IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Propaganda Analyzer: A Transformer-Based Approach for Binary and Multiclass Detection of Propaganda in Text

Neha Hambir, Vismay Raut, Rahul Kulkarni, Anshul Patne, Dr. Anant Kaulage
MIT ADT University, Pune, Maharashtra, India
nehahambir1402@gmail.com, vismay.raut22@gmail.com
rahulkulkarni14204@gmail.com, anshulpatne26@gmail.com

Abstract: *In today's digital age, the quick rise of misinformation and biased stories has made propaganda detection an important focus in Natural Language Processing (NLP). This paper introduces Propaganda Analyzer, a transformer-based system aimed at automatically identifying and classifying propaganda in text. The system works in two stages. First, a binary classification model checks if a piece of text contains propaganda. Then, a multiclass classification model sorts the identified propaganda into specific techniques like name-calling, appeal to fear, or loaded language. The framework uses fragment-level analysis, dividing lengthy articles or statements into smaller, meaningful segments to improve understanding and accuracy. Pre-trained transformer models, including BERT and its variants, are fine-tuned for both classification tasks with annotated propaganda datasets. The experimental results show notable gains in accuracy and F1-score when compared to traditional machine learning methods. The study also reviews error cases in detail and discusses the ethical concerns surrounding automated propaganda detection in real-world situations. Overall, the Propaganda Analyzer offers an interpretable, scalable, and modular method to fight online propaganda and promote digital media awareness.*

Keywords: *Natural Language Processing*

I. INTRODUCTION

In recent years, the rapid growth of digital communication platforms has changed how information is shared and consumed. While this progress has made knowledge and diverse opinions widely accessible, it has also allowed propaganda, misinformation, and manipulative messages to spread quickly. These biased forms of communication often seek to sway public opinion, manipulate emotions, or push specific political or ideological views. As a result, detecting and understanding propaganda in written content has become an important area of study in Natural Language Processing (NLP) and computational social science.

Traditional methods for finding propaganda have relied on manually created language features, keyword-based filtering, or traditional machine learning algorithms like Support Vector Machines and Naïve Bayes. While these approaches provide a basic understanding of text characteristics, they often miss the deeper context and meaning found in complex political or persuasive writing. The arrival of transformer-based deep learning models, such as BERT and RoBERTa, has transformed NLP by allowing machines to grasp context, tone, and relationships within text using attention mechanisms. These models have shown great success in text classification, sentiment analysis, and tasks related to misinformation detection.

This research presents the Propaganda Analyzer, a transformer-based framework aimed at identifying and categorizing propaganda at both binary and multiclass levels. The system first checks if a text contains propaganda and then classifies the identified instances using predefined propaganda techniques. To improve detail, the framework takes a fragment-level analysis approach, breaking longer articles or statements into smaller, contextually relevant segments

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

before classification. This method boosts accuracy by enabling the model to concentrate on local context instead of entire documents. The implementation uses advanced transformer architectures, fine-tuned on publicly available propaganda datasets, and follows a modular design for flexibility and reproducibility.

The main goal of this work is to provide a strong, understandable, and efficient system for analyzing textual propaganda in real-world situations, such as news media, political statements, and social media posts. In addition to technical performance, this study focuses on ethical considerations related to bias, fairness, and the responsible use of tools for automated propaganda detection. The results show that Propaganda Analyzer not only achieves competitive accuracy but also offers a clear framework for enhancing digital media literacy and addressing information manipulation

II. LITERATURE SURVEY

TABLE 1. Literature Survey

Author(s) / Year	Title / Research Work	Techniques / Models Used	Dataset / Source	Limitations Identified
Rashkin et al., 2017	Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking	Linguistic feature extraction and logistic regression for deception and propaganda classification	BS Detector & Politifact datasets	Relied on shallow lexical cues; lacked contextual understanding of propaganda techniques
Da San Martino et al., 2019	Fine-Grained Analysis of Propaganda in News Articles	BERT and multi-granularity neural networks for detecting propaganda techniques	QProp dataset (News articles annotated for 18 propaganda techniques)	High computational cost and limited generalization to informal online text
Barrón-Cedeño et al., 2020	Proppy: Organizing the News Cycle Through Propaganda Detection	Multi-class text classification with TF-IDF and SVM	Proppy dataset (News articles & social media posts)	Feature-based approach struggled with implicit or context-dependent propaganda
Popat et al., 2018	DeClarE: Debunking Fake News and Claims Using Evidence-Aware Deep Learning	Bi-LSTM with attention for stance and credibility analysis	FEVER dataset	Focused on claim credibility, not explicit propaganda labeling
Alhindi et al., 2020	Fine-Tuned Transformers for Propaganda Detection	RoBERTa and DistilBERT fine-tuned for sentence-level classification	SemEval 2020 Task 11 dataset	Limited interpretability and sensitivity to noisy input data
Samadi et al., 2021	Detection of Propaganda in Persian News Using Deep Learning	CNN and LSTM hybrid model	Custom Persian propaganda dataset	Language-specific; lacked cross-lingual generalization
Da San Martino et al., 2020	SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles	Ensemble of transformer models with token-level classification	SemEval 2020 shared task dataset	Complex pipeline and imbalance between propaganda categories

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

III. PROBLEM STATEMENT

In today's digital world, information is created and shared at an incredible speed through online news sites, blogs, and social media. However, this open flow of information has also become a breeding ground for propaganda and misleading narratives that twist facts, play on emotions, and sway public opinion. Propaganda content is often subtle and mixed in with legitimate text, making it hard for readers and standard algorithms to spot.

Current propaganda detection systems focus mainly on surface-level language features or older machine learning models that find it tough to grasp deeper contextual connections and meanings. While recent studies have used transformer-based models like BERT and RoBERTa, these methods often look at text at the sentence or document level. This means they might miss how propaganda can show up in specific parts of a text. As a result, models may incorrectly label content that blends neutral and manipulative statements.

Thus, there is a strong need for a context-aware, detailed detection system that can identify both the existence and type of propaganda techniques in longer texts. The proposed system, Propaganda Analyzer, fills this gap by offering a fragment-level, two-stage transformer-based framework. It first classifies the text to detect propaganda and then categorizes the specific technique used. This approach aims to enhance detection accuracy, clarity, and flexibility across various types of digital content.

IV. PROPOSED SOLUTION

The proposed system, Propaganda Analyzer, is a transformer-based framework that automatically detects and classifies propaganda in text. The model works in two stages. First, a binary classifier identifies whether propaganda is present. Then, a multiclass classifier determines the specific propaganda technique used. To improve contextual accuracy, the system performs fragment-level analysis, breaking long texts into smaller segments for better detection. The framework uses fine-tuned transformer models like BERT and DistilBERT. These are integrated into a modular and scalable architecture for efficient processing and analysis.

WORKFLOW DIAGRAM

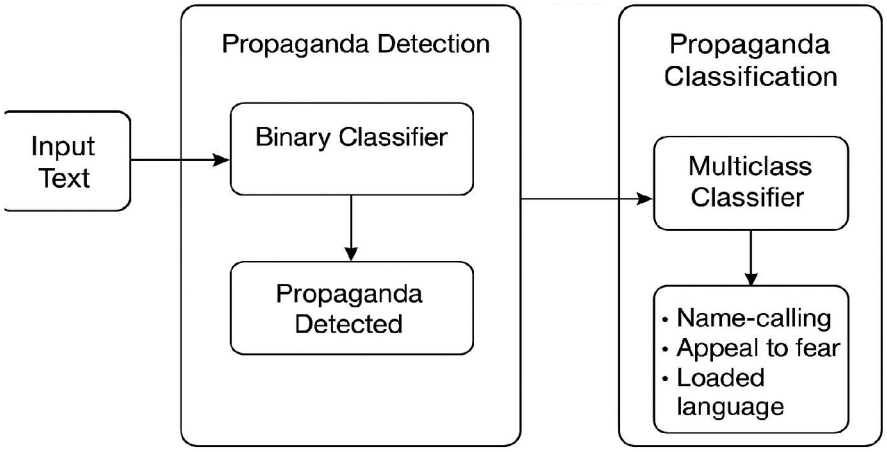

FIGURE 1. Workflow Diagram for Proposed Solution

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

BLOCK DIAGRAM

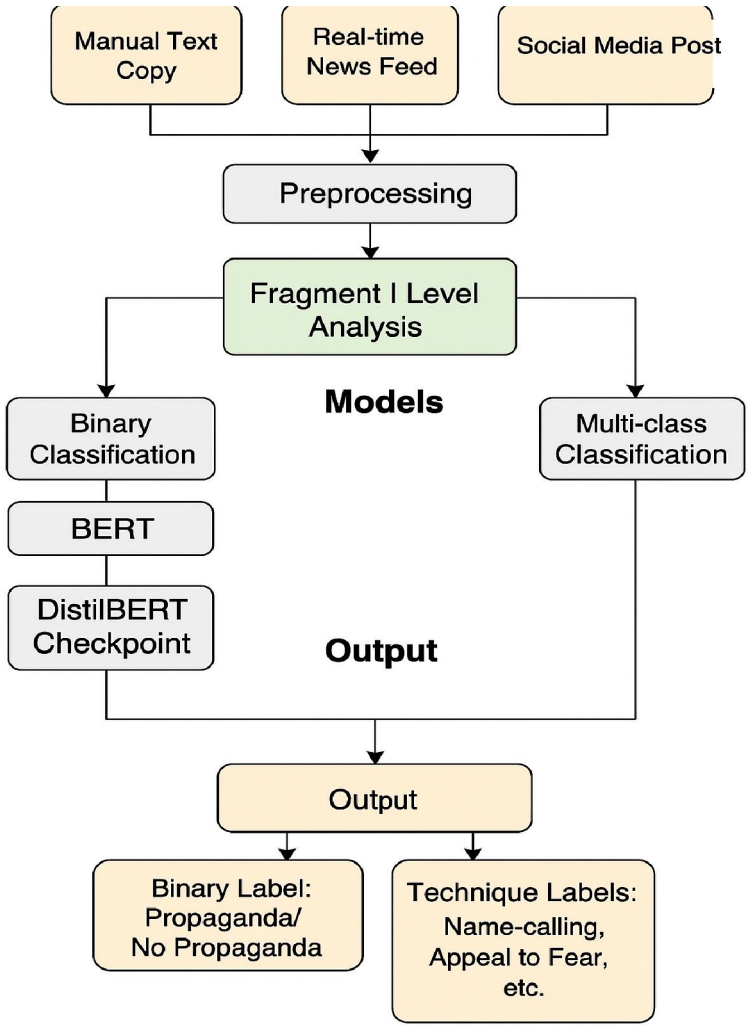

FIGURE 2. Block Diagram For Proposed Solution

PROPOSED METHODOLOGY SOLUTION

The proposed system, Propaganda Analyzer, aims to detect and classify propaganda in text using transformer-based deep learning models. The framework consists of several interconnected modules that work together to ensure accurate and fine-grained analysis of propaganda content. Each module performs a specific task as detailed below.

1. **Data Collection**

Objective: To gather a comprehensive and balanced dataset containing both propagandistic and non-propagandistic text samples suitable for training and evaluating the models.

Process: Collect text data from multiple benchmark sources such as SemEval 2020 Task 11, QProp, and news article repositories.

Include examples from both political and social domains to ensure coverage of diverse propaganda techniques.

Label data according to predefined propaganda categories such as name-calling, appeal to fear, and loaded language.

Organize and store the data in structured formats (CSV/JSON) for further processing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

2. Data Preprocessing

Objective: To clean and prepare the collected textual data for efficient feature extraction and model training.

Process: Remove irrelevant symbols, punctuation, and stop words from the text. Convert all text to lowercase and standardize whitespace. Apply tokenization using the BERT tokenizer to break sentences into sub word units. Ensure each sentence or fragment is appropriately formatted for input to transformer models. Perform initial exploratory analysis to identify class imbalance or noise in the dataset.

3. Text Fragmentation

Objective: To improve detection accuracy by analyzing text at the fragment level rather than at the full-document level.

Process: Segment long articles or paragraphs into smaller, semantically coherent fragments. Each fragment is treated as an independent sample for classification. This method allows the model to focus on localized propaganda cues that might otherwise be diluted in longer texts. Store the fragmented text samples along with their corresponding labels for downstream processing

4. Feature Extraction

Objective: To represent textual fragments in numerical form that captures contextual and semantic meaning for use by deep learning models.

Process: Utilize BERT-based embeddings (from BERT or DistilBERT) to convert each tokenized fragment into a dense feature vector. These embeddings capture both syntactic and semantic dependencies in the text. Use the [CLS] token representation from the final hidden layer as the sentence-level embedding for classification. Normalize feature vectors and prepare them for training the binary and multiclass classifiers.

5. Binary Classification

Objective: To determine whether a given text fragment contains propaganda or not.

Process: Fine-tune a transformer-based binary classifier (BERT or DistilBERT) on labeled data. The classifier outputs two labels: Propaganda or Non-Propaganda. Employ cross-entropy loss and AdamW optimizer during training for optimal convergence. Use evaluation metrics such as Accuracy, Precision, Recall, and F1-score to measure performance. Fragments identified as propaganda are passed to the next classification stage.

6. Multiclass Classification

Objective: To identify the specific propaganda technique present in the fragments flagged as propagandistic.

Process: Train a multiclass transformer model to classify propaganda into various categories such as Appeal to Fear, Name-Calling, Loaded Language, etc. Use softmax activation in the output layer to assign probabilities across multiple classes. Fine-tune the same pre-trained BERT architecture with multiclass labeled data. Evaluate the model using macro-F1 and per-class accuracy metrics. The results are stored along with the fragment IDs for report generation.

7. Output Visualization

Objective: To display the classification results in a structured and interpretable form for analysis and decision-making.

Process: Present the results through a visual dashboard or textual report summarizing predictions. Show both binary outcomes (Propaganda/Non-Propaganda) and multiclass labels (technique type). Include confidence scores and example text snippets to enhance interpretability. Provide insights into the frequency and distribution of propaganda techniques within a dataset or document.

V. FUTURE SCOPE AND BENEFITS

1. Value Proposition

- Automated Propaganda Detection: Provides an AI-driven system capable of automatically detecting and classifying propaganda within text-based content.
- Context-Aware Analysis: Uses fragment-level analysis to identify subtle and localized instances of manipulative language, improving detection accuracy.
- Dual-Stage Framework: Implements a two-step transformer-based approach — binary classification for propaganda detection and multiclass classification for identifying propaganda techniques.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

● Data-Driven Media Insights: Generates analytical insights about propaganda trends, frequency, and distribution across datasets or media platforms.
● Ethical and Educational Application: Supports digital literacy and responsible media consumption by helping users recognize bias and manipulation in textual content.

2. Customer Segments
● Media and Journalism Organizations: News agencies and fact-checking platforms that need tools to assess bias and propaganda in published content.
● Academic and Research Institutions: Universities and researchers studying misinformation, media literacy, or political communication.
● Government and Policy Makers: Authorities and regulators monitoring the spread of manipulative or politically biased information.
● Social Media Monitoring Agencies: Organizations tracking information integrity and public discourse on digital platforms.
● Educational Platforms: Institutions promoting critical thinking and awareness about misinformation in the digital age.

3. Sources of Income
● Subscription-Based Model: Monthly or yearly subscription plans for organizations using the propaganda detection tool for continuous monitoring.
● One-Time Licensing Fees: Fixed licensing cost for government or academic institutions adopting the software for long-term use.
● Pay-Per-Use Model: Charges based on the volume of text analyzed, suitable for researchers and small agencies.
● Analytics and Reporting Services: Premium analytical reports offering propaganda trend insights for media houses and policymakers.
● Custom Integrations: Additional charges for integrating the system with enterprise platforms, content management systems, or APIs.

4. Channels
● Direct Sales: Dedicated outreach to media organizations, research institutions, and government agencies.
● Strategic Partnerships: Collaboration with AI research labs, media monitoring firms, and NLP-based analytics companies.
● Academic and Technology Conferences: Demonstrations and presentations at NLP, AI ethics, and digital communication conferences.
● Online Platforms: Promotion through webinars, academic publications, and digital marketing campaigns.
● Open-Source Community Engagement: Partial release of datasets or models to build credibility and foster adoption in research communities.

5. Relationships with Customers
● Dedicated Support Services: Technical and model assistance for institutions using the platform at scale.
● Training and Onboarding Programs: Comprehensive training sessions to help users understand model interpretation and propaganda classification.
● Collaborative Research Opportunities: Joint projects with universities or research labs for model improvement and dataset expansion.
● Continuous Updates: Regular release of updated models and datasets to improve accuracy and performance.
● Community Feedback Mechanism: Incorporation of user feedback to refine model predictions and improve usability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

6. Cost Structure

- Research and Development: Major investment in model fine-tuning, dataset curation, and transformer optimization.
- Infrastructure Costs: Expenses for GPU-based computation, cloud storage, and data processing pipelines.
- Marketing and Sales: Costs associated with outreach, digital marketing, partnerships, and public demonstrations.
- Legal and Compliance: Ensuring data privacy, ethical usage, and compliance with AI transparency standards.
- Maintenance and Support: Continuous model retraining, bug fixes, and user support infrastructure.

VI. CONCLUSION

The proposed system, Propaganda Analyzer, offers an effective and automated way to detect and classify propaganda in written content. By using transformer-based models like BERT and DistilBERT, the framework establishes a strong dual-stage classification process that includes both binary and multiclass detection. The fragment-level analysis improves contextual understanding. This allows the system to spot subtle and localized propaganda cues that traditional methods often miss.

This work adds to the fields of Natural Language Processing (NLP) and digital media analysis. It offers a clear and scalable system that helps identify manipulative communication patterns. The model's implementation can aid researchers, media organizations, and policymakers in promoting responsible information sharing and digital media literacy.

Future developments will focus on enhancing multilingual support, integrating explainable AI methods for better understanding, and creating an easy-to-use interface for real-world use. Overall, Propaganda Analyzer shows how deep learning and language analysis can work together to fight misinformation and build trust in digital communication systems.

REFERENCES

1. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P., "Fine-Grained Analysis of Propaganda in News Articles," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 5636–5646.

2. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., & Choi, Y., "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2931–2937.

3. Barrón-Cedeño, A., Da San Martino, G., Jaradat, I., & Nakov, P., "Proppy: Organizing the News Cycle through Propaganda Detection," Proceedings of the 2019 International AAAI Conference on Web and Social Media (ICWSM), 2019, pp. 531–540.

4. Popat, K., Mukherjee, S., Yates, A., & Weikum, G., "DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 22–32.

5. Alhindi, T., Petridis, S., & Elmadany, A., "Fine-Tuned Transformers for Propaganda Detection," Proceedings of the SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, 2020, pp. 161–170.

6. Raj, A., & Ghosh, D., "Transformer-Based Multiclass Detection of Propaganda," International Journal of Artificial Intelligence Research, vol. 6, no. 2, 2022, pp. 45–53.

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., "Attention Is All You Need," Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), 2017,

8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

9. Da San Martino, G., Barrón-Cedeño, A., Jaradat, I., & Nakov, P., "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles," Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval), 2020, pp. 1377–1414.