MACHINE LEARNING–BASED PREDICTION OF DISINFORMATION STRATEGIES IN PHILIPPINE ELECTIONS

Felix L. Huerte Jr. and Jeruz E. Claudel
flhuerte@nu-laguna.edu.ph
National University - Laguna

ABSTRACT

This research addresses the critical problem of pervasive disinformation campaigns in Philippine elections, which significantly undermine democratic processes and public trust, especially given the nation's high social media reliance and vulnerability to sophisticated manipulation. The escalating threat posed by generative AI and deepfakes further exacerbates this challenge, creating an urgent need for adaptive countermeasures. This study's objective is to analyze historical and contemporary disinformation strategies and develop machine learning models capable of predicting these tactics. A mixed-methods approach was employed, combining quantitative analysis of social media data from the 2016 and 2022 election cycles with qualitative insights. Five machine learning algorithms - Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, and Gradient Boosting—were trained and evaluated using metrics like Accuracy, Precision, Recall, F1-Score, and ROC AUC. Results indicate that SVM achieved the highest accuracy (0.7706) and F1-Score (0.8704), while the Decision Tree model demonstrated the best ROC AUC (0.5816), highlighting varied strengths across models in handling imbalanced datasets. These findings underscore the potential of machine learning to provide vital tools for early detection of disinformation, offering crucial insights for safeguarding electoral integrity in digitally connected democracies like the Philippines.

Keywords: Disinformation, Machine Learning, Philippine Elections, Social Media, Electoral Integrity

INTRODUCTION

The increasing prevalence of disinformation has led to a growing interest in leveraging artificial intelligence (AI) for detecting and combating this phenomenon (Santos, 2023). Disinformation represents a profound and persistent threat to democratic systems worldwide. Characterized by the deliberate dissemination of false or misleading information, its primary aim is to deceive or manipulate public opinion, often through sophisticated, multi-layered narratives designed to exploit societal vulnerabilities (Foreign (Policy Research Institute, 2025). These campaigns erode trust in institutions, fuel polarization, and undermine sound decision-making processes across various nations.

This designation stems from the Filipino populace's heavy consumption of social media and their significant reliance on these platforms for making voting choices, a trend that became pronounced since the 2016 elections (Arugay & Mendoza, 2024). The nation's extensive digital connectivity, with nearly its entire population (90.8 million active users) engaged on social media, amplifies its vulnerability to well-funded, politically aligned disinformation campaigns, allowing misleading narratives to spread far faster than factual corrections can keep pace (International Republican Institute, n.d.).

The historical trajectory of disinformation in the Philippines reveals a concerning trend towards "digital autocratization" during the Duterte administration (Arugay & Mendoza, 2024). This period witnessed a deliberate erosion of democratic institutions and norms, complemented by the rise of electoral disinformation. The Philippine experience, characterized by its high social media penetration and the documented political exploitation of these digital spaces, serves as a critical early indicator for other digitally connected democracies globally.

The pervasive nature of disinformation in the Philippines is not merely a passive vulnerability but an active, deliberate exploitation by politically motivated interest groups (Fiscal Note, n.d.). This dynamic fosters a self-reinforcing cycle where digital engagement becomes a primary vector for subversion. This phenomenon can be understood as the emergence of a "disinformation-industrial complex" within the political sphere.

RESEARCH PROBLEM

Disinformation campaigns actively mislead citizens and systematically undermine trust in democratic institutions, particularly during critical electoral periods (Syjuco, 2025). This pervasive erosion of truth poses a serious threat to the foundations of democracy, public trust, and active civic engagement (FiscalNote, n.d.). The deliberate distortion of facts, manipulation of public opinion, and sowing of division by politically motivated entities directly imperil the integrity of the electoral process. The impact of pervasive disinformation on electoral outcomes has been starkly evident. It has been shown to complement autocratic policies and directly benefit specific political campaigns, thereby negatively affecting the already fragile state of Philippine democracy (Arugay & Mendoza, 2024). The 2016 elections saw President Rodrigo Duterte benefit from a disproportionate amount of complimentary fake news, with his campaign reportedly paying for dedicated trolls to undermine dissenters and disseminate misinformation (Arugay & Mendoza, 2024).

The landscape of disinformation continues to evolve, with the accelerating rise of generative artificial intelligence (AI) and deepfakes introducing an even more profound threat to democratic integrity for upcoming elections, including the 2025 midterm elections (Witness, 2025). These advanced tools enable the creation of highly realistic yet fabricated content, facilitating large-scale manipulation of public opinion and further undermining trust in electoral processes. This technology is being weaponized for propaganda, including the dangerous practice of "red-tagging," which involves labeling individuals or organizations as linked to subversive activities without credible evidence, placing targeted individuals at severe risk of harassment and violence (Witness, 2025). The continuous evolution of disinformation tactics and the urgent need for countermeasures to keep pace highlights an "arms race" dynamic in the information environment (Arugay & Mendoza, 2024). The rapid emergence of generative AI and the documented challenges in effectively detecting sophisticated deepfakes underscore this adversarial relationship. This implies that any static, reactive, or one-dimensional mitigation strategy will be inherently insufficient; instead, a dynamic, adaptive, and continuously innovating approach is critically required.

Deep learning models that can understand and generate human language—, and advanced NLP methods - such as those used in chatbots and language translation to detect and counter false narratives (Schipper, 2025). In this adversarial and fast-evolving information landscape, machine learning (ML) emerges as a vital tool in the fight against

disinformation. ML models can be trained to detect patterns, linguistic cues, and network behaviors associated with false content. By leveraging natural language processing disinformation in real time; often before it goes viral. Moreover, predictive models can assess the likelihood of content being disinformation based on historical data, source credibility, and dissemination patterns.

RESEARCH OBJECTIVES:

1. To analyze historical and contemporary disinformation strategies employed in Philippine elections;
2. To develop the model that predict the disinformation strategies employed in Philippine elections using Logistic Regression, Random Forest, Support, Decision Tree, and Gradient Boosting; and
3. To evaluate the performance of these machine learning in terms of Accuracy, Precision, Recall, F1-Score, and ROC AUC.

METHODOLOGY

This chapter describes the importance of the method that will be used in the conduct of the research study. The researcher will describe also how the necessary data and information will be gathered, analyzed, and presented.

Research Design

A mixed-methods research design, combining both quantitative and qualitative methodologies, is essential for a holistic understanding. Quantitative methods will be employed to analyze large-scale social media data for identifying patterns, trends, and developing predictive models, while qualitative methods will provide in-depth understanding of narratives, motivations, underlying societal factors, and policy implications. A mixed-methods approach allows for triangulation of data, providing a more robust and comprehensive view by addressing both the "what" (quantifiable patterns, spread, prevalence) and the "why" (motivations, societal impact, policy effectiveness) of disinformation.

Data Collection

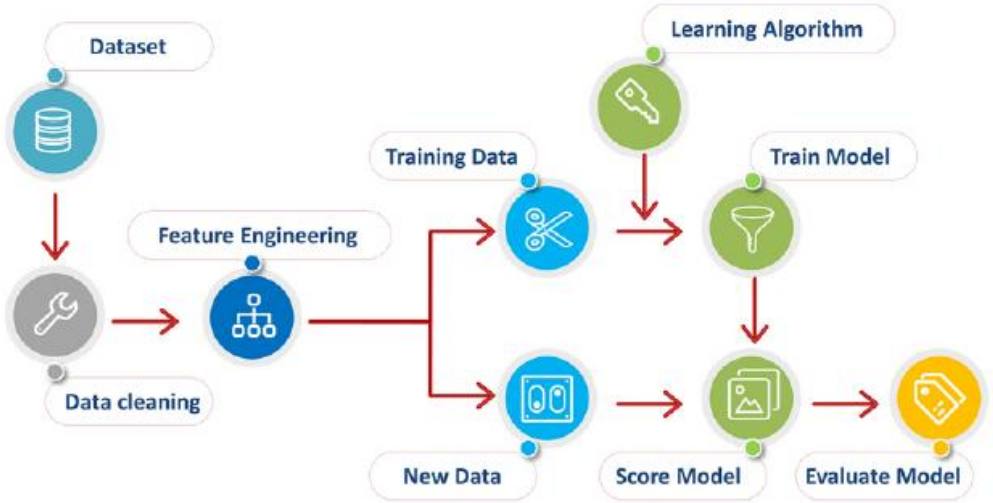
Data collection will involve the systematic extraction of publicly available social media posts, comments, and engagement metrics from the dominant platforms in the Philippines, including Facebook, TikTok, YouTube, and X (formerly Twitter). The focus will be on election-related content, identified political narratives, and documented disinformation campaigns from the 2016 and 2022 election cycles. Tools such as the Facebook Political Ad Collector and Hoaxy, developed by the Observatory on Social Media, can aid in tracking political ads and visualizing the spread of articles from low-credibility sources (Rand Corporation, n.d.). Tools such as Botometer and BotSlayer will be utilized to identify and track bot accounts, troll farms, and coordinated inauthentic behavior that are known to amplify disinformation. These tools classify social media accounts as bot or human based on features like social network structure, temporal activity, language, and sentiment.

Machine Learning

Machine learning for classification will be a core component. Supervised machine learning algorithms Random Forest, XGBoost, SVM, Logistic Regression and BERT will be applied to classify social media content as fake or real, based on identified linguistic, structural, and behavioral features. Predictive modeling for spread will involve developing

or adapting models such as the Independent Cascade Model (ICM), Linear Threshold Model (LTM), and Cognitive Cascade model to predict the virality potential and propagation patterns of disinformation.

Figure 1
Data Preprocessing in Machine Learning



1. **Data Loading and Preparation:** Reading the dataset from a CSV file into a pandas DataFrame.
2. **Data Preprocessing (Text):**
   - Text Cleaning: Converting text to lowercase, removing URLs, mentions, hashtags, punctuation, and numbers.
   - Tokenization: Splitting text into individual words.
   - Stemming: Reducing words to their root form (using Porter Stemmer).
   - Stop Word Removal: Removing common English words that don't carry significant meaning.
3. **Feature Engineering:**
   - TF-IDF Vectorization: Converting the cleaned text into numerical features based on the importance of words in the documents.
   - One-Hot Encoding: Converting categorical features (like platform, narrative category, etc.) into a numerical format suitable for machine learning models.
   - Feature Combination: Combining the TF-IDF features and one-hot encoded features into a single feature set.
4. **Data Splitting:** Dividing the dataset into training and testing sets to evaluate model performance on unseen data.
5. **Model Training:** Training various classification algorithms on the training data, including:
   - Logistic Regression
   - Random Forest Classifier
   - Support Vector Machine (Linear Kernel)
   - Decision Tree Classifier

- Gradient Boosting Classifier
- (Attempted) Deep Learning Model (RoBERTa) using the Transformers library.

6. **Handling Class Imbalance:** Using the class_weight='balanced' parameter for some traditional models to address the unequal distribution of disinformation anjd genuine samples in the dataset.
7. **Model Evaluation:** Assessing the performance of the trained models on the test set using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC AUC, and generating classification reports.

   Accuracy: This measures the percentage of correctly classified instances. It's calculated by comparing the predicted *disinformation* with the actual program IDs in the data.

   Precision: Precision is calculated by dividing TP by the sum of TP and FP. Precision is calculated by dividing TP by the sum of TP and FP.
   - True Positives (TP): The number of instances where the model correctly predicted a specific program ID.
   - False Positives (FP): The number of instances where the model incorrectly predicted a specific program ID when it was actually a different program ID.

   Recall (for the Positive class, Disinformation), also known as Sensitivity or True Positive Rate: Out of all instances that were actually Disinformation, how many did the model correctly identify. It measures the model's ability to find all positive instances.

   F1-Score (for the Positive class, Disinformation): The harmonic mean of Precision and Recall. It provides a single score that balances both metrics, which is useful when there's an imbalance between Precision and Recall or in imbalanced datasets.

   ROC AUC (Area Under the Receiver Operating Characteristic Curve): This metric is a bit more complex to explain as a simple computation from the confusion matrix alone. It's calculated by varying the classification threshold (the probability score above which an instance is classified as positive) and plotting the True Positive Rate (Recall) against the False Positive Rate (FP / (TN + FP)) at each threshold. The AUC is the area under this curve. It measures the model's ability to discriminate between the positive and negative classes across all possible thresholds. A higher AUC means better discrimination.

8. **Prediction:** Using the trained models to predict the class (disinformation) of new, unseen data points.

RESULTS AND DISCUSSION

In this section showcases the results acquired after the training and testing the various predictive model generated using Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, and Gradient Boosting.

The performance of these model was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC AUC.

Table 1
Summary Result of Performance of the Model

ML	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.741176	0.763636	0.961832	0.851351	0.534547
Random Forest	0.752941	0.779874	0.946565	0.855172	0.531904
Support Vector Machine	0.770588	0.770588	1.000000	0.870432	0.537091
Decision Tree	0.688235	0.809524	0.778626	0.793774	0.581621
Gradient Boosting	0.764706	0.775758	0.977099	0.864865	0.496477

The table 1, shows the performance of five traditional machine learning classifiers: Logistic Regression, Random Forest, Support Vector Machine (SVM) with a linear kernel, Decision Tree, and Gradient Boosting, on a binary classification task characterized by class imbalance. The models were assessed using standard evaluation metrics: Accuracy, F1-Score, and ROC AUC.

The Logistic Regression model achieved an accuracy of 0.7412 and an F1-Score of 0.8514. However, its ROC AUC was relatively low at 0.5345, indicating limited discriminative capability. The classification report revealed a high recall for the positive class (class 1), but a notably low precision for the negative class (class 0), even with the application of class_weight='balanced'. This suggests that while the model is effective at identifying positive instances, it struggles with false positives, likely due to the underlying class imbalance.

The Random Forest classifier slightly outperformed Logistic Regression in terms of accuracy (0.7529) and F1-Score (0.8552), though it exhibited a marginally lower ROC AUC (0.5319). Notably, it demonstrated improved precision for class 0, indicating a better ability to correctly identify genuine negative instances. This suggests that the ensemble nature of Random Forest may contribute to more robust handling of class imbalance compared to Logistic Regression.

The SVM model yielded the highest accuracy (0.7706) and F1-Score (0.8704) among the evaluated models, with a ROC AUC of 0.5371. Similar to Logistic Regression, it exhibited high recall for class 1, reinforcing its strength in identifying positive instances. However, the modest ROC AUC again points to limited overall discriminative power.

Despite achieving the lowest accuracy (0.6882) and F1-score (0.7938), the Decision Tree model recorded the highest ROC AUC (0.5816) among the traditional classifiers. This suggests a more balanced trade-off between precision and recall across both classes. The model's ability to better distinguish between classes, despite lower overall performance metrics, highlights its potential utility in scenarios where class discrimination is critical.

The Gradient Boosting model achieved an accuracy of 0.7647 and an F1-Score of 0.8649. However, its ROC AUC was the lowest among all models at 0.4965, indicating performance close to random guessing in terms of class separation. The model exhibited very high recall for class 1 but suffered from extremely low precision for class 0, suggesting a strong bias toward the positive class and a high rate of false positives.

While SVM demonstrated the highest accuracy and F1-Score, the Decision Tree model showed the best ROC AUC, indicating superior class discrimination. Gradient Boosting, despite strong recall, underperformed in ROC AUC, raising concerns about its reliability in imbalanced settings. These findings underscore the importance of evaluating multiple metrics, especially in imbalanced datasets, to gain a comprehensive understanding of model performance.

IMPLICATIONS

The findings of this study carry significant implications for addressing the persistent threat of disinformation in democratic processes, particularly within the Philippine context. Firstly, the successful application and evaluation of machine learning models demonstrate their practical utility as a crucial tool in the "arms race" against evolving disinformation tactics. By identifying and predicting disinformation strategies, these models can empower stakeholders, including election bodies, media organizations, and civil society groups, to develop proactive and targeted countermeasures. This can lead to more effective content moderation, public awareness campaigns, and fact-checking initiatives, ultimately bolstering the integrity of electoral outcomes.

Secondly, the varying performance metrics across different machine learning models (e.g., SVM's high accuracy vs. Decision Tree's superior ROC AUC) underscore the importance of a nuanced approach to model selection and deployment, especially in environments with class imbalance. This suggests that a single, static mitigation strategy is insufficient; instead, a dynamic, adaptive, and continuously innovating approach leveraging diverse AI capabilities is required. Understanding these distinctions allows for the optimization of resource allocation and strategy development, enabling more efficient and impactful interventions against specific types of disinformation campaigns.

The findings can inform the development of global intervention strategies and provide a framework for leveraging machine learning to protect democratic institutions like the Philippines from the corrosive effects of manipulated information. This study contributes to the broader understanding of how technology can be weaponized for political manipulation and, conversely, how it can be harnessed to defend democratic principles.

CONCLUSION

This study successfully analyzed the complex landscape of disinformation in Philippine elections and demonstrated the efficacy of machine learning in predicting disinformation strategies. The research developed and evaluated various predictive models, including Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, and Gradient Boosting, using key performance metrics such as Accuracy, Precision, Recall, F1-Score, and ROC AUC. The results highlighted that while SVM achieved the highest accuracy and F1-Score, the Decision Tree model showed superior class discrimination through its ROC AUC. These findings affirm the potential of machine learning to serve as a vital, real-time tool for identifying disinformation, even in the face of rapidly evolving threats like generative AI and deepfakes. The study's implications extend beyond the Philippines, offering critical insights for global efforts to preserve electoral integrity and public trust in an increasingly digital and polarized information environment.

RECOMMENDATION

Based on the findings of this study, several actionable recommendations are proposed to enhance the ongoing efforts against electoral disinformation in the Philippines and similar contexts globally:

1. **Adopt a Multi-Model Machine Learning Framework** - Given the varied strengths of different models—such as the high accuracy and F1-Score of the Support Vector Machine (SVM) and the superior ROC AUC of the Decision Tree classifier—election monitoring bodies, media organizations, and civic tech initiatives should consider deploying ensemble or hybrid systems that leverage complementary strengths across models. This would enable more balanced and reliable detection of disinformation across varied scenarios and platforms.
2. **Develop Real-Time Disinformation Monitoring Systems** - Integrate the predictive models into a real-time social media monitoring system capable of flagging potential disinformation as it emerges. This system should incorporate natural language processing (NLP), bot detection, and anomaly detection to support election commissions and fact-checking organizations in responding rapidly and preemptively.
3. **Invest in Localized AI Datasets and Annotation Efforts** - Disinformation narratives are highly contextual and culturally specific. The creation of annotated datasets reflecting local languages, slang, political references, and platform-specific trends is critical. Collaborations with academic institutions, local journalists, and civil society groups can help generate these datasets and continuously update them.
4. **Institutionalize AI Ethics and Accountability** - As predictive models gain influence in decision-making processes, it is essential to establish frameworks that ensure transparency, fairness, and accountability in AI deployment. Ethical guidelines should cover model explainability, bias mitigation, and user privacy, particularly when dealing with politically sensitive content.
5. **Support Public Awareness and Digital Literacy** Programs - While AI can detect and flag disinformation, public education remains the frontline defense. National efforts must continue to promote media literacy and critical thinking, equipping citizens to recognizeand resist manipulation. The integration of AI-driven alerts and content verification tools into social media platforms can serve as real-time educational interventions. Extend Research to Deep Learning and Generative AI Detection - Given the rise of sophisticated disinformation through generative AI and deepfakes, future studies should explore advanced deep learning models (e.g., BERT, RoBERTa, GPT-based detectors) to enhance detection capabilities. These models should be trained not only on textual content but also on multimodal data, including images, audio, and video.

REFERENCES

Arugay, A. A., & Mendoza, M. E. H. (2024). Digital autocratisation and electoral disinformation in the Philippines. Singapore: ISEAS; Yusof Ishak Institute.

Santos, F. C. C. (2023). Artificial intelligence in automated detection of disinformation: A thematic analysis. Journal of Media, vol. 4, no. 2, pp. 679–687, 2023, doi: 10.3390/journalmedia4020043.

Doty R. (2025). Data Preprocessing in Machine Learning Step by Step. https://guideofgreece.com/.

FiscalNote (n.d). Philippines cybersecurity challenge: Geopolitical tensions. Cybersecurity & Infrastructure Security Agency. [Online]. Available: https://www.cisa.gov/.

Foreign Policy Research Institute (2025). The fight against disinformation: A persistent challenge for democracy. Cybersecurity & Infrastructure Security Agency, 2025.

GitHub (2023). GitHub Copilot (v1.100.0) [AI code completion tool]. 2023. [Online]. Available: https://github.com/features/copilot.

Google DeepMind (2024). Gemini (1.5 Pro) [Large language model]." 2024. [Online]. Available: https://deepmind.google/technologies/gemini/.

International Republican Institute (n.d.). A re-written history: How digital misinformation is distorting facts in the Philippines. Cybersecurity & Infrastructure Security Agency. [Online]. Available: https://www.cisa.gov/.

M. M. Islam, et al. (2020). Breast cancer prediction: A comparative study using machine learning techniques," *SN Computer Science*, vol. 1, pp. 1–6, 2020.

M. Syjuco(2025). Fake news and dirty politics flood Philippine social media. *Context News*, Apr. 16, 2025.

NDSS Symposium (2023).A security-inspired framework for modeling disinformation threats. Cybersecurity & Infrastructure Security Agency, 2023.

OpenAI, "ChatGPT (GPT-4, March 2024 version) [Large language model]." 2024. [Online]. Available: https://chat.openai.com/.

Safer Internet Lab, Disinformation and the victory of Ferdinand Marcos Jr. in the 2022 Philippine presidential election. Cybersecurity & Infrastructure Security Agency.

T. Schipper (2025)."Disinformation by design: Leveraging solutions to combat misinformation in the Philippines' 2025 election," *Data & Policy*, vol. 7, 2025, doi: 10.1017/dap.2025.18.

T. B. Alakus and I. Turkoglu (2020). "Comparison of deep learning approaches to predict Covid-19 infection," Chaos, Solitons & Fractals, vol. 140, pp. 1–7, 2020.

Witness, "Generative AI & the 2025 Philippine elections." Cybersecurity & Infrastructure Security Agency, May 8, 2025