliberately craft neutral-appearing messages that contain hidden persuasive elements, exploiting the identified weaknesses in our error analysis.

Privacy Concerns. While our dataset uses publicly available political communications, the ability to systematically analyze persuasive patterns at scale raises concerns about political profiling and targeted manipulation.

Mitigation Strategies

Access Controls. The dataset will be made available exclusively for academic research purposes through a formal request process that includes intended use statements and institutional verification.

Usage Guidelines. We provide comprehensive documentation explicitly prohibiting use cases that could harm democratic processes, manipulate public opinion, or violate privacy rights.

Technical Limitations. We openly discuss our model's limitations and biases (particularly the false positive tendency with polite language) to prevent overreliance on automated decisions. We recognize the potential implications of our work on the broader context of digital persuasion and are committed to promoting ethical standards in computational linguistic research about political communication. We encourage researchers using our resources to consider the societal impact of their applications and to prioritize transparency and accountability in their work.

Acknowledgments

This work is partially supported by the Australian Research Council (ARC) Centre of Information Resilience (Grant No. IC200100022) and by an ARC Future Fellowship Project (Grant No. FT240100022).

References

2016. Detection of Promoted Social Media Campaigns. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*.

2024. The Persuasive Power of Large Language Models. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM)*.

Ali, M.; Sapiezynski, P.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging. *arXiv preprint arXiv:1912.04255*.

Angelopoulos, A. N.; Bates, S.; Fannjiang, C.; Jordan, M. I.; and Zrnic, T. 2023. Prediction-powered inference. *Science*, 382(6671): 669–674.

Appel, R.; Matz, S. C.; and Kosinski, M. 2024. The Persuasive Effects of Political Microtargeting in the Age of Personality Profiling. *PNAS Nexus*, 3(2): pgae035.

Artetxe, M.; Goswami, V.; Bhosale, S.; Fan, A.; and Zettlemoyer, L. 2023. Revisiting Machine Translation for Cross-lingual Classification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6489–6499. Singapore: Association for Computational Linguistics.

Barrett, B.; and McGregor, S. C. 2024. No Better Than Soup? Comparing Null Experimental Effects of Political Facebook Ads Across Persuasive and Instrumental Measures of Effectiveness. *Social Media + Society*, 10(1): 20563051251316117.

Barrón-Cedeño, A.; Jaradat, I.; Da San Martino, G.; and Nakov, P. 2019. Proppy: A system to unmask propaganda in online news. *Information Processing & Management*, 56(5): 1849–1864.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 440–447. Prague, Czech Republic: Association for Computational Linguistics.

Boerboom, C. 2020. Cambridge Analytica: The Scandal on Data Privacy. *Augustana Center for the Study of Ethics Essay Contest*.

Chernyavskiy, A.; Ilvovsky, D.; and Nakov, P. 2021. Transformers:"the end of history" for natural language processing? In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, 677–693. Springer.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Da San Martino, G.; Yu, S.; Barrón-Cedeño, A.; Petrov, R.; and Nakov, P. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5636–5646.

Darmon, A. N.; Bazzi, M.; Howison, S. D.; and Porter, M. A. 2021. Pull out all the stops: Textual analysis via punctuation sequences. *European Journal of Applied Mathematics*, 32(6): 1069–1105.

Dimitrov, D.; Ali, B. B.; Shaar, S.; Alam, F.; Silvestri, F.; Firooz, H.; Nakov, P.; and Martino, G. D. S. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Edwards, A.; and Camacho-Collados, J. 2024. Language Models for Text Classification: Is In-Context Learning Enough? arXiv:2403.17661.

Egami, N.; Hinck, M.; Stewart, B.; and Wei, H. 2024. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36.

Gao, W.; McKeown, K.; and Shivade, N. 2011. Towards a Unified Lexicon of Persuasion Tactics. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 21–29. Portland, Oregon: Association for Computational Linguistics.

Glowacki, M.; Narayanan, V.; Maynard, S.; Hirsch, G.; Kollanyi, B.; Neudert, L.-M.; Howard, P.; Lederer, T.; and Barash, V. 2018. News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. *The Computational Propaganda Project*.

Habernal, I.; and Gurevych, I. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics*, volume 43, 125–179. MIT Press.

Hristakieva, K.; Cresci, S.; Da San Martino, G.; Conti, M.; and Nakov, P. 2022. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM Web Science Conference 2022*, 191–201.

Hromadka, T.; Smolen, T.; Remis, T.; Pecher, B.; and Srba, I. 2023. Kinitveraai at semeval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. *arXiv preprint arXiv:2304.11924*.

Jose, J.; and Greenstadt, R. 2024. Are Large Language Models Good at Detecting Propaganda? *AAAI Conference on Artificial Intelligence*.

Li, Y.; Goldwasser, D.; and Lee, L. 2019. A Survey on Domain Adaptation for Neural Networks: From Single-Task Learning to Multi-Task Learning. *Journal of Machine Learning Research*, 20(129): 1–50.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Meguellati, E.; Zeghina, A.; Sadiq, S.; and Demartini, G. 2025. LLM-based Semantic Augmentation for Harmful Content Detection. arXiv:2504.15548.

Nakov, P.; Alam, F.; Shaar, S.; Da San Martino, G.; and Zhang, Y. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 997–1009.

Nawrot, P.; Tworkowski, S.; Tyrolski, M.; Kaiser, Ł.; Wu, Y.; Szegedy, C.; and Michalewski, H. 2021. Hierarchical transformers are more efficient language models. *arXiv preprint arXiv:2110.13711*.

Perloff, R. 2017. *The Dynamics of Persuasion: Communication and Attitudes in the Twenty-First Century*. Routledge, 6th edition.

Piskorski, J.; Stefanovitch, N.; Da San Martino, G.; and Nakov, P. 2023a. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–2361.

Piskorski, J.; Stefanovitch, N.; Nikolaidis, N.; Da San Martino, G.; and Nakov, P. 2023b. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3001–3022.

Purificato, A.; and Navigli, R. 2023. APatt at SemEval-2023 Task 3: The Sapienza NLP System for Ensemble-based Multilingual Propaganda Detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 382–388.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937.

Rasooli, M.; and Sproat, R. 2018. Exploring Text Classification for Short Social Media Messages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 296–306. Melbourne, Australia: Association for Computational Linguistics.

Salman, M. U.; Hanif, A.; Shehata, S.; and Nakov, P. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. *arXiv preprint arXiv:2305.14534*.

Silva, M.; Oliveira, L. S. d.; Andreou, A.; Vaz de Melo, P. O.; Goga, O.; and Benevenuto, F. 2020. Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. *arXiv preprint arXiv:2001.10581*.

Sosnovik, V.; and Goga, O. 2021. Understanding the Complexity of Detecting Political Ads. In *Proceedings of the Web Conference 2021*, WWW '21, 2002–2013. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Subbiah, M.; and McKeown, K. 2021. Understanding Identity Signalling in Persuasive Online Text. In *Proceedings of the ICWSM 2021 Workshop on Social Media, Persuasion, and Identity*.

Tardaguila, C.; Benevenuto, F.; and Ortellado, P. 2018. Fake news is poisoning Brazilian politics. WhatsApp can stop it. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html. Accessed: 15 Feb 2024.

Xu, P.; Kumar, D.; Yang, W.; Zi, W.; Tang, K.; Huang, C.; Cheung, J. C. K.; Prince, S. J.; and Cao, Y. 2021. Optimizing Deeper Transformers on Small Datasets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2089–2102. Online: Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR*, abs/1906.08237.

Yu, S.; Martino, G. D. S.; Mohtarami, M.; Glass, J.; and Nakov, P. 2021. Interpretable propaganda detection in news articles. *arXiv preprint arXiv:2108.12802*.

Zarouali, B.; Dobber, T.; De Pauw, G.; and de Vreese, C. H. 2020. Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research*, 47(8): 1050–1076.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Further Discussion

Why did we Use the French Subset?

As shown in Table 1, French has the highest Persuasive techniques per document compared to the other languages where we observed a trend of the higher the concentration of the techniques per document the better the models perform. We experimented with different subsets. As it can be seen below, there is a pattern where a higher average number of persuasive techniques in the document corresponds to a higher F1-micro score:

- **French:** 40.62% F1-micro, avg. persuasive techniques per doc: 29.9
- **German:** 38.01% F1-micro, avg. persuasive techniques per doc: 27.5
- **Italian:** 36.8% F1-micro, avg. persuasive techniques per doc: 25.4
- **All languages:** 39.71% F1-micro, avg. persuasive techniques per doc: 25.9

Additionally, Table 4 shows that the performance of the French-to-English subset has a comparable performance with a slight advantage in most cases compared to All languages translated to English set

Why did we Use Machine Translation?

Our approach builds upon recent advancements in cross-lingual classification, particularly emphasizing the benefits of machine translation (MT) and focusing on data efficiency. Artetxe et al. (2023) demonstrated, the translate-set approach can perform substantially better when utilizing a more robust MT system and addressing the mismatch between training on original text and inferencing on translated text (see the effect of removing translation in the ablation study in Table 3. This revisited perspective challenges the prevailing dominance of multilingual models, revealing that MT-based baselines can be highly effective. By translating the dataset from French to English, we leverage these insights to potentially improve classification accuracy while maintaining computational efficiency As shown in Table 4 where XLNet-Base with french-to-englsih subset significantly outperforms XLM-RoBERTa-Large. This strategy aligns with the growing body of evidence suggesting that well-implemented translation methods can rival or surpass the performance of complex multilingual models in certain cross-lingual tasks, offering a compelling alternative in scenarios where resource optimization is crucial.

Why not use newer models?

The approach we propose in this work aligns with research showcasing the enduring competitiveness of older transformer models. For example, Nawrot et al. (2021) introduced the Hourglass model, a hierarchical transformer that optimizes the baseline transformer for efficiency, while Xu et al. (2021) explored deeper transformers optimized for small datasets, demonstrating that older models can be enhanced through optimization techniques, illustrating how BERT and its derivatives remain effective methods for text classification tasks.

Our decision to utilize XLNet in specific, was primarily driven by considerations of data efficiency and model performance. While RoBERTa-base demonstrated superior performance, it required the entire dataset of 1.16 million tokens to achieve its results. In contrast, XLNet-base achieved comparable performance using only 153,000 tokens, which is approximately 13% of the data used by RoBERTa. This significant reduction in data requirements, coupled with similar performance outcomes as shown in Table 4, highlights XLNet's efficiency in learning from limited data. Such data efficiency is particularly valuable in scenarios where labeled data is scarce or expensive to obtain, as is often the case in specialized domains or low-resource languages.

Why Traditional Statistical Tests Were Not Applied?

The structure of this particular SemEval task presents unique challenges that preclude the application of traditional statistical significance testing:

1. **Blind Test Set:** The evaluation is conducted on a completely blind test set. Participants submit predictions on unlabeled data and receive only aggregate scores, not per-instance results. This lack of detailed feedback limits our ability to perform instance-level statistical analyses.
2. **Development-Test Set Mismatch:** As noted in our limitations section and Figure 2, we observed a significant discrepancy between the performance on the development and test sets. This mismatch undermines the reliability of using the development set as a proxy for significance testing on the test set.
3. **Limited Access to Comparative Data:** We do not have access to other models' per-instance predictions on the test set. Furthermore, the unavailability of complete implementation details for competing models prevents us from accurately replicating their performance, which is necessary for methods such as McNemar's test.
4. **Focus on Resource Efficiency:** Our primary contribution lies in demonstrating that a lightweight model can achieve high performance with considerably lower resource requirements. This point is effectively illustrated by our results (Table 2), making the statistical significance between top-performing models less central to our main argument.

Why did we Binarize the Annotations?

Due to differences in the annotation schemes between our APA22 dataset and the SemEval-23 dataset, performing multi-label classification directly was not feasible. The APA22 dataset does not contain the same detailed persuasive technique labels as SemEval-23, making a fair multi-label classification impractical. To address this, we binarized both datasets into persuasive and non-persuasive classes before training our models.

Binarizing the data prior to training allowed us to focus on the presence or absence of persuasion, facilitating a direct comparison between the two datasets. This decision was made for the following reasons:

Direct domain adaptation: Binarizing before training enabled us to assess the model's ability to adapt to a new domain without the confounding effects of differing label granularities. It provided a clearer evaluation of how well a model trained on one dataset (SemEval-23) could generalize to another (APA22) when both are considered in terms of persuasive versus non-persuasive content.

Computational efficiency: Simplifying the classification task to a binary problem reduced the complexity of the model training process. Training on binary labels requires fewer computational resources compared to multi-label classification involving 23 distinct labels, as in the original SemEval task. This efficiency was particularly beneficial given the scale of the datasets.

While binarizing predictions from models trained on the full label set could offer additional insights, we found that binarizing the data before training was more appropriate for our study's objectives. This approach allowed us to directly evaluate the model's performance in detecting persuasive content across different domains and datasets.

Why 20/80 Ratio?

In our analysis of the APA22 dataset, we sought to categorize advertisements based on their level of persuasive content to better understand the strategies employed in political advertising. Initially, we considered a binary classification approach where an ad would be labeled as *persuasive* if it contained at least one persuasive sentence and *non-persuasive* otherwise. However, this method resulted in less than 15% of the total ads being classified as non-persuasive. This low percentage was intuitive, given that advertisements are inherently designed to persuade audiences. Such a classification did not align with the fundamental purpose of political ads and risked oversimplifying the persuasive dynamics present in the dataset.

To address this challenge, we introduced a more nuanced classification scheme that distinguishes between *low persuasion* and *high persuasion* ads based on the proportion of persuasive sentences within each ad. Specifically, we defined:

- **Low persuasion ads**: Advertisements where 20% or fewer of the sentences are classified as persuasive.
- **High persuasion ads**: Advertisements where 80% or more of the sentences are classified as persuasive.

This threshold-based categorization allows us to capture a spectrum of persuasive intensity, acknowledging that ads may employ persuasion to varying degrees. The choice of the 20% and 80% thresholds was informed by exploratory analysis and experimentation with different ratios, such as 10%–90% and 30%–70%. We observed that while varying the thresholds slightly affected the distribution of ads between the categories, the overall patterns and insights remained consistent. The 20%–80% split provided a meaningful distinction between ads with minimal persuasive content and those heavily relying on persuasive techniques.

By adopting this approach, we ensured that our analysis reflects the inherent persuasive intent of political advertising while allowing for a granular examination of how the degree of persuasion correlates with other variables, such as ad spending, impressions, and campaign duration. This methodology aligns with practices in computational social science and natural language processing research, where nuanced categorization facilitates more accurate modeling and interpretation of complex social phenomena.

Furthermore, the 20%-80% threshold provides important robustness against classification errors at the sentence level. Our error analysis,in the appendix, revealed that the sentence classifier achieves 81.8% accuracy with a tendency toward false positives. By requiring 80% of sentences to be classified as persuasive for an ad to be considered "high persuasion," we create a buffer against individual misclassifications. This threshold ensures that occasional errors in sentence-level predictions do not cascade into ad-level misclassifications, as an ad would need multiple correctly identified persuasive sentences to meet the threshold. Conversely, the 20% threshold for "low persuasion" ads provides similar protection against false negatives, ensuring that ads with minimal persuasive content are not mislabeled due to a few erroneous sentence classifications. This approach effectively transforms our sentence-level predictions into a robust voting mechanism at the ad level, where the collective evidence from multiple sentences provides more reliable categorization than any single sentence classification.

Text Pre-processing Setup

In Study 2, we applied a series of pre-processing steps to the text data obtained from the ads to facilitate linguistic analysis using TF-IDF and n-grams. The pre-processing pipeline included converting the text to lowercase, removing punctuation, emojis, and links, and stripping whitespaces. Additionally, we filtered out stopwords to focus on more semantically meaningful terms, and tokenized the text into individual units. These pre-processing steps were useful for reducing noise and improving our linguistic analysis.

TF-IDF

For the analysis of the textual component of the ads, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique to assess the importance of words within the corpus. This method measures the frequency of a word in a single advertisement and also accounts for the popularity of the word across all ads.

To ascertain a global significance score for each word across all ads, we computed the average TF-IDF score for each term across the dataset. This average score represents the overall importance of each word within the entire corpus. The results of our TF-IDF analysis are summarized by listing the 10 words with the highest average TF-IDF scores for both high and low persuasion ads (see Table 8).

Bi-gram

Our methodology also involves the extraction and evaluation of n-grams from text data to uncover frequently occurring word patterns. These n-grams are sorted to ensure that identical phrases with words in different orders are counted as the same n-gram. For instance, ('word1', 'word2') and ('word2', 'word1') are counted as a single bi-gram. We

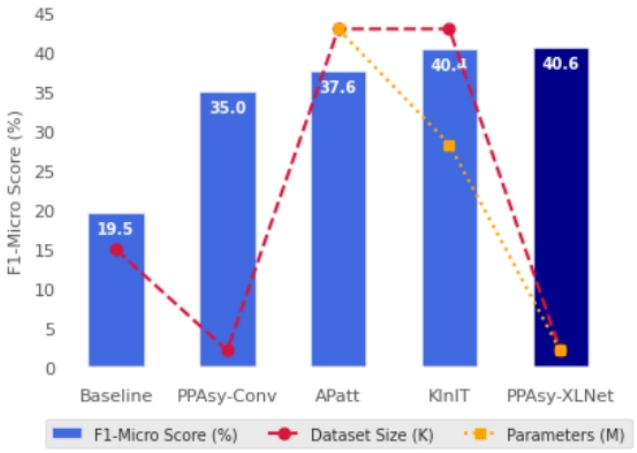


Figure 5: Performance comparison of models on the English test set, highlighting F1-Micro scores, dataset size, and parameter count.

chose bi-grams because they provide an optimal balance between capturing meaningful lexical relationships and ensuring a sufficient frequency of occurrences. Bi-grams allow us to observe the most significant word pairs, providing insights into the structure of persuasive language commonly used in political advertising.

While word co-occurrence matrices provide useful insights, particularly in identifying broader semantic connections and themes within large text corpora, they can sometimes reduce the focus on direct linguistic interactions, which are crucial in advertising. In contrast, N-gram analysis offers a more targeted approach by directly highlighting the sequences of words that frequently occur together. This provides a clearer view of common linguistic patterns and potentially persuasive language strategies used in online advertisements.

Persuasion Technique	**Distribution (%)**
Loaded Language	42.65
Name Calling-Labeling	33.17
Doubt	24.95
Questioning the Reputation	13.43
Exaggeration-Minimisation	11.61
Appeal to Fear-Prejudice	11.47
Conversation Killer	6.95
Appeal to Hypocrisy	6.50
Repetition	6.31
Appeal to Authority	6.11
Appeal to Values	5.11
Slogans	4.95
Guilt by Association	4.87
Flag Waving	4.85
Causal Oversimplification	4.45
False Dilemma-No Choice	3.18
Obfuscation-Vagueness-Confusion	2.80
Appeal to Popularity	2.61
Consequential Oversimplification	2.44
Straw Man	2.37
Red Herring	1.63
Whataboutism	1.18
Appeal to Time	1.16

Table 7. Distribution of Persuasion Techniques of the SemEval Dataset

	High Persuasion	**Low Persuasion**
TF-IDF	labor	community
	vote	local
	government	support
	morrison	team
	future	labor
	need	time
	better	nsw
	australia	join
	community	service
	plan	vote

Table 8. Ten highest TF-IDF words for high and low persuasive ads respectively

	High Persuasion	**Low Persuasion**
Bi-grams	morrison scott	election federal
	government labor	job vacancy
	change climate	melbourne vic
	better future	government labor
	aged care	strait torres
	cost living	islander strait
	petition sign	aboriginal torres
	economy strong	business small
	future stronger	health mental
	government morrison	candidate labor

Table 9. Top ten bi-grams for high and low persuasive ads respectively

Error Analysis

To better understand our model's behavior beyond the 81.8% accuracy metric, we conducted a comprehensive error analysis on the test set predictions. The confusion matrix (Figure 6) reveals an interesting asymmetry in the error distribution, with 20 false positives (neutral sentences classified as persuasive) compared to 10 false negatives (persuasive sentences classified as neutral). This 2:1 ratio suggests our model has developed a heightened sensitivity to persuasive linguistic features, which merits further investigation.

Analysis of the confidence distributions (Figure 7, top left panel) provides valuable insights into the model's calibration characteristics. While the overall performance is strong, we observe that errors are not uniformly distributed across confidence levels. Specifically, 60% of errors occur in the high confidence range (above 0.7 or below 0.3), indicating areas where the model exhibits strong conviction despite being incorrect. The box plot analysis (Figure 7, top right panel) clearly illustrates this pattern, showing that false positives have significantly higher median confidence compared to false negatives, which cluster around the 0.2-0.4 range. False positives show notably higher average confidence (0.793) compared to false negatives (0.317), suggesting the model is more decisive when incorrectly identifying persuasive content.

The error rate analysis by confidence level (Figure 7, middle right panel) reveals a concerning pattern: the highest error rates occur in the 0.4-0.6 confidence range (approximately 55%), which represents the model's uncertainty zone. However, substantial error rates persist even in high-confidence regions, with the 0.8-1.0 range showing approximately 15% error rate. This indicates that while the model's confidence generally correlates with accuracy, it maintains a non-trivial error rate even when highly confident.

Linguistic analysis of the misclassified sentences reveals interpretable patterns that explain these errors. In the false positive category, we frequently observe sentences containing politeness markers ("please," "thank you"), future-oriented language ("will reduce," "will mean"), and expressions of support or commitment. For example, simple courtesies like "Have a great day!" and "Thank you for your support!" achieved confidence scores of 0.909 and 0.849 respectively, suggesting the model associates these social conventions with persuasive intent. This pattern likely emerges from the prevalence of polite language in genuine persuasive content within our training data.

Conversely, false negatives often include sentences with interrogative structures, technical references (URLs, statistics), and complex multi-clause constructions. The model appears less confident when encountering persuasive content that deviates from typical call-to-action patterns, such as questions like "Do you live in Camden LGA, Wollondilly, Goulburn...?" (confidence: 0.197) or factual statements containing numerical data.

Sentence length analysis (Figure 7, middle left panel) reveals additional patterns, with false positives generally appearing in shorter sentences with high confidence, while false negatives are distributed across various lengths but consistently show low confidence. The word count analysis (Figure 7, bottom panel) further confirms this pattern, demonstrating that false positives tend to be concise statements with high confidence, while false negatives often involve longer, more complex constructions that the model struggles to classify confidently. This suggests the model may have learned to associate brevity and directness with persuasive intent, while missing more nuanced persuasive strategies in longer texts.

These findings highlight the inherent challenges in distinguishing between genuine persuasive intent and conventional polite discourse in textual communication. The model's behavior aligns with human annotation challenges in this domain, where context and subtle linguistic cues play crucial roles. The observed patterns provide valuable directions for future improvements, including enhanced feature engineering to better capture contextual nuances and potential data augmentation strategies to address the identified biases. Overall, while the model demonstrates strong baseline performance, this analysis illuminates specific areas where targeted improvements could enhance both accuracy and calibration.

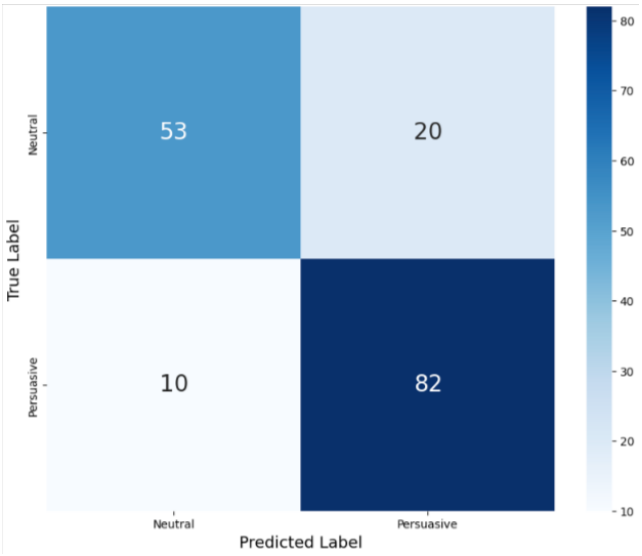


Figure 6: Confusion matrix showing the distribution of predictions. The asymmetry between false positives (20) and false negatives (10) indicates the model's heightened sensitivity to persuasive features.

Computational Resources Used for the Experiments

The experiments were performed using a Google Colab environment with an NVIDIA A100 GPU (40 GB of GPU RAM) and 83.5 GB of system RAM.

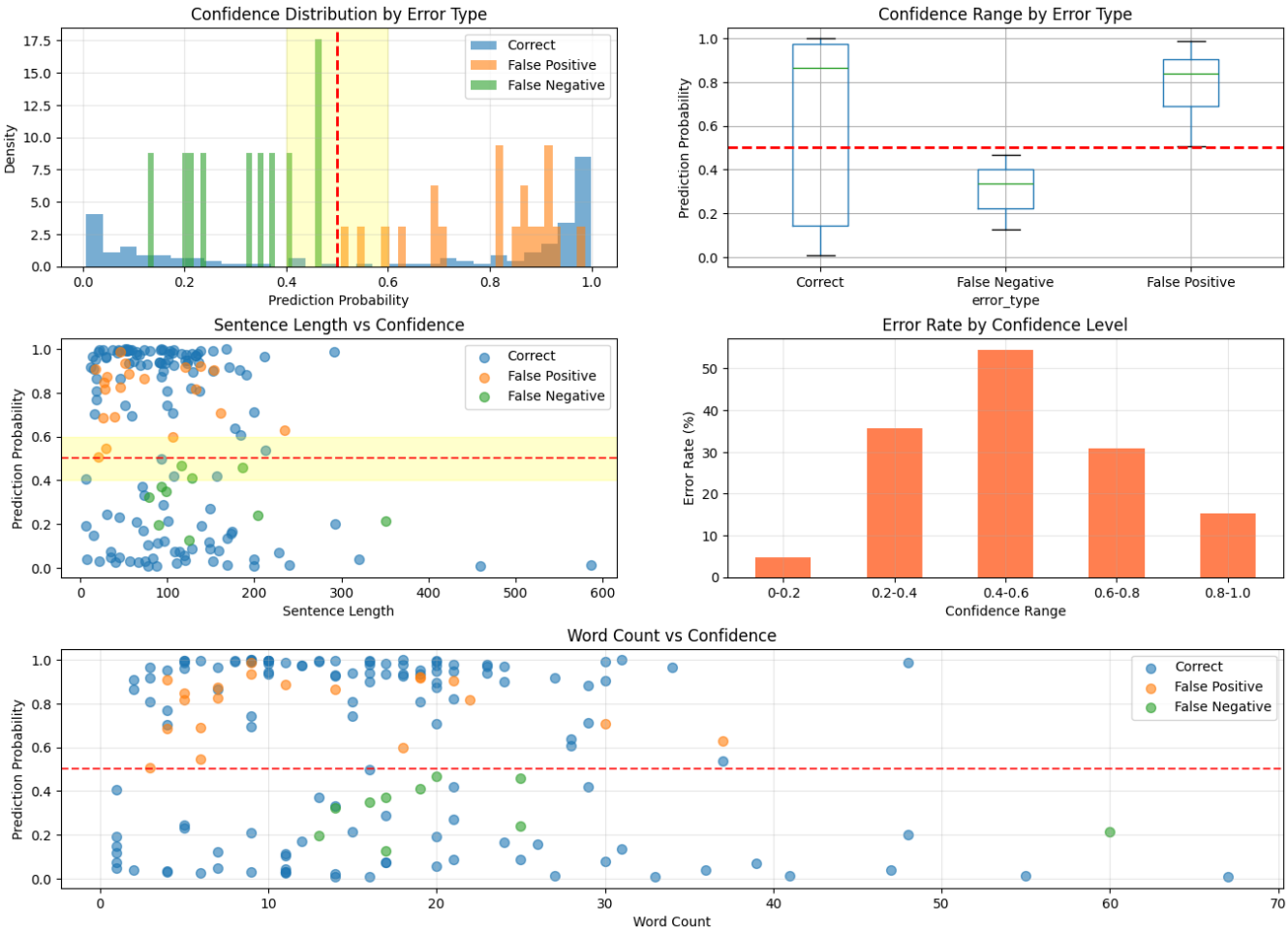


Figure 7: Comprehensive error analysis showing confidence distributions, sentence characteristics, and error patterns across different prediction categories. The borderline confidence zone (0.4-0.6) is highlighted in yellow where applicable.

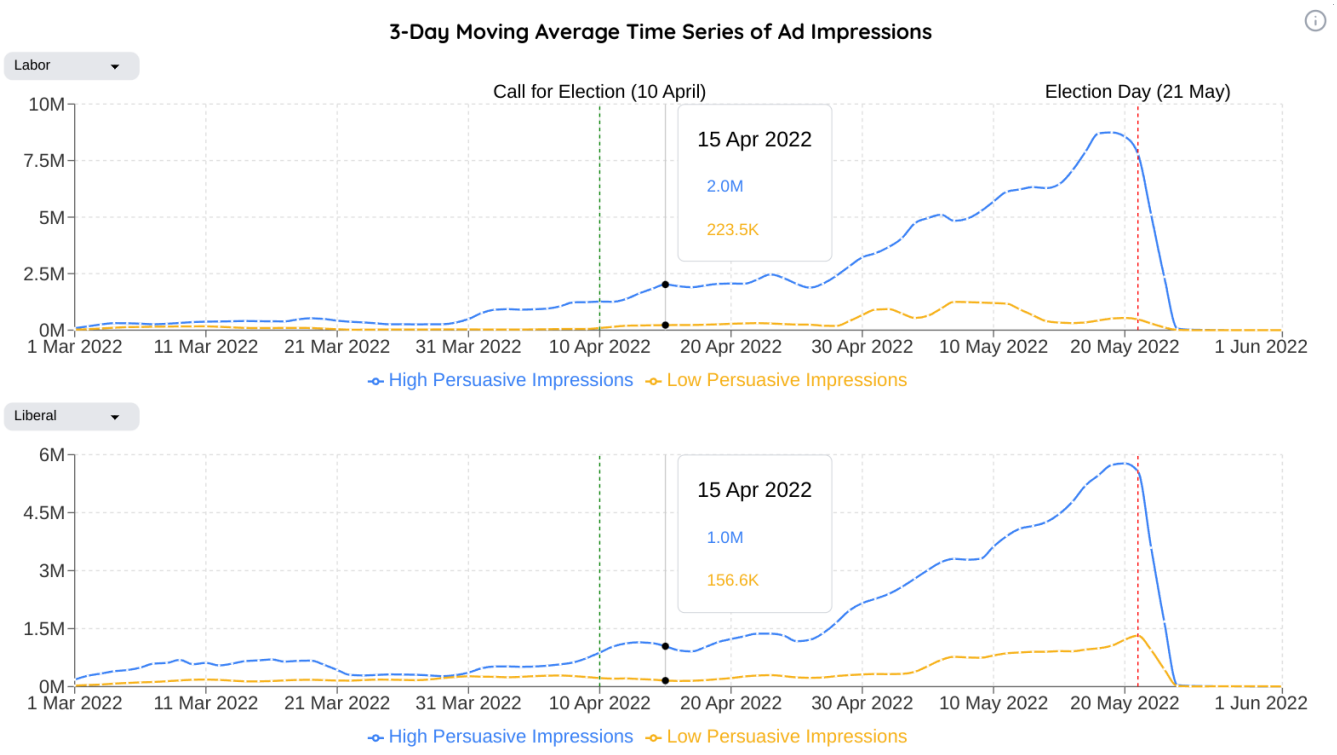


Figure 8: 3-Day Moving Average Time Series showing Impressions of High and Low Persuasive content for Labor and Liberal parties during the 2022 Australian Federal Election campaign.

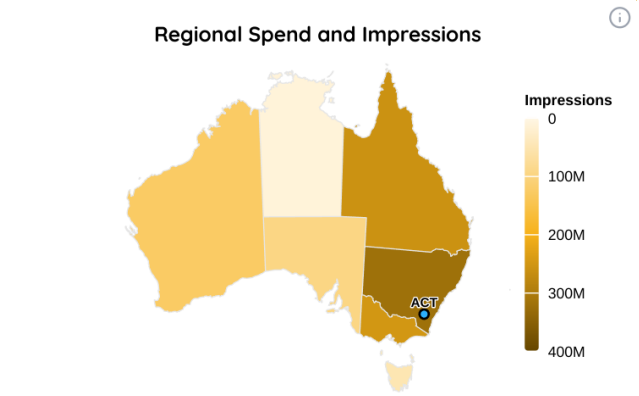


Figure 9: Geographical breakdown of ad impressions during the 2022 Australian Federal Election, showing regions with higher concentrations of political ad spending.

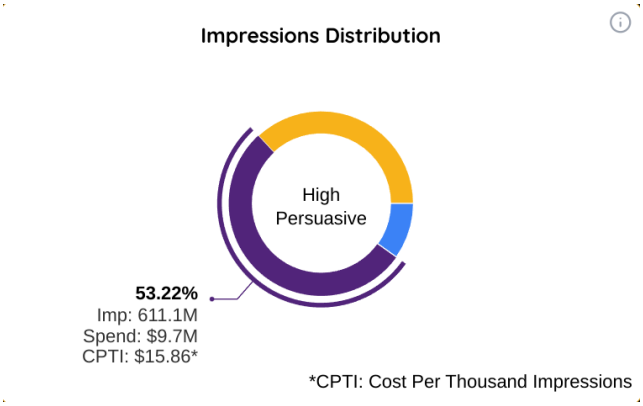
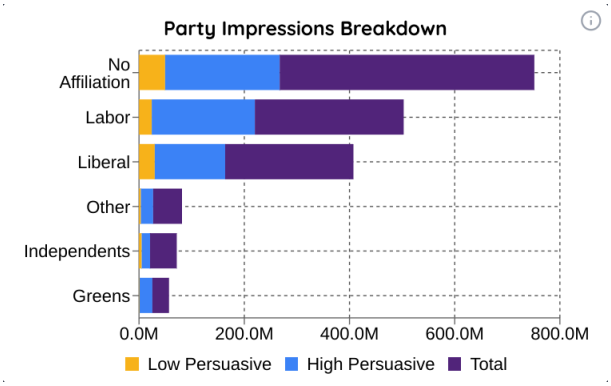
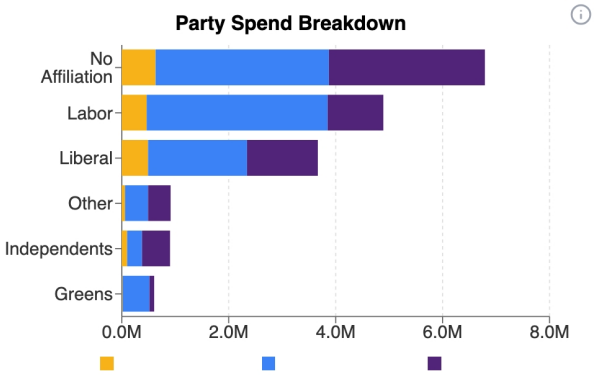


Figure 10: Percentage and cost distribution of impressions for high persuasion ads, illustrating the strategic focus of campaigns on persuasive messaging.

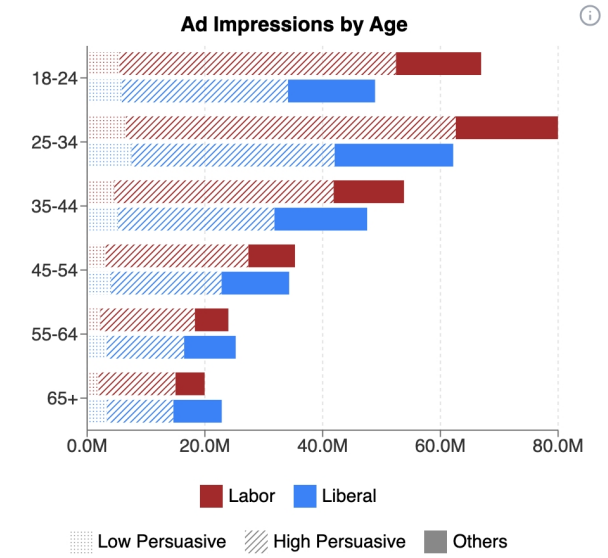


(a) Distribution of impressions for low persuasion, high persuasion, and total ads across different political parties during the 2022 Australian Federal Election.

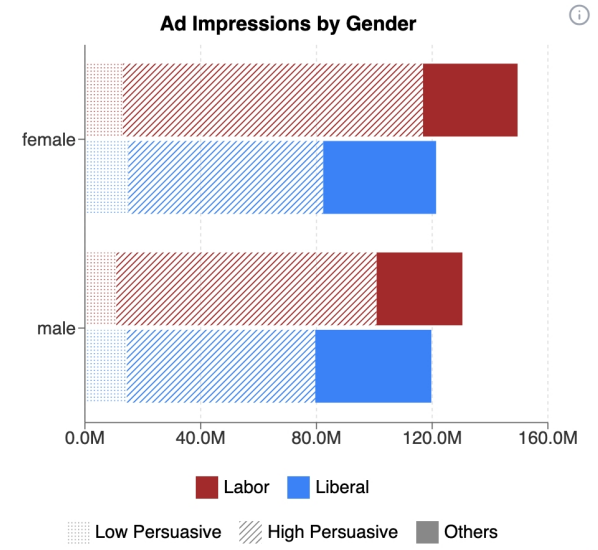


(b) Party spend breakdown across different political parties and affiliations during the 2022 Australian Federal Election.

Figure 11: Comparison of political party impressions and spending during the 2022 Australian Federal Election.



(a) Liberal (blue) vs Labor (red) ad impressions by age group during the 2022 Australian Federal Election.



(b) Liberal (blue) vs Labor (red) ad impressions by gender during the 2022 Australian Federal Election.

Figure 12: Comparison of Liberal (blue) and Labor (red) ad impressions by age and gender during the 2022 Australian Federal Election.

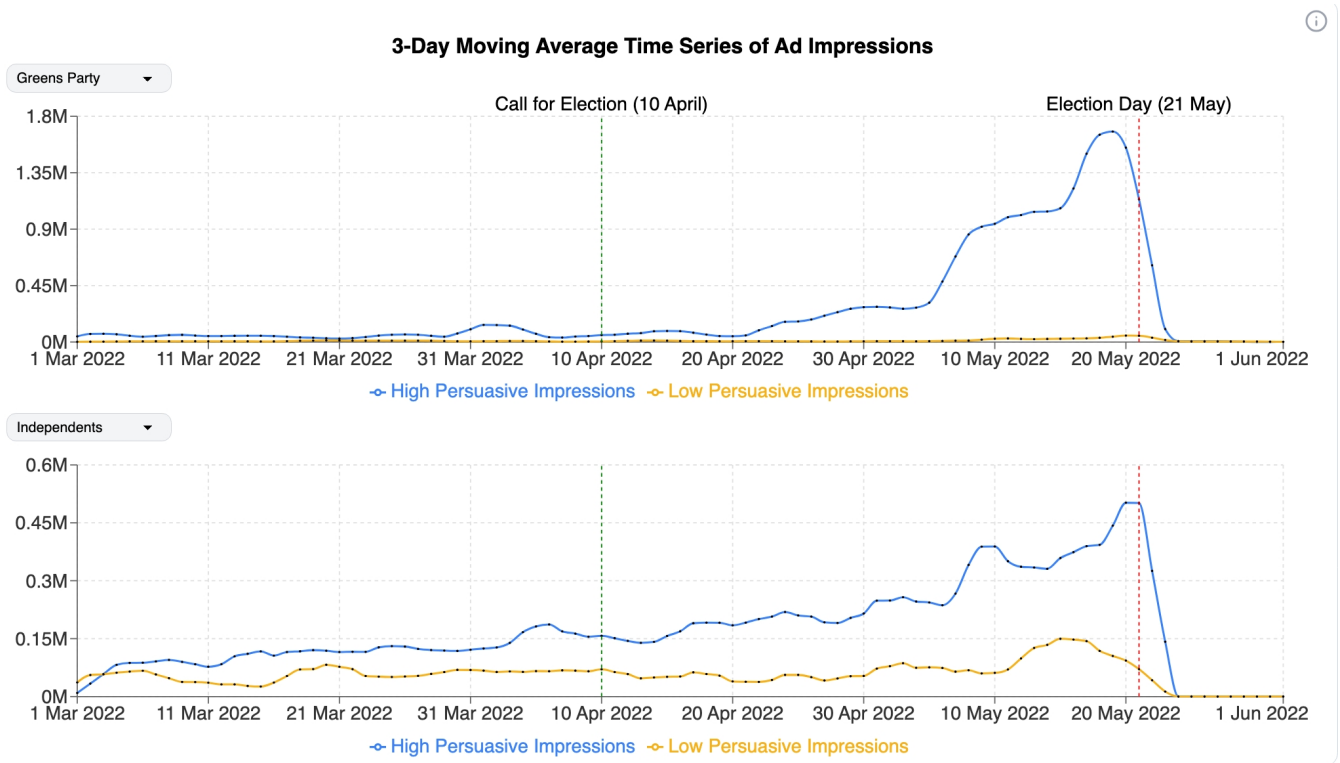


Figure 13: 3-Day Moving Average Time Series showing Impressions of High and Low Persuasive content for Greens and Independents parties during the 2022 Australian Federal Election campaign.