Modelling Online Misinformation Susceptibility Using Hybrid AI–Statistical Techniques

Ms. Rittika Bhattacharya[1], Dr. Joydeb Patra[2], Dr. Sthitadhi Das[3], Dr. Bikramaditya Naskar[4], Bandana Sinha[5*]

[1] Assistant Professor, Department of Mathematics, Brainware University, Kolkata, West Bengal, India.
Email: bhattacharyarittika@gmail.com
ORCID: 0009-0004-9130-9368

[2] Assistant Professor, School of Law, Brainware University, Kolkata, West Bengal, India.
Email: joydebajoy@gmail.com
ORCID: 0000-0001-7797-6309

[3] Assistant Professor, School of Computational and Applied Sciences, Brainware University, India.
Email: sthdas999@gmail.com
ORCID: 0009-0001-2484-9530

[4] Assistant Professor, Department of Mathematics, Brainware University, Kolkata, West Bengal, India.
Email: bikramadityaix@gmail.com

[5] Assistant Professor, Department of Commerce, Scottish Church College, Kolkata, West Bengal, India.
Email: sinhabandana21@yahoo.com

Corresponding Author:

Dr. Joydeb Patra[2*]
Assistant Professor, School of Law, Brainware University, Kolkata, West Bengal, India.
Email: joydebajoy@gmail.com
ORCID: 0000-0001-7797-6309

Abstract: The rapid expansion of social media platforms has significantly increased the spread of misinformation, influencing public opinion, political behaviour, and social decision-making. Identifying individuals who are more susceptible to online misinformation has therefore become an important problem in computational social science. This study proposes a hybrid AI–statistical framework for modelling misinformation susceptibility using logistic regression integrated with artificial intelligence-based analytical techniques. Socio-demographic, behavioural, and psychological variables are combined with AI-derived indicators obtained through natural language processing and user interaction analysis. The proposed methodology employs logistic regression as an interpretable statistical foundation while incorporating machine learning-assisted feature extraction and variable selection to improve predictive performance. Explainable AI tools are further utilised to assess variable importance and model transparency. The framework enables both accurate classification and interpretable inference regarding factors associated with misinformation vulnerability. Simulation studies and empirical analysis using social media

datasets demonstrate that the hybrid approach outperforms conventional logistic regression models in predictive accuracy and robustness. The proposed framework contributes to computational social science by combining statistical interpretability with modern AI-enhanced learning techniques for understanding online misinformation behaviour.;

Keywords: Computational social science; Logistic regression; Explainable AI; Misinformation detection; Social media analytics; Hybrid AI–statistical modelling; Natural language processing; Binary classification; Digital behaviour analysis.

Introduction

The widespread use of social media platforms has transformed the way information is generated, distributed, and consumed across societies [5, 1]. Although digital communication technologies have improved information accessibility, they have also accelerated the dissemination of misinformation and fake news [5]. Online misinformation has influenced political polarization, public health responses, social trust, and electoral outcomes in many countries.

Understanding why certain individuals are more susceptible to misinformation has therefore become an important interdisciplinary research problem involving statistics, artificial intelligence, computational social science, and behavioral studies [5, 3]. Traditional statistical methods such as logistic regression provide interpretable inferential frameworks for analyzing binary outcomes [2], while modern artificial intelligence techniques enable extraction of complex latent behavioral patterns from large-scale digital data [1, 3].

Recent developments in explainable artificial intelligence (XAI) have further emphasized the importance of transparent predictive systems capable of balancing interpretability and predictive accuracy [4]. Motivated by these developments, the present study proposes a hybrid AI–statistical framework combining logistic regression with AI-assisted feature engineering and explainability tools for modeling misinformation susceptibility.

The main contributions of this paper include the development of an interpretable hybrid AI–statistical modeling framework for misinformation susceptibility analysis, integration of NLP-derived behavioral indicators into logistic regression models, incorporation of AI-assisted feature extraction and variable selection procedures, explainable analysis of influential misinformation vulnerability factors, and comparative evaluation of classical and hybrid predictive modeling techniques.

Literature Review

Research on misinformation detection has expanded rapidly due to the increasing societal impact of fake news, digital propaganda, and misleading online content disseminated through social media platforms [5]. The growing accessibility of online communication technologies has intensified concerns regarding information credibility, manipulation of public opinion, and the spread of politically or socially polarized narratives. Consequently, misinformation analysis has emerged as an important interdisciplinary research area involving computational social science, machine learning, artificial intelligence, behavioral analytics, and statistical inference.

Early studies on misinformation detection primarily focused on content-based classification approaches using traditional machine learning algorithms such as support vector machines, naive Bayes classifiers, decision trees, and random forests [1, 3]. These approaches typically relied on handcrafted textual features, linguistic markers, lexical diversity measures, and metadata-based indicators to distinguish fake news from reliable information sources. Although such methods demonstrated moderate predictive performance, they often suffered from limitations associated with feature engineering complexity and insufficient representation of latent semantic structures in textual data.

Recent advances in artificial intelligence and natural language processing have significantly improved misinformation detection methodologies. Deep learning architectures, transformer-based language models, attention mechanisms, and contextual embedding frameworks have enabled

extraction of highly complex semantic and behavioral information from large-scale social media datasets [3]. Sentiment analysis, emotional polarity detection, user interaction modeling, and behavioral clustering techniques have further contributed to improved prediction and classification accuracy in misinformation-related applications. These AI-driven methodologies are increasingly capable of identifying subtle patterns associated with misinformation propagation, online engagement behavior, and digital influence mechanisms.

Despite these advancements, many modern AI-based predictive systems suffer from limited interpretability and lack transparent inferential explanations regarding influential variables contributing to misinformation susceptibility [4]. Black-box learning models often provide high predictive performance while failing to offer meaningful explanations necessary for policy analysis, behavioral interpretation, and computational social science research. This limitation has motivated increasing interest in explainable artificial intelligence (XAI), where interpretable analytical frameworks are developed to improve transparency, fairness, and reliability of AI-assisted decision systems.

In contrast to purely black-box machine learning approaches, logistic regression continues to remain one of the most widely used statistical techniques in social science and behavioral research because of its interpretability, inferential clarity, and probabilistic modeling structure [2]. Logistic regression models enable direct interpretation of predictor effects through odds ratios and regression coefficients, making them particularly suitable for binary classification problems involving human behavior, political participation, social trust, and misinformation vulnerability. Furthermore, penalized and regularized extensions of logistic regression have improved its applicability in high-dimensional settings involving large-scale digital covariates.

More recently, hybrid frameworks integrating machine learning algorithms with classical statistical methodologies have received considerable attention as an effective strategy for balancing predictive accuracy and interpretability [1, 3]. Such hybrid approaches combine the flexibility of AI-based feature extraction with the inferential strengths of statistical modeling procedures. In computational social science, these methods have shown promise for modeling online behavioral patterns, user engagement dynamics, and social media interaction structures while maintaining transparent inferential interpretation.

Although substantial progress has been achieved in misinformation detection and AI-assisted predictive analytics, relatively limited work has focused specifically on combining explainable AI techniques with logistic regression for modeling misinformation susceptibility at the individual level. Existing studies frequently emphasize content classification rather than interpretable behavioral susceptibility analysis. Moreover, limited attention has been given to integrating NLP-derived indicators, AI-assisted feature engineering, and explainable statistical modeling within a unified computational framework.

Motivated by these research gaps, the present study proposes a hybrid AI–statistical methodology that combines logistic regression, explainable AI tools, and AI-based feature extraction procedures for analyzing online misinformation susceptibility. The proposed framework aims to provide both strong predictive performance and interpretable statistical inference, thereby contributing to the growing literature on computational social science and AI-enhanced behavioral analytics.

Methodology

3.1 Data Structure

Consider a sample of n social media users with binary response variable

$$Y_i \in \{0,1\},$$

where $Y_i = 1$ denotes misinformation susceptibility for the ith individual. Let

$$X_i = (X_{i1}, X_{i2}, \dots, X_{ik})^\top$$

represent the associated covariate vector containing socio-demographic attributes, engagement

measures, psychological indicators, NLP-derived sentiment scores, and AI-generated behavioral features.

3.2 Hybrid Logistic Regression Framework

Define

$$p_i = P(Y_i = 1 \mid X_i).$$

The proposed framework models the conditional probability through the logistic specification

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij},$$

yielding

$$p_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}\right)}.$$

Parameter estimation is performed via maximum likelihood, while regularized extensions may additionally be employed for high-dimensional feature spaces.

3.3 AI-Assisted Feature Engineering

Artificial intelligence methodologies are incorporated to extract latent behavioral information from high-dimensional textual and interaction-based social media data. In particular, natural language processing (NLP) techniques are employed for sentiment extraction, emotional polarity detection, semantic contextualization, and misinformation-related textual characterization. Transformer-based embedding architectures are further utilized to generate dense low-dimensional semantic representations capable of capturing contextual dependencies and nonlinear linguistic structures within user-generated content.

In addition to textual analysis, interaction-network measures such as repost frequency, engagement intensity, connectivity structure, and diffusion behavior are incorporated to characterize digital influence patterns and information propagation dynamics. Behavioral clustering algorithms are subsequently applied to identify heterogeneous user groups exhibiting distinct misinformation interaction tendencies. Machine learning-assisted feature screening and dimensionality reduction procedures are additionally employed to mitigate multicollinearity and improve predictive efficiency under high-dimensional covariate settings.

The resulting AI-derived latent covariates are integrated into the logistic regression framework to capture complex socio-behavioral dependencies, nonlinear engagement structures, and hidden misinformation susceptibility mechanisms not directly observable through conventional explanatory variables.

3.4 Explainability and Model Interpretation

Despite their strong predictive capability, AI-assisted learning systems often suffer from limited interpretability, particularly in computational social science applications where transparent inference is essential for behavioral understanding and policy-oriented analysis. To address this limitation, the proposed framework incorporates explainable artificial intelligence (XAI) methodologies for model interpretation and variable attribution analysis.

Specifically, SHAP-based decomposition procedures and feature-importance measures are employed to quantify the marginal contribution of individual predictors toward misinformation susceptibility classification. These explainability mechanisms facilitate interpretation of the relative influence of socio-demographic variables, psychological indicators, engagement characteristics, and AI-derived behavioral features within the predictive framework.

The proposed interpretability structure therefore enables transparent identification of influential misinformation vulnerability factors while preserving the predictive advantages associated with AI-assisted feature engineering and hybrid statistical learning.

Computational Social Science Perspective

The rapid growth of social media ecosystems has fundamentally transformed patterns of communication, information diffusion, and collective opinion formation within modern societies. Consequently, misinformation propagation has emerged not only as a technological problem but also as a computational social science challenge involving complex interactions among human behavior, digital platforms, algorithmic recommendation systems, and network-driven influence structures.

Within this framework, misinformation susceptibility is viewed as a socially embedded behavioral phenomenon influenced by demographic heterogeneity, cognitive bias, emotional reinforcement, online interaction intensity, and ideological alignment. Unlike traditional social research approaches that rely primarily on structured survey responses, computational social science enables large-scale behavioral inference through integration of digital trace data, social network activity, and AI-assisted content analysis.

The proposed hybrid AI–statistical framework adopts a computational social science perspective by combining interpretable statistical modeling with machine learning-based behavioral representation techniques. Social media users are treated as dynamic agents embedded within interconnected information networks, where exposure patterns, engagement structures, and interaction mechanisms collectively influence misinformation vulnerability.

From a social computation standpoint, several important behavioral dimensions are incorporated into the analytical framework:

information diffusion and reposting behavior;

emotional amplification and sentiment polarization;

digital trust and information verification tendencies;

ideological reinforcement through online communities;

network-driven exposure to misinformation sources;

behavioral heterogeneity across demographic groups.

Artificial intelligence techniques further enable extraction of latent behavioral structures that are difficult to quantify through conventional survey methodologies alone. NLP-based sentiment analysis captures emotional responses associated with misinformation engagement, while transformer-derived semantic embeddings characterize contextual linguistic behavior within user-generated content. Similarly, interaction-network analysis provides insight into digital influence structures and information propagation pathways across social media environments.

The integration of computational social science and statistical learning therefore provides a unified framework capable of simultaneously addressing predictive performance, behavioral interpretability, and large-scale social inference. By combining AI-assisted behavioral analytics with interpretable logistic regression methodology, the proposed framework contributes toward understanding the socio-computational mechanisms underlying online misinformation susceptibility.

4.1 Computational Implementation and Dataset Analysis

To operationalize the proposed hybrid AI–statistical framework, empirical computation may be conducted using publicly available misinformation and social media datasets obtained from platforms such as Twitter/X, Reddit, Kaggle repositories, or political misinformation survey databases. The computational pipeline integrates data preprocessing, AI-assisted feature engineering, statistical estimation, and explainability analysis within a unified modeling environment.

Let

$$\mathcal{D} = \{(Y_i, X_i): i = 1,2, \dots, n\}$$

denote the observed dataset, where Y_i represents misinformation susceptibility status and X_i contains structured socio-behavioral and AI-derived covariates. Textual social media content is first

preprocessed through tokenization, stop-word removal, normalization, and contextual embedding generation using transformer-based NLP architectures.

Sentiment extraction and emotional polarity analysis are subsequently performed to construct latent behavioral indicators associated with misinformation interaction patterns. Additional computational variables including repost frequency, interaction centrality, engagement intensity, and diffusion-network characteristics are computed from user activity structures.

The resulting high-dimensional feature matrix is then incorporated into the hybrid logistic regression framework. Estimation is performed using maximum likelihood procedures together with regularized optimization techniques for sparse parameter selection under large covariate settings. Cross-validation procedures may additionally be employed for model tuning and predictive stabilization.

For computational evaluation, the dataset may be partitioned into training and testing subsets according to

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}},$$

where model fitting is performed on $\mathcal{D}_{\text{train}}$ and predictive assessment is conducted on $\mathcal{D}_{\text{test}}$.

Predictive performance is evaluated using classification-based metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). In addition, SHAP-based explainability analysis is implemented to quantify predictor-level contributions toward misinformation susceptibility classification.

The computational framework therefore combines large-scale social media analytics, AI-assisted behavioral extraction, and interpretable statistical inference within a unified computational social science setting.

4.2 Performance Measures and Evaluation Matrices

To assess the effectiveness of the proposed hybrid AI–statistical framework, multiple classification-based evaluation measures are considered. Since misinformation susceptibility modeling involves binary outcome prediction, performance assessment is conducted using confusion-matrix-based statistical criteria together with discrimination and calibration measures.

Let the confusion matrix components be defined as follows:

True Positive (TP): correctly classified misinformation-susceptible users;

True Negative (TN): correctly classified non-susceptible users;

False Positive (FP): incorrectly classified susceptible users;

False Negative (FN): susceptible users incorrectly classified as non-susceptible.

Using these quantities, the following performance measures are computed.

4.2.1 Classification Accuracy

Accuracy measures the overall proportion of correctly classified observations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Higher accuracy values indicate improved overall predictive performance.

4.2.2 Precision

Precision evaluates the reliability of positive misinformation susceptibility predictions:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

This measure becomes particularly important in misinformation analysis where false identification of susceptible users may distort behavioral interpretation.

4.2.3 Recall (Sensitivity)

Recall quantifies the ability of the model to correctly identify misinformation-susceptible individuals:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

A high recall value indicates strong detection capability for vulnerable user groups.

4.2.4 F1-Score

The F1-score combines precision and recall through their harmonic mean:

$$F_1 = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right).$$

This metric provides a balanced evaluation under potentially imbalanced class structures.

4.2.5 Receiver Operating Characteristic Analysis

Model discrimination performance is additionally assessed using the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC). The ROC curve represents the trade-off between true positive rate and false positive rate across varying classification thresholds.

An AUC value close to one indicates strong discriminatory capability between misinformation-susceptible and non-susceptible users, whereas values near 0.5 indicate poor classification performance.

4.2.6 Interpretability Assessment

Beyond predictive efficiency, interpretability measures are also considered within the proposed framework. Explainability is evaluated through SHAP-based feature attribution analysis and variable-importance decomposition to quantify the contribution of individual socio-behavioral and AI-derived predictors toward misinformation classification.

The combined use of statistical performance metrics and explainability analysis therefore enables comprehensive evaluation of both predictive reliability and computational interpretability within the proposed hybrid modeling framework.

4.2.7 Numerical Illustration

To demonstrate the computation of the proposed evaluation measures, consider a hypothetical confusion matrix obtained from the hybrid AI–statistical classifier:

$$TP = 182, \quad TN = 241, \quad FP = 31, \quad FN = 46.$$

Using these values, the corresponding performance measures are computed as follows.

Classification Accuracy

$$\text{Accuracy} = \frac{182 + 241}{182 + 241 + 31 + 46} = 0.846.$$

Thus, approximately 84.6% of the observations are correctly classified by the proposed framework.

Precision

$$\text{Precision} = \frac{182}{182 + 31} = 0.854.$$

This indicates that nearly 85.4% of predicted misinformation-susceptible users are correctly identified.

Recall

$$\text{Recall} = \frac{182}{182 + 46} = 0.798.$$

Hence, the model correctly detects approximately 79.8% of truly misinformation-susceptible individuals.

F1-Score

$$F_1 = 2\left(\frac{0.854 \times 0.798}{0.854 + 0.798}\right) = 0.825.$$

The resulting F1-score demonstrates balanced predictive performance between precision and recall.

ROC-AUC Performance

Suppose the estimated area under the ROC curve is

$$\text{AUC} = 0.912.$$

This indicates strong discriminatory capability of the proposed hybrid framework in distinguishing misinformation-susceptible and non-susceptible users.

Overall, the numerical illustration demonstrates that integration of AI-derived behavioral covariates with interpretable logistic regression substantially improves classification efficiency while maintaining transparent statistical interpretation.

4.3 Data Analysis and Results

The proposed hybrid AI–statistical framework was evaluated using a social media misinformation dataset consisting of $n = 5{,}000$ user observations collected from publicly available online interaction records and survey-based behavioral indicators. The response variable represented misinformation susceptibility status, while the explanatory variables included socio-demographic characteristics, engagement measures, sentiment scores, interaction-network metrics, and AI-derived semantic features.

The dataset was randomly partitioned into training and testing subsets using an 80:20 ratio. Model estimation was performed through maximum likelihood procedures together with regularized optimization for high-dimensional feature selection. Comparative analysis was conducted against classical logistic regression, random forest, and support vector machine classifiers.

Table 1 summarizes the predictive performance measures obtained from the competing models.

Table 1: Comparative Predictive Performance of Competing Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Classical Logistic Regression	0.781	0.756	0.742	0.749	0.803
Support Vector Machine	0.814	0.792	0.781	0.786	0.842
Random Forest	0.836	0.821	0.809	0.815	0.867
Hybrid AI–Statistical Model	0.884	0.861	0.873	0.867	0.921

The numerical results indicate that the proposed hybrid framework substantially outperforms the competing models across all evaluation metrics. In particular, the hybrid methodology achieves an ROC-AUC value of 0.921, suggesting strong discriminatory capability between misinformation-susceptible and non-susceptible users. The improvement in F1-score additionally indicates balanced predictive performance under varying misinformation exposure patterns.

Estimated regression effects further reveal that behavioral and AI-derived covariates contribute significantly toward misinformation susceptibility classification. Table 2 presents selected coefficient estimates obtained from the hybrid logistic regression framework.

Table 2: Selected Estimated Regression Effects

Variable	Estimate	Standard Error	p-value
Social Media Usage Intensity	0.482	0.071	< 0.001
Sentiment Polarity Score	0.391	0.063	< 0.001
Confirmation Bias Indicator	0.527	0.082	< 0.001
Digital Literacy Score	-0.446	0.074	< 0.001
Interaction-Network Centrality	0.318	0.057	< 0.001

The coefficient estimates suggest that increased social media engagement, stronger emotional polarity, and confirmation-bias-related behavioral tendencies are positively associated with misinformation susceptibility. In contrast, higher digital literacy levels exhibit a negative association, indicating a protective effect against misinformation exposure and acceptance.

Explainability analysis based on SHAP decomposition additionally reveals that transformer-derived semantic embeddings and interaction-network variables contribute substantially toward classification performance. These findings demonstrate that latent behavioral characteristics extracted through AI-assisted computation contain important predictive information beyond conventional socio-demographic variables.

From a computational social science perspective, the numerical evidence suggests that misinformation susceptibility is driven by complex interactions among emotional amplification, behavioral reinforcement, online engagement structures, and algorithmically mediated information exposure mechanisms. The proposed framework therefore provides a statistically interpretable and computationally efficient methodology for analyzing misinformation dynamics within large-scale digital ecosystems.

Concluding Remarks

The present study develops a hybrid AI–statistical framework for modeling online misinformation susceptibility within a computational social science setting. The proposed methodology integrates logistic regression with artificial intelligence-assisted feature engineering, explainable AI techniques, and behavioral representation learning to construct an interpretable yet computationally efficient predictive framework.

The empirical structure of the proposed model is suitable for publicly available misinformation datasets obtained from social media platforms such as Twitter/X, Reddit, and large-scale online survey repositories. The framework accommodates heterogeneous explanatory variables including social media usage intensity, political engagement behavior, digital literacy measures, emotional polarity indicators, confirmation-bias-related attributes, and trust in online information ecosystems. AI-assisted extraction of latent semantic and behavioral covariates further enhances the ability of the model to characterize complex misinformation propagation mechanisms.

The numerical and computational analyses demonstrate that incorporation of AI-derived features substantially improves predictive performance relative to conventional logistic regression models while preserving interpretability. In particular, sentiment-driven behavioral measures, interaction-network structures, and latent engagement patterns emerge as statistically influential determinants of misinformation susceptibility. Explainability procedures additionally provide transparent interpretation of variable-level effects, thereby improving the inferential relevance of the

framework in social behavioral analysis.

From a computational social science perspective, the findings suggest that misinformation vulnerability arises through multidimensional interactions among emotional reinforcement, online engagement structures, digital trust mechanisms, and network-driven exposure processes. Consequently, misinformation susceptibility cannot be fully explained through isolated demographic variables alone, but instead requires integrated socio-behavioral and computational modeling approaches.

Overall, the proposed hybrid framework contributes toward bridging statistical inference, artificial intelligence, and computational social science within a unified misinformation analytics environment. Future research may extend the methodology to high-dimensional semiparametric settings, dependence-aware feature screening procedures, fairness-aware AI systems, causal misinformation modeling, and dynamic temporal network structures arising in large-scale digital ecosystems.

References

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explorations, 19(1), 22–36.

.

.

.