Detecting Propaganda Techniques in News Articles

Isaacs Adeyemo
Computer Science Department,,
University of Calgary
Calgary, Canada
Isaacs.adeyemo@ucalgary.ca

Saad Elkadri
Computer Science Department,
University of Calgary
Calgary, Canada
Saad.Elkadri@ucalgary.ca

***Abstract*—Manipulative narratives and news pieces in mainstream media has created global challenges regarding what information can be trusted or not. Insufficient media literacy is more rampant today due to the propagation of fake news and propaganda. Propaganda is a way of disseminating biased information usually with the goal of influencing people in a particular way. News articles is a popular tool used in propaganda efforts as it is a trusted source of information for many and can be quickly disseminated around the globe. Propaganda can take many forms like false information being presented as fact, biased reporting, or sensationalized headlines. Tackling propaganda at the text level is a crucial step towards preventing the spread of propaganda. In this paper, we aim to combat the spread of misinformation through detecting several propaganda techniques in various news articles through the state-of-the-art pre-trained language model, RoBERTa. The results show that the word-embedding pre-trained model of RoBERTa can detect various propaganda techniques and achieve an F1 score of 60%**

Keywords—Propaganda, technique, detection, fragment, level, RoBERTa, text classification, F1 score.

I. Introduction

In today's digital age, propagation of information has become extremely easy and dangerous. While news articles grant timely and unprecedented access to news, they also serve as breeding grounds for misinformation and propaganda [1]. With social media becoming more popular as a primary news source, the boundaries between truth and fabrication blur, allowing deliberate agendas to infiltrate public discourse. "Fake News" is not merely a nuisance, but a potent tool wielded for political and financial gain, with implications of undermining democratic values [2].

At the core of this issue lies propaganda, an age-old tactic reinvented for the digital era. Through carefully crafted psychological manipulation and narrative, propaganda seeks to sway public opinion and solidify ideological allegiance [3]. Yet, its insidious nature lies in its disguise of objectivity, masking its true intent behind seemingly impartial messages [4].

To combat this rampant threat, we employed deep learning models to detect propaganda techniques within news articles. Leveraging the power of pre-trained language models like RoBERTa, we delve into intricate nuances of propaganda, aiming to unravel its subtle yet impactful presence [5]. Through a meticulous exploration of various methodologies, including fine-tuning pre-existing training models along with utilizing data augmentation, we strive to illuminate the intricate web of propaganda techniques.

This paper is structured as follows: Section II offers a comprehensive review of existing literature on propaganda detections, providing valuable insights into the current landscape. In section III, we talk about our methodology, detailing the diverse array of approaches employed in our pursuit of propaganda detection and the dataset utilized. Section IV presents our findings and engages in a discussion of the implications therein. Finally, Section V offers reflections on our discoveries and charts a course for future investigations.

II. Related Work

Various models have been developed in an attempt to identify propaganda techniques in news articles. In a study [6], three different parameters; text analytics to identify instances of propaganda, real-time user behavior after reading article to understand context [7], and authentic source indexing [8] were the primary basis for their methodology. This solution architecture produced a lower loss value of 0.5087 as well as an accuracy score of 0.7424.

In another work [9], a methodology including two subtasks of span identification and technique identification were utilized. Span identification is used for specific text fragments in articles that contain propaganda and maps them to a binary sequence, while text classification determines the propaganda techniques the text employs using 14 defined predefined techniques [10] for each specific text fragment. The architecture relied on pre-trained models for feature extraction and learning [11] and then sequential models created from pre-trained learned language learning model employ deep learning techniques for detection of propaganda techniques. Results of this experiment showed segment identification tended to be easier with most of the system using it showing improvements, while the more difficult task of technique classification limited many teams from being able to surpass the baseline performance. Another result of this was that the performance was greatest when teams utilized pre-trained transformer models (BERT) and similar deep learning techniques. The concluding statement of findings from the authors stated that the act of programming to detect classification techniques proved quite difficult

In a recent study [12], a dataset split into 371 training articles and 75 testing sets where fine-tuned versions of GPT-3 and GPT-4 models [13] were experimented with using various prompt engineering strategies to evaluate metrics such as the F1, Precision and Recall Score. Initially, the program finds various metrics as listed above using prompt engineering for OpenAI's GPT-X models for both training and testing evaluations. From here, the language learning model RoBERTa compares the performance of the GPT models being used to find which one 'best-fits' the analysis of propaganda detection [14]. The experiments run by the authors found that both the GPT-3 and GPT-4 models proved quite successful yet both also had their own setbacks. GPT-4 prompts led to much higher accuracy when detecting propaganda techniques in that it was able to successfully classify labels with judiciousness. On the other hand, the GPT-3 model was not as successful as results revealed that it had issues with, "repetitions, hallucinations and unanticipated terminations." All in all, results were hindered by both of the models' inability to accept any text (including lengthy datasets for training) past a preset token length which led to inconsistency with testing results.

A paper focusing on English news articles [15] had a methodology centered around the use of a RoBERTa model, utilizing a dataset from SemEval 2020 Task 11 that includes training, development, and test articles annotated with 14 different propaganda techniques. The solution architecture consisted of pre-processing which involves cleaning text and tokenizing using RoBERTa Tokenizer, adjusting input to fixed length for uniformity before training. The experiment achieved a significant improvement over baseline models with an F1 score of 60.2%, highlighting RoBERTa's effectiveness in understanding nuanced language use associated with propaganda.

Lastly, an article [16] that utilized BERT model to tackle span identification and technique classification in news articles. Approach is augmented by over-sampling and Exploratory Data Analysis (EDA) strategies to address dataset imbalances. The solution employed sentence-level feature concatenation to improve the detection accuracy. For technique classification, a dimensionality reduction layer is added before classification. The Result of the model was an outperformance of existing models in detecting propaganda techniques, achieving state of the art performance with an F1 score of 0.575729 due in large part to EDA.

III. METHODOLOGY

Our methodology for detecting propaganda techniques in news articles is encompassed by using techniques and strategies, namely data augmentation, ensemble learning for detection and feature extraction methodologies.

We initialize our methodology by deploying the Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) model [17] to detect propaganda techniques. Our attempt at utilizing and leveraging BERT's contextual understanding and ability to discern relationships in sentences in combination with RoBERTa's byte-level tokenization [18] gives us a model that achieves a high-level propaganda identification accuracy.

Feature extraction leverages RoBERTa's pre-trained layers and from within this, is able to successfully generate embeddings, which are vector representations, for input text from news articles. These embeddings capture contextual semantics of input tokens that have been generated from the given news articles allowing the RoBERTa model to comprehend and detect the language and propaganda techniques being utilized. Finally, the input text is encapsulated by the model's transformer architecture allowing it to preprocess data using tokenization and segment IDs. [19]

We adopted an ensemble learning approach through the utilization of XGBoost and linear regression models to further alleviate the performance of our detection model. These classifiers are trained independent of each other using different subsets of data and extractions of features. After their independent training, they are combined to improve the overall performance of the model's accuracy rate through leveraging their individual strengths which furthermore mitigates the possibility of any model bias.

Data augmentation through WordNet [20] was a technique used to enhance the diversity and accuracy of our model by incorporating synonyms into our dataset to enlarge our overall testing and training datasets. With this strategy, we aimed to expose the RoBERTa model to a larger dataset, more specifically, a dataset that has now doubled in size, to ensure the model had adequate data for text and language understanding.

Our RoBERTa model was extensively trained on the dataset from the "Hack the News Datathon Case for Propaganda Detection" [21]. The dataset is comprised of articles containing textual information that has been annotated with propaganda techniques to allow for learning through models as well as '.labels' files which are tab separated files consisting of columns headers: article number, propaganda techniques identified, starting character index in which the segment of text relating to the propaganda technique and the ending character of the segment of text relating to the propaganda technique, all of which links to a corresponding text file. The model training for this dataset involves using a suitable loss function and an optimization algorithm to enhance the usage of the RoBERTa model's input parameters.

IV. FINDINGS

Our RoBERTa model achieved an overall accuracy score of 66%, which was higher than 90% of all the articles we have reviewed. This significant number indicates the ability of our model to accurately detect and classify propaganda techniques used in news articles. The metrics used for our model's results and most of the models in the literature reviews are precision, recall, and the F1-score. [22]

Precision metric is used to measure the percentage of instances correctly predicted as positive for any of the given propaganda techniques such as loaded language and name calling [23]. Our RoBERTa model achieved a score of 66% for this metric showcasing an inductiveness of the model to accurately classify techniques.

Recall metric is used to measure the percentage of correctly identified instances belonging to a given class over another [24]. This metric is extremely similar to the precision metric in the sense that precision goes hand-in-hand with recall as they are capturing the percentage of similar instances, hence our recall score of 66%. This number shows that the model was able to correctly predict 66% of varying instances.

F1-Score is the mean of all sets of calculated results observed through our model's predictions, more specifically, it is the average of both precision and recall resulting in a balanced measure of the RoBERTa model's overall propaganda detection performance [25]. Our F1-score ranged from 13% to 76% showcasing that the model struggled classifying a small number of techniques but with an overall F1-score of 66%, the final version of the RoBERTa model proved to be very successful in the overall detection of propaganda techniques in news articles.

Our high resulting findings led by the effectiveness of RoBERTa's embeddings paired with ensemble learning with XGBoost implicate a successful overall model which was able to correctly identify propaganda techniques through linguistic patterns. However, we also have to note the disparity in ranges between different techniques such as repetition and straw man. Repetition was detected with a very high precision and recall of 0.69 and 0.86 indicating the model's success in using linguistic embeddings to detect repetition while straw man had a high precision as well but a low recall of only 0.07 indicative of much greater difficulty to distinguish this technique most likely due to the model's low support training for straw man.

Overall, our findings and results shed light on the model's high potential to effectively detect propaganda techniques in news articles [26]. While the numbers looked great on paper, there is still more room to increase this model's performance using additional techniques, different regression models and more support for all popular techniques and we as a team plan to use these strategies to further increase the detection rate and classification accuracy.

V. CONCLUSION

This paper presents a significant step forward in the battle against misinformation and propaganda in news articles. By utilizing advanced deep learning techniques, particularly the RoBERTa model, we have displayed the effectiveness of detecting various propaganda techniques with a remarkable F1 score of 60%. Through thorough analysis and experimentation through trial and error, we have highlighted the potential of machine learning models in texts indicative of propaganda. Our findings emphasize the importance of media literacy and technological solutions in combating the spread of bias and manipulative narratives [27]. While our model showed remarkable results, there remains ample room for improvement. Included in things that can be done to further improve are providing better support for detecting all propaganda techniques, utilizing a combination of deep learning methods, and training our datasets using more powerful methods. As we continue to refine our methodologies, we aim to further enhance the accuracy and effectiveness of propaganda detection in news articles, contributing to a more informed and discerning society [28].

REFERENCES

[1] A. Abedalla, A. Al-Sadi, and M. Abdullah, "A closer look at fake news detection: A deep learning perspective," in Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence, pp. 24– 28, 2019.

[2] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[3] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. "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, pages 2931–2937, Copenhagen, Denmark, September. Associationfor Computational Linguistics. 2017.

[4] M. Orlov and M. Litvak, "Using behavior and text analysis to detect propagandists and misinformers on twitter," in *Annual International Symposium on Information Management and Big Data*, pp. 67–74, Springer, 2018.

[5] Bias, symbolism, and Propaganda (no date) Education. Available at: https://education.nationalgeographic.org/resource/resource-library-bias-symbolism-and-propaganda

[6] Patil, Megharani, Hrishikesh Yadav, Mahendra Gawali, Jaya Suryawanshi, Jaikumar Patil, Anjali Yeole, Prathik Shetty, and Jayesh Potlabattini. "A Novel Approach to Fake News Detection Using Generative AI." *International Journal of Intelligent Systems and Applications in Engineering* 12, no. 4s (2024): 343-354.

[7] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, "SpotFake: A Multi-modal Framework for Fake News Detection," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 2019, pp. 39-47, doi: 10.1109/BigMM.2019.00-44.

[8] Edell, A. (2018) "Trained Fake News Detection AI with >95% Accuracy, and Almost Went Crazy." Towards Data Science

[9] Martino, G., Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. "SemEval-2020 task 11: Detection of propaganda techniques in news articles." *arXiv preprint arXiv:2009.02696* (2020).

[10] Rajaswa Patil, Somesh Singh, and Swati Agarwal. 2020. BPGC at SemEval-2020 Task 11: Propaganda detection in news articles with multi-granularity knowledge sharing and linguistic features based ensemble learning. In Proceedings of the 14th International Workshop on Semantic Evaluation, SemEval '20, Barcelona, Spain.

[11] Mayank Raj, Ajay Jaiswal, Rohit R.R, Ankita Gupta, Sudeep Sahoo, Vertika Srivastava, and Kim Yeon Hyang. 2020. Solomon at SemEval-2020 Task 11: Novel ensemble architecture for fine-tuned propaganda detection in news articles. In Proceedings of the 14th International Workshop on Semantic Evaluation, SemEval '20, Barcelona, Spain.

[12] Sprenkamp, K., Jones, D.G. and Zavolokina, L. (2023) *Large language models for propaganda detection*, *arXiv.org*. Available at: https://arxiv.org/abs/2310.06422

[13] Liu et al. (2023a) Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9):1–35.

[14] Da San Martino et al. (2019) Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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), pages 5636–5646.

[15] M. Abdullah, O. Altiti and R. Obiedat, "Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers," 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2022, pp. 301-308.,

[16] Li, W. et al. (2021) SPAN identification and technique classification of Propaganda in news articles - complex & intelligent systems, SpringerLink.Available: https://link.springer.com/article/10.1007/s40747-021-00393-y

[17] Y. Liu et al., "Roberta: A robustly optimized Bert pretraining approach," arXiv.org, https://arxiv.org/abs/1907.11692

[18] N. NLP, "Byte-pair Encoding Tokenization - hugging face NLP course," Byte-Pair Encoding tokenization - Hugging Face, https://huggingface.co/learn/nlp-course/en/chapter6/5

[19] D. Sharma, "A gentle introduction to Roberta," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2022/10/a-gentle-introduction-to-roberta/

[20] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," arXiv.org, https://arxiv.org/abs/1805.06201

[21] G. Preslav, "Hack the news Datathon case - propaganda detection," Data Science Society, https://www.datasciencesociety.net/hack-news-datathon-case-propaganda-detection/

[22] T. Tigerschiold, "What is accuracy, precision, recall and F1 score?," What is Accuracy, Precision, Recall and F1 Score?, https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score#:~:text=F1%20Score%20is%20a%20measure,models%20make%20that%20 trade%2Doff

[23] Lador, S.M. (2017) *What metrics should be used for evaluating a model on an imbalanced data set?*, *Medium*. Available at: https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

[24] N. Shazeer and M.Stern, "Adafactor: Adaptive learning rates with sub-linear memory cost," in International Conference on Machine Learning, pp. 4596-4604, PMLR, 2018

[25] F1 Score in Machine Learning: Intro & Calculation (no date) V7. Available at: https://www.v7labs.com/blog/f1-score-guide

[26] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning (2015). A large annotated corpus for learning natural language inference. In Emperical Methods in Natural Language Processing (EMNLP).

[27] Lord, K.M., Vogt, K. and Kristin M. Lord (@kristin_lord) is president and CEO of IREX (no date) Strengthn media literacy to win the fight against misinformation (SSIR), Stanford Social Innovation Review: Informing and Inspiring Leaders of Social Change. Available at: https://ssir.org/articles/entry/strengthen_media_literacy_to_win_the_fight_against_misinformation

[28] F.-J. Rodrigo, L.Plaza, "HModeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets." Available: https://ceur-ws.org/Vol-3496/dipromats-paper7.pdf