International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Textual Persuasion Analysis for Identifying Manipulative Languages

Bhavana R K[1], Sandeep albert Mathias[2]

[1]Student, School of Computer Science and Engineering, Presidency University
[2]Assistant Professor, School of Computer Science and Engineering, Presidency University

Abstract
Manipulative language, often used in political, commercial, and social discourse, can distort public opinion by appealing to emotions, authority, or fallacious reasoning. This paper presents a comprehensive framework for detecting such manipulation using natural language processing. Leveraging a taxonomy of 25 persuasion techniques, we fine-tune the XLM-RoBERTa transformer model to perform multi-label classification of manipulative content. Additionally, we develop a counter-narrative generation module using a T5-based model to suggest ethically persuasive responses. Our results show promising accuracy in classification across varied techniques and demonstrate the effectiveness of automated counter-speech generation. This work contributes to efforts in enhancing digital media literacy and resisting propaganda in online platforms.

Keywords: Manipulative Language, Textual Persuasion Analysis, Natural Language Processing (NLP) XLM-RoBERTa, Multi-label Classification

1. Introduction

The rapid expansion of online communication has dramatically increased the influence of persuasive and manipulative language across digital platforms. Political campaigns, advertising agencies, and ideological groups frequently exploit linguistic techniques that manipulate emotions, distort logic, or appeal to authority to influence public opinion. As misinformation and propaganda become more sophisticated, the need for automated tools that can detect and counteract such rhetorical strategies grows urgent.

Textual persuasion analysis involves identifying specific language patterns used to subtly or overtly influence readers. Unlike sentiment analysis or fake news detection, which often focus on emotional polarity or factual accuracy, persuasion analysis delves into the structure and intent of discourse. Manipulative techniques like *Loaded Language*, *Appeal to Fear*, and *False Equivalence* operate by triggering cognitive biases or exploiting logical fallacies, making them challenging to detect using surface-level linguistic cues alone.

Recent advances in deep learning, particularly transformer-based language models, offer powerful capabilities for understanding context, semantics, and implicit meaning. These models can be fine-tuned to recognize persuasive techniques and even generate appropriate counter-speech to mitigate their influence. However, the lack of comprehensive datasets annotated with detailed persuasion strategies and the multi-label nature of such texts present unique challenges.

This study addresses these issues by proposing a dual-model NLP framework for detecting manipulative persuasion techniques and generating corresponding counter-narratives. We leverage the multilingual

XLM-RoBERTa model for multi-label classification and a T5-based model for counter-speech generation. Our contributions are threefold:

1. We adopt and refine a 25-label taxonomy of persuasion techniques suitable for detailed rhetorical analysis.
2. We build and augment a dataset encompassing political, health, and social discourse to enable robust classification and response generation.
3. We evaluate the system's ability to both detect manipulation and produce ethically grounded counter-narratives that promote critical thinking and factual dialogue.

By addressing both detection and response generation, our work aims to support efforts in media literacy, content moderation, and AI-assisted critical discourse.

2. Related Work

Research on manipulative language and misinformation has gained momentum with the increasing impact of online media. Several efforts have been directed toward detecting propaganda, disinformation, and fallacious argumentation in textual content. Notably, the work by Da San Martino et al. (2019) introduced a fine-grained annotation schema and dataset for propaganda detection, forming the basis for subsequent studies in the field.

Transformer-based models like BERT, RoBERTa, and especially multilingual models such as XLM-RoBERTa have shown strong performance across various natural language understanding tasks, including sentiment analysis, stance detection, and misinformation classification. These models benefit from pretraining on large corpora and can generalize well to unseen texts.

Argument mining is another closely related area, focusing on identifying argumentative components and their structures in texts. However, it typically deals with well-structured arguments, unlike the often emotionally manipulative and logically flawed content in propaganda. Recent efforts in multi-label classification for rhetorical and persuasive technique identification have addressed the challenge of overlapping persuasive strategies, but most do not go beyond detection.

Counter-speech generation has been explored using both rule-based and neural models, with growing interest in using sequence-to-sequence models like T5 and GPT-2. Nevertheless, integrating classification with response generation in a unified framework remains relatively unexplored. Our work bridges this gap by providing both detection and automatic response mechanisms in a coherent NLP pipeline.

3. Dataset and Taxonomy

Our dataset is composed of three components: (1) the annotated dataset from the SemEval Propaganda Detection shared tasks, (2) manually augmented examples with additional labels for underrepresented persuasion techniques, and (3) a curated set of counter-narratives corresponding to each labeled instance. The taxonomy includes 25 distinct persuasive strategies, such as Guilt by Association, Appeal to Fear, Red Herring, False Equivalence, and Repetition. Each text may be tagged with multiple labels, reflecting the real-world complexity of manipulative messaging.

This dataset spans various domains including political discourse, public health communication (especially COVID-19 vaccine debates), and online opinion forums. Annotations were validated by multiple annotators to ensure label reliability and narrative consistency.

4. Methodology

4.1 Classification Model To perform persuasive technique detection, we fine-tune XLM-RoBERTa using a multi-label classification head. Binary cross-entropy loss is employed for each of the 25 labels. A sigmoid activation function is used at the output layer to obtain label-wise probabilities, which are thresholded using validation-set-optimized values. The model is trained using the Adam optimizer with early stopping to prevent overfitting.

4.2 Counter-Narrative Generation For generating responses, we employ a T5-small model fine-tuned on input-output pairs where the input is the manipulative text with associated technique labels and the output is the counter-narrative. We apply data augmentation techniques including back-translation and synonym replacement to enrich the training set. Outputs are evaluated using BLEU, ROUGE-L, and a human review rubric focusing on ethical reasoning, clarity, and relevance.

5. Experiments and Evaluations

We split the dataset into 70% training, 15% validation, and 15% test sets. For classification, macro-averaged precision, recall, and F1-score are computed. For the generation task, we compare our model outputs against human-written counter-narratives using automatic metrics and a manual quality check.

In addition, ablation studies were conducted to evaluate the effect of multilingual input, data augmentation, and label-specific thresholding. The robustness of the classifier was also tested under domain shift by introducing out-of-distribution inputs from newly collected sources.

6. Results

Persuasion Technique	Precision	Recall	F1-Score
Loaded Language	0.88	0.89	0.88
Appeal to Fear	0.85	0.83	0.84
Name Calling	0.80	0.78	0.79
Overall (Macro Average)	0.82	0.81	0.82

Table:6.1 results obtained

For the generation task, BLEU score averaged 31.6 and ROUGE-L score was 46.8. Human evaluators rated 86% of responses as contextually appropriate and ethically grounded.

7. Discussions

Our integrated approach shows that multi-label classification using XLM-RoBERTa can effectively handle complex, overlapping persuasive techniques. Techniques that rely on emotion (e.g., Loaded Language, Appeal to Fear) are detected more accurately than those that rely on subtle logical flaws (e.g., Red Herring, False Equivalence). Challenges persist in distinguishing between semantically similar categories, suggesting a need for deeper semantic reasoning.

Counter-narrative generation proved to be an effective tool for fostering critical discourse. However, ensuring factuality and stylistic appropriateness under varying cultural contexts remains an open challenge. Further improvements could involve integrating retrieval-augmented generation and reinforcement learning from human feedback.

8. Conclusion

In this study, we presented a comprehensive framework for the detection and mitigation of manipulative language in textual content through advanced NLP techniques. By integrating a multi-label classification

model based on XLM-RoBERTa with a counter-narrative generation model using T5, we successfully demonstrated the ability to not only identify subtle and complex persuasive strategies but also to generate meaningful responses that promote ethical and logical discourse.

Our approach leverages a rich 25-label taxonomy of persuasion techniques, capturing the multifaceted nature of propaganda and rhetorical manipulation. The classification model achieved strong performance metrics, particularly in detecting emotionally charged tactics such as Loaded Language and Appeal to Fear. Meanwhile, the counter-narrative module produced responses that were rated highly for contextual accuracy and ethical clarity, highlighting the potential of generative models in combating misinformation and disinformation.

This dual-model architecture offers significant contributions to the fields of media literacy, automated content moderation, and digital civic education. Unlike prior work that treats persuasion detection and response as separate tasks, our system provides an end-to-end solution capable of both analysis and intervention.

However, challenges remain. The detection of nuanced techniques like Red Herring or False Equivalence requires deeper semantic understanding and discourse-level reasoning, which current models only partially address. Likewise, ensuring that generated responses remain culturally sensitive, non-confrontational, and factually accurate across languages and domains warrants further exploration.

Future directions include expanding the multilingual capabilities of the models, incorporating user feedback to improve adaptive learning, and deploying the system in real-world settings such as online forums, educational platforms, and social media moderation pipelines. Through these enhancements, we aim to contribute toward a more informed, critical, and resilient digital society.

References

1. Da San Martino, G., et al. (2019). Fine-Grained Analysis of Propaganda in News Articles. *EMNLP*.
2. Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.
3. Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.
4. Alhindi, T., et al. (2018). Where is Your Evidence: Improving Fact-checking by Justification Modeling. *EMNLP*.
5. Naderi, N., et al. (2021). Generating Counter Narratives Against Hate Speech: Data and Strategies. *ACL*.
6. **Bender, E. M., et al. (2021).** "Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021).*
7. **Zhou, X., et al. (2020).** "A Survey on Deep Learning for Argumentation Mining." *Computers, 9(3), 42.*
8. **Horne, B. D., et al. (2020).** "The Imminence of Disinformation: A Comparative Study of Fake News and Propaganda." *Proceedings of the 2020 International Conference on Computational Social Science (ICCSS 2020).*
9. **Saha, K., et al. (2020).** "Leveraging Multimodal Data for Misinformation Detection." *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM 2020).*
10. **Binns, R., et al. (2021).** "Understanding the Limits of Counter-Speech: A Study on the Ethical Impli

cations of Automated Responses." *Proceedings of the 2021 ACM Conference on Human-Computer Interaction (CHI 2021).*

11. **Kemp, L. J., & Solomon, A. A. (2021).** "Using Deep Learning Models to Detect Manipulative Discourse in Political Texts." *Journal of Political Communication, 38(4), 467-487.*
12. **Patel, R., & Jackson, S. (2020).** "Counter-Narrative Generation for Social Media Platforms." *Journal of AI Ethics, 8(2), 195-212.*
13. **Volkova, S., et al. (2020).** "Detecting and Mitigating Harmful Bias in Text Generators." *Proceedings of the 2020 International Conference on Learning Representations (ICLR 2020).*