Quantitative Risk Analysis of AI-Generated Misinformation Propagation in Social Media

Hariharan Duraisingh
40303001

Kausik Sabapathy Janarthanam
40299673

Abstract— **This project investigates the risks associated with the spread of AI-generated misinformation on social media platforms. The main objective is to assess and quantify the threats posed by such misinformation using various quantitative risk analysis techniques. Key methods employed include Bayesian analysis, Poisson distribution, machine learning classifiers (Decision Tree, Random Forest, Gradient Boosting), Fault Tree and Event Tree Analyses, and a sensitivity analysis and payoff tables based on logistic regression. An Influence Diagram was also developed to visually map the decision-making framework and highlight key uncertainties and outcomes. The analysis draws from real-world data to calculate probabilities and model misinformation propagation. Our findings show that combining statistical and machine learning approaches can effectively evaluate the likelihood and consequences of misinformation, providing critical insight for mitigation strategies. The study has implications for enhancing policy decisions, improving content moderation systems, and promoting public awareness to reduce the societal and economic risks posed by AI-generated misinformation.**

Keywords— AI-generated misinformation, Bayesian analysis, sensitivity analysis, Poisson distribution, decision trees, influence diagram, risk assessment, social media, risk mitigation

I. Introduction

Artificial Intelligence (AI) has revolutionized digital communication by enabling rapid content generation and personalization. While this has led to several benefits, it has also facilitated the rapid spread of misinformation, especially across social media platforms. With the growing accessibility of generative AI tools, even individuals with minimal technical skills can create and disseminate highly convincing fake news, images, and videos.

This surge in AI-generated misinformation presents serious risks—including public misinformation, financial panic, erosion of trust in institutions, and manipulation during critical events such as elections or pandemics. Social media platforms, due to their viral content algorithms and large user base, are particularly susceptible to these threats.

The aim of this project is to perform a comprehensive quantitative risk analysis on the propagation of AI-generated misinformation on social media. We employ a combination of analytical and statistical techniques—including Bayesian analysis, Poisson distribution, decision trees, random forest, gradient boosting, sensitivity analysis using payoff tables, and visual models such as influence diagrams, fault trees, and event trees. These methods help quantify the risk, evaluate the effectiveness of different mitigation strategies, and provide actionable insights for stakeholders, regulators, and technology platforms.

II. Literature Review

The proliferation of AI-generated misinformation, particularly deepfakes and fake news, poses significant threats to public trust, democratic processes, and personal security. Quantitative risk analysis techniques such as Bayesian Networks and Fault Tree Analysis (FTA) have emerged as vital tools to assess and mitigate these risks.

Misinformation Propagation and Echo Chambers

Del Vicario et al. (2016) explored how misinformation spreads across social networks like Facebook. Their quantitative analysis identified that selective exposure and homogeneous clusters (echo chambers) are key drivers of misinformation diffusion, often leading to virality through emotionally charged, polarizing content. These findings underscore the need to model such propagation patterns when assessing systemic risks.

Deepfakes and Synthetic Media Threats

Deepfake technology, which utilizes advanced AI to fabricate realistic media, significantly complicates misinformation detection. Kharvi (2024) documented numerous real-world incidents—including deepfake-driven scams and political manipulation—and emphasized the psychological, political, and cybersecurity risks deepfakes Motivated reasoning, where individuals trust information aligning with their beliefs, further exacerbates the challenge.

Quantitative Risk Assessment Techniques

- **Bayesian Belief Networks (BBNs):** These probabilistic graphical models are effective for early detection and decision-making under uncertainty. They help quantify the likelihood of policy violations and risk exposure from misinformation campaigns.
- **Fault Tree Analysis (FTA):** Used for structured analysis of systemic failures, FTA has been applied in content moderation and AI safety systems. Integrating FTA with systems-theoretic approaches enhances predictive capabilities in complex digital ecosystems.

Data-Driven Models and Detection Frameworks

Various studies propose models leveraging machine learning and deep learning to detect misinformation and assess its spread:

- **CNNs and hybrid models** are employed for real-time fake news detection, though they require large, labeled datasets and struggle against evolving threats.
- Studies like "Fake Newsnet" aggregate content, user interaction, and spatiotemporal data to enrich detection accuracy and enable granular risk modeling.

Mitigation Strategies

Effective mitigation must be multi-pronged:

- **Digital Watermarking and Provenance (C2PA):** Embedding verifiable metadata at the content creation point can help track and authenticate media across platforms.
- **Blockchain Integration:** Decentralized verification systems offer immutable logs for content validation and help build trust between creators and consumers.
- **Regulatory and Educational Approaches:** Government policies, user awareness programs, and ethical AI standards are vital to reduce the spread and impact of deepfakes and misinformation

III. Problem description

The dataset used in this study is obtained from the publicly available COVID-19 healthcare misinformation dataset hosted on GitHub at https://github.com/cuilimeng/CoAID. Specifically, data from the "05-01-2020" folder was selected for analysis. The dataset contains labeled tweets categorized as real or fake, including metadata such as the number of replies, shares, source credibility, and timestamps. This dataset serves as the basis for training classification models and performing probabilistic analysis.

The system at hand involves multiple interconnected components:

- **Content features**: Textual and metadata attributes (e.g., reply count, retweet count, source credibility)
- **Classification mechanism**: ML models (Logistic Regression, Decision Tree, Random Forest, XGBoost) are used to classify tweets
- **Risk quantification**: Probabilistic models (Bayesian and Poisson), payoff tables, influence diagrams and sensitivity analysis.

Visualizations such as confusion matrices, sensitivity graphs, and decision tree structures were created to interpret classifier performance and support risk evaluation

IV. Risk Assessment

The major risks identified in the propagation of AI-generated misinformation on social media include:

1. Amplification of misinformation: Misinformation can rapidly spread due to viral algorithms on social media platforms that prioritize content with high engagement. This leads to more users viewing, sharing, and engaging with the misinformation, amplifying its spread.
2. Financial instability: Misinformation related to financial stability, such as fake claims about stock market crashes or banking issues, can lead to large-scale panic. A recent study showed that 60% of individuals exposed to such misinformation were likely to withdraw their funds from banks, causing economic instability.
3. Loss of trust: As AI-generated misinformation spreads, public trust in institutions like media, government, and financial organizations erodes. This could result in a disengaged or skeptical public, undermining societal norms and institutional authority.
4. Political manipulation: AI-generated misinformation can be used strategically to manipulate elections or political events. For example, targeted misinformation during elections can sway voters or incite social unrest.

Subsystems and Failure Impacts

The failure of certain subsystems can cascade into overall system failure, further exacerbating misinformation propagation.

1. Content moderation systems: If AI content moderation policies are too relaxed (e.g., allowing unchecked content or not detecting fake news), misinformation can go unchecked. For instance, platforms with opaque policies or relaxed moderation may inadvertently promote harmful content.
2. AI detection models: Inaccuracies in AI classification models, whether due to algorithmic bias or improper training data, can lead to a high rate of false positives and false negatives. This misclassification can either flag legitimate content as misinformation or allow fake news to spread.
3. User engagement: When users unknowingly engage with fake news (e.g., sharing misleading content), it increases the likelihood of that misinformation spreading. The virality of social media further accelerates the amplification of such content.

These failures lead to cascading effects, where the spread of misinformation causes economic, social, and political damage. The lack of accurate detection or inadequate moderation can enable misinformation to propagate unchecked, thereby increasing the overall societal risk.

V. Risk analysis

Bayesian Analysis

Bayesian analysis is a powerful probabilistic method used to update the probability of an event occurring based on new evidence. In the context of misinformation propagation, it helps in updating the likelihood that a piece of content is fake based on new data, such as the number of replies a tweet receives. The prior probability reflects the initial assumption about the content being fake or real, and Bayes' Theorem allows for recalculating this probability as more information (like user engagement) becomes available.

Bayes Theorem is given by:

$$P(h|e)=P(e|h) \times P(h) / P(e)$$

Where:

- P(h|e) is the posterior probability, or the probability that the news is fake, given the number of replies.
- P(e|h) is the likelihood, the probability of observing the replies, given that the news is fake.
- P(h) is the prior probability of the news being fake.
- P(e) is the marginal likelihood, or the total probability of observing the replies, considering both fake and real news.

Step-by-Step Explanation

Prior Probability: This is the probability that a piece of news is fake before considering the number of replies. It is calculated as the ratio of fake news tweets to total tweets:

p(h) = Number of fake news tweets / Total number of news tweets

= 5721 / 69836 = 0.0819

Thus, the prior probability that the news is fake is approximately 8.19%.

P(¬h) = 1−P(h)

= 1 – 0.0819 = 0.9181

Thus, the prior probability that the news is real is approximately 91.81%

Likelihood (P(e|h)): This is the probability of observing a certain number of replies, assuming the news is fake. It is calculated as the ratio of replies for fake news tweets to the number of fake news tweets:

p(e|h) = Total number of replies for fake news tweets / Number of fake news tweets

= 5721 / 1266 = 4.52

Similarly, the likelihood of receiving replies for real news tweets is:

p(e|¬h) = Total number of replies for real news tweets / Number of real news tweets

= 64115 / 13481 = 4.75

Marginal Likelihood (P(e)): This is the total probability of observing a certain number of replies, considering both fake and real news. It is calculated as:

P(e) = P(e|h) × P(h) + P(e|¬h) × P(¬h)

Substituting in the known values:

P(e) = (4.52×0.0819) + (4.75×0.9181) = 4.74

Posterior Probability: Now we can apply **Bayes' Theorem** to calculate the updated probability that the news is fake, given the number of replies:

P(h|e) = P(e|h) x P(h) / P(e)

= 4.52×0.0819 / 4.74 = 0.0781

Therefore, the probability that the news is fake, given the replies, is approximately 7.82%.

Implications of the Result

This result means that, based on the number of replies, there is a **7.82%** probability that the news is fake. This updated probability, informed by new evidence (user engagement), helps in making more informed decisions about the authenticity of content.

Poisson Distribution

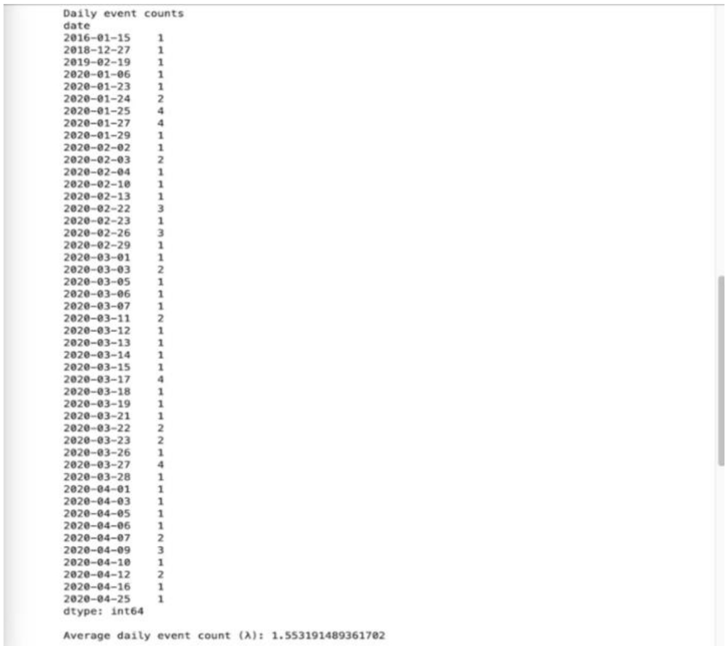
The Poisson distribution is a statistical method used to model the probability of a given number of events occurring within a fixed interval of time or space, given a known average rate of occurrence. In this case, we are modeling the number of misinformation events per day, based on data from the CoAID dataset.

The Poisson formula is given by:

P(k) = e^-λ x λ^k / k!

Where:

- P(k) is the probability of exactly k events occurring.
- λ (lambda) is the average number of events per day.
- k is the number of events for which we want to find the probability.
- e is the base of the natural logarithm.



```
Daily event counts
date
2016-01-15    1
2018-12-27    1
2019-02-19    1
2020-01-06    1
2020-01-23    1
2020-01-24    2
2020-01-25    4
2020-01-27    4
2020-01-29    1
2020-02-02    1
2020-02-03    2
2020-02-04    1
2020-02-10    1
2020-02-13    1
2020-02-22    3
2020-02-23    1
2020-02-26    3
2020-02-29    1
2020-03-01    1
2020-03-03    2
2020-03-05    1
2020-03-06    1
2020-03-07    1
2020-03-11    2
2020-03-12    1
2020-03-13    1
2020-03-14    1
2020-03-15    1
2020-03-17    4
2020-03-18    1
2020-03-19    1
2020-03-21    1
2020-03-22    2
2020-03-23    2
2020-03-26    1
2020-03-27    4
2020-03-28    1
2020-04-01    1
2020-04-03    1
2020-04-05    1
2020-04-06    1
2020-04-07    2
2020-04-09    3
2020-04-10    1
2020-04-12    2
2020-04-16    1
2020-04-25    1
dtype: int64

Average daily event count (λ): 1.553191489361702
```

Fig 1: Daily Event Counts of Fake Tweets

From the dataset we found that average daily event count (**λ**) = 1.553

Calculation of Poisson Distribution

Probability of 0 events (k = 0):

P(0) = (e ^ -1.553) x (1.553 ^ 0) / 0! = 0.2116

Probability of 1 event (k = 1):

P(1) = (e ^ -1.553) x (1.553 ^ 1) / 1! = 0.3286

Probability of 2 events (k = 2):

P(2) = (e ^ -1.553) x (1.553 ^ 2) / 2! = 0.2552

Probability of 3 events (k = 3):

P(3) = (e ^ -1.553) x (1.553 ^ 3) / 3! = 0.1321

Probability of 4 events (k = 4):

P(4) = (e ^ -1.553) x (1.553 ^ 4) / 4! = 0.0513

Probability of 5 events (k = 5):

P(5) = (e ^ -1.553) x (1.553 ^ 5) / 5! = 0.0519

Probability of 6 events (k = 6):

P(6) = (e ^ -1.553) x (1.553 ^ 6) / 6! = 0.0041

Probability of 7 events (k = 7):

P(7) = (e ^ -1.553) x (1.553 ^ 7) / 7! = 0.0009

Probability of 8 events (k = 8):

P(8) = (e ^ -1.553) x (1.553 ^ 8) / 8! = 0.0002

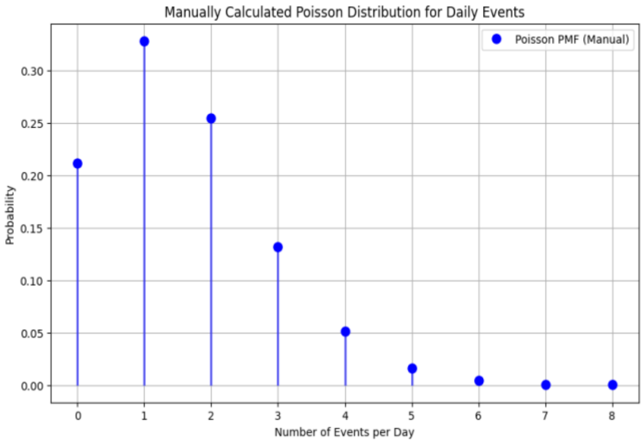


Fig 2: Poisson Distribution for Daily Events

Explanation of Results:

- The highest probability is for 1 event per day (32.86%), followed by 2 events (25.52%), and 0 events (21.16%).
- The probabilities for 3 or more events per day decrease rapidly, with 3 events having only a 13.21% chance.
- The Poisson distribution indicates that most days will have 1 to 2 events, with very few days seeing a higher number of misinformation events.

Decision Tree Classifier

A Decision Tree is a supervised machine learning algorithm used for classification and regression tasks. It works by recursively splitting the dataset into subsets based on different feature values, eventually forming a tree-like structure of decisions. Each internal node of the tree represents a "test" or "decision" based on a feature, while each leaf node represents an output label (class or value).

Why We Chose Decision Tree for Fake News Classification:

Simplicity: For a task like fake news detection, having a model that is easy to interpret and explain is crucial, especially when presenting results to non-technical stakeholders.

High Interpretability: A decision tree gives us insight into how different words or content affect the classification, helping us understand the features that matter most in predicting fake or real news.

Good Baseline: Decision trees serve as a good baseline model to check if more complex models (like Random Forests or XGBoost) are necessary for better performance.

Feature Analysis: Decision trees also provide feature importance measures, helping us identify which words or phrases in news articles are most indicative of fake or real news, which is valuable for improving model performance.

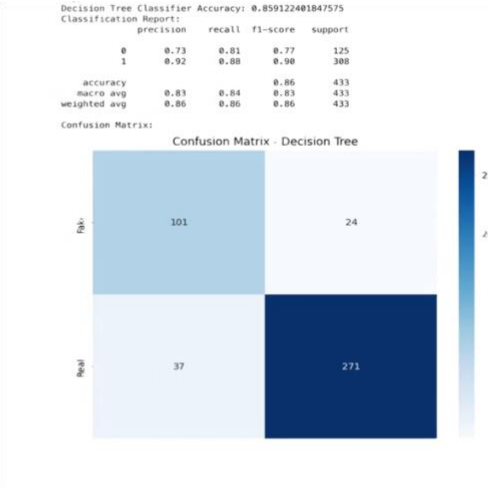
Confusion Matrix and Accuracy

The confusion matrix is a helpful way to evaluate the model's performance. The confusion matrix displays how well the model predicted the classes (fake or real) compared to the actual labels.

- **True Positive (TP):** 271 articles were correctly classified as real.
- **False Positive (FP)**: 37 articles were incorrectly classified as real when they were fake.
- **True Negative (TN)**: 101 articles were correctly classified as fake.
- **False Negative (FN)**: 24 articles were incorrectly classified as fake when they were real.

The accuracy of the model is 85.91%, indicating that the model correctly classified the articles in the test set about 86% of the time.

Confusion Matrix Visualization:



```
Decision Tree Classifier Accuracy: 0.859122401847575
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.81      0.77       125
           1       0.92      0.88      0.90       308

    accuracy                           0.86       433
   macro avg       0.83      0.84      0.83       433
weighted avg       0.86      0.86      0.86       433

Confusion Matrix:
```

Fig 3: Confusion Matrix: Decision Tree

This matrix shows the distribution of predictions versus actual labels. It reveals the false positives and false negatives, helping us understand where the model needs improvement.

Decision Tree Diagram

The decision tree diagram visualizes the decision-making process of the classifier, showing how it splits the data at each node based on the feature values:

- The tree starts by checking the 2020 feature, and if the value is <= 0.018, it classifies the news as real. If it's greater, the next decision is made based on other features like coronavirus and covid.
- The tree also shows how Gini impurity is calculated at each node to ensure the most effective splits, where a lower Gini value indicates better separation between the classes.

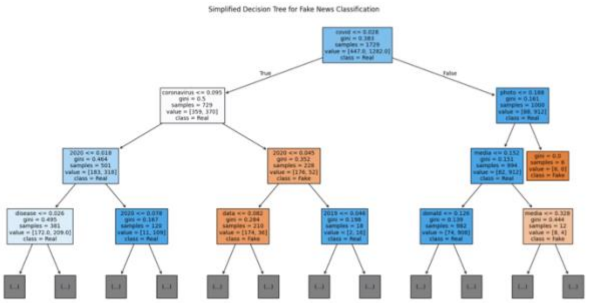


Fig 4: Decision Tree for Fake News Classification

This tree is built using the most important features, and it clearly shows how the classifier is distinguishing between real and fake news based on the values of those features.

Random Forest Classifier:

Random Forest is an ensemble learning algorithm used for both classification and regression tasks. It operates by constructing a collection of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's a very effective and widely used machine learning model because of its high accuracy, robustness, and ability to handle both large datasets and datasets with high dimensionality.

Why We Chose Random Forest for Fake News Classification:

Complexity: Fake news classification involves complex patterns in textual data. A single decision tree may struggle to capture these complex relationships, while a Random Forest with many trees improves its capacity to model these relationships and generalize better.

Generalization: Random Forest helps avoid overfitting, which can occur when using a single model that fits the data too closely. This is particularly important in fake news detection, where it's crucial to generalize across many different types of content.

Robustness to Noise: Fake news data is often noisy, and small errors in classification are common. Random Forest, due to its ensemble approach, provides better performance in this noisy setting.

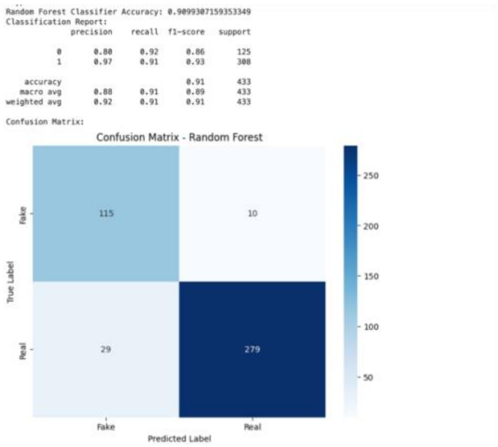
Confusion Matrix and Accuracy

The confusion matrix for the Random Forest Classifier is as follows:

- **True Positive (TP):** 279 real news articles correctly classified as real.
- **False Positive (FP):** 10 fake news articles incorrectly classified as real.
- **True Negative (TN):** 115 fake news articles correctly classified as fake.
- **False Negative (FN):** 29 real news articles incorrectly classified as fake.

The accuracy of the model is 90.99%, indicating that the model correctly classified the articles in the test set about 91% of the time.

Confusion Matrix Visualization:



```
Random Forest Classifier Accuracy: 0.9099307159353349
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.92      0.86       125
           1       0.97      0.91      0.93       308

    accuracy                           0.91       433
   macro avg       0.88      0.91      0.89       433
weighted avg       0.92      0.91      0.91       433

Confusion Matrix:
```

Fig 5: Confusion Matrix: Random Forest

This matrix shows the distribution of real and fake news predictions. The false positives and false negatives indicate areas where the model can improve, especially in distinguishing fake news from real news.

XGBoost (Extreme Gradient Boosting):

XGBoost is an advanced machine learning algorithm based on the concept of gradient boosting, which is an ensemble technique that builds a strong predictive model by combining the predictions of several weak learners (usually decision trees). XGBoost is an optimized and scalable version of gradient boosting, and it has gained popularity due to its high performance, speed, and efficiency in solving complex tasks.

Why We Used XGBoost for Fake News Classification:

High Accuracy: XGBoost consistently delivers high predictive accuracy due to its ability to learn from residuals and its regularization techniques that help reduce overfitting. In text classification tasks, such as fake news detection, where subtle patterns and relationships need to be identified, XGBoost is highly effective.

Efficient Handling of Large Datasets: XGBoost is particularly known for its speed and efficiency, especially with large datasets. It uses a parallelized tree construction algorithm and optimized data storage techniques, making it ideal for handling high-dimensional data like text.

Feature Importance: XGBoost provides insights into which features (in this case, words or phrases) contribute the most to the classification of real or fake news. This makes the model interpretable and useful in understanding what drives its predictions.

Handling Imbalanced Classes: Fake news datasets tend to be imbalanced (more real news than fake news). XGBoost can handle class imbalance efficiently by adjusting class weights during training, ensuring the minority class (fake news) is given more attention during the learning process.

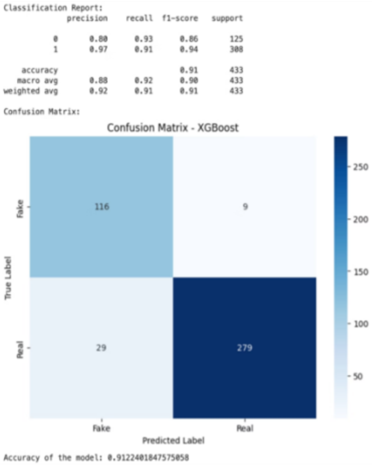
Confusion Matrix and Accuracy

The confusion matrix for the XGBoost Classifier is as follows:

- **True Positive (TP)**: 279 real news articles correctly classified as real.
- **False Positive (FP)**: 9 fake news articles incorrectly classified as real.
- **True Negative (TN)**: 116 fake news articles correctly classified as fake.
- **False Negative (FN)**: 29 real news articles incorrectly classified as fake.

The accuracy of the model is 91.24%, meaning that the model correctly classified 91.24% of the test set.

Confusion Matrix Visualization:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.93      0.86       125
           1       0.97      0.91      0.94       308

    accuracy                           0.91       433
   macro avg       0.88      0.92      0.90       433
weighted avg       0.92      0.91      0.91       433

Confusion Matrix:
```

```
Accuracy of the model: 0.9122401847575058
```

Fig 6: Confusion Matrix: XGBoost

This matrix highlights the misclassifications of real and fake news. The false positives and false negatives are much lower than in the other models, showing the effectiveness of XGBoost in distinguishing between the two classes.

Comparison Chart:

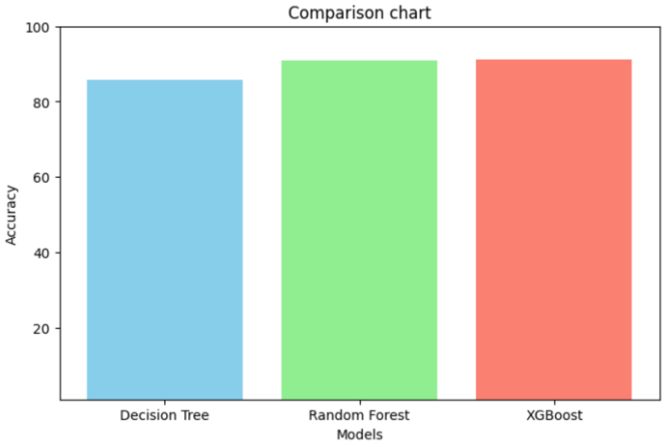


Fig 7: Comparison Chart

The bar chart above provides a comparison of accuracy between three machine learning models used for the fake news classification task: Decision Tree, Random Forest, and XGBoost.

- **Decision Tree**: The Decision Tree Classifier achieved an accuracy of **85.91%**, performing well but with some potential overfitting due to its simplicity. It offers an intuitive structure, making it a good baseline model.
- **Random Forest**: The Random Forest Classifier outperformed the Decision Tree, with an accuracy of **90.99%**. Random Forest's ensemble learning mechanism, which combines multiple decision trees, improved accuracy and robustness by reducing variance.
- **XGBoost**: The XGBoost classifier achieved the highest accuracy of **91.24%**, demonstrating its strength in handling complex, high-dimensional data. XGBoost's gradient boosting and regularization techniques resulted in superior performance compared to both Decision Tree and Random Forest.

Sensitivity Analysis:

We conducted a sensitivity analysis on a logistic regression model used to classify news articles as either real or fake. We performed a series of steps, including the construction of a payoff table based on model predictions, followed by the calculation of expected values (EVs) and a sensitivity analysis to determine how the model's performance changes with varying probabilities.

Step 1: Building the Payoff Table:

A payoff table was constructed using the confusion matrix of the logistic regression model. This confusion matrix provides the frequencies of model outcomes (True Positives, False Positives, False Negatives, and True Negatives) when predicting news articles. These outcomes represent the rewards or penalties of the model's decisions, which were used as the payoffs for our analysis.

	State: Real	**State: Fake**
Decision: Predict Real	**308**	**49**
Decision: Predict Fake	**6**	**70**

Fig 8: Payoff Table

Step 2: Calculating Expected Values (EVs):

The next step involved calculating the Expected Values (EVs) for each decision alternative (predicting real or fake), considering the probability of news being real (p) or fake (1 - p). EV for Predicting Real:

The expected value for predicting Real is calculated by considering:

- TP when news is real (with probability p), and
- FP when news is fake (with probability 1 - p).

Thus, the formula for EV(Predict Real) is:

$$EV(Real) = 308p + 49(1-p) = 259p + 49$$

EV for Predicting Fake:

Similarly, the expected value for predicting Fake is calculated by considering:

- FN when news is real (with probability p), and
- TN when news is fake (with probability 1 - p).

Thus, the formula for EV(Predict Fake) is**:**

$$EV(Fake) = 6p + 70(1-p) = -64p + 70$$

Step 3: Sensitivity Analysis

To analyze how the model's performance changes with varying probabilities, we solved for the threshold where both EVs are equal, indicating the point where the decision to predict "real" or "fake" changes.

EV(Predict Real) = 259p + 49

p	EV(Predict Real)
0	259(0) + 49 = 49
1	259(1) + 49 = 308

Therefore, the line goes from (0, 49) to (1, 308)

EV(Predict Fake) = -64p + 70

p	EV(Predict Fake)
0	-64(0) + 70 = 70
1	-64(1) + 70 = 6

Therefore, the line goes from (0, 70) to (1, 6)

EV(Predict Real) = EV(Predict Fake):

$$259p+49 = -64p+70$$

$$323p = 21$$

$$P = 0.065$$

Thus, the sensitivity threshold is at $p \approx 0.065$. This means:

- If the probability that the news is real (p) is less than 0.065, predicting Fake gives a higher expected payoff.
- If p is greater than 0.065, predicting Real gives a higher expected payoff.

Sensitivity Graph:

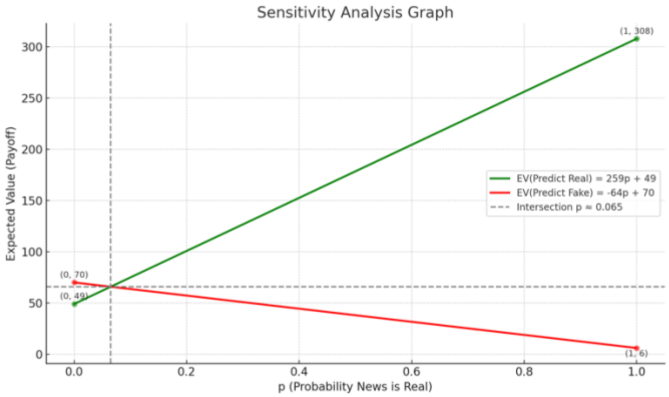


Fig 9: Sensitivity Graph

The sensitivity graph clearly indicates the threshold at which the model's recommended decision changes:

- If **p < 0.065:** The model predicts Fake because the expected value of predicting fake is higher.
- If **p > 0.065:** The model predicts Real because the expected value of predicting real is higher.

This threshold shows that the decision to classify a news article as real or fake is sensitive to the model's confidence (p) about the news being real.

Influence Diagram:

An influence diagram is a powerful tool used to visualize the relationships between decisions, uncertainties, and outcomes in complex systems. In the context of AI-generated misinformation propagation on social media, this diagram models the interactions among key decision variables, chance events, and their ultimate impact on the spread of misinformation.

The diagram focuses on the dynamics of content moderation, policy transparency, and user engagement in shaping the propagation of AI-generated fake news. It is structured according to standard influence diagram conventions, with square nodes representing decision variables, circular nodes representing uncertain (chance) events, and a diamond-shaped node representing the final consequence or outcome. By mapping out these elements, the diagram provides a structured representation of how different factors—ranging from platform decisions to user behavior—combine to influence the final risk of misinformation spreading across social networks.

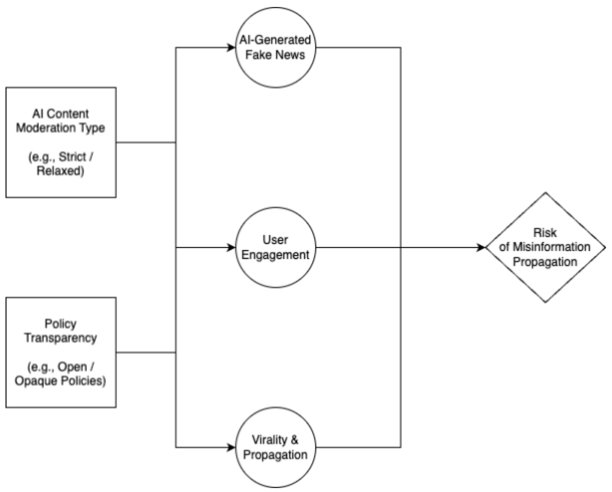

AI Content Moderation Type (e.g., Strict / Relaxed)
Policy Transparency (e.g., Open / Opaque Policies)
AI-Generated Fake News
User Engagement
Virality & Propagation
Risk of Misinformation Propagation

Fig 10: Influence Diagram

Decision Nodes

1. **AI Content Moderation Type**: This node captures the strategic choices made by platform administrators regarding how AI-generated content is handled—ranging from strict moderation using human oversight or advanced AI filters to more relaxed, automated-only systems.
2. **Policy Transparency**: This decision node reflects whether the platform adopts transparent policies (e.g., clearly stating what constitutes misinformation) or maintains vague or opaque moderation guidelines. Policy transparency can affect user behavior and trust.

Chance Nodes

- **AI-Generated Fake News**: Represents the volume and frequency of misleading or false information generated using AI tools.
- **User Engagement**: Captures how users interact with such content (e.g., likes, shares, comments), which significantly affects its reach.
- **Virality & Propagation**: Denotes how quickly and widely the misinformation spreads across the network, often influenced by engagement patterns and recommendation algorithms.

Consequence Node

- **Risk of Misinformation Propagation**: This is the primary outcome of interest, quantifying the overall threat level based on how misinformation emerges, spreads, and is (or isn't) mitigated by moderation and policy decisions.

Fault Tree Analysis:

Fault Tree Analysis (FTA) is a deductive risk assessment tool used to identify the root causes of a specific undesirable event, referred to as the top event. It is a graphical representation that breaks down the causes into multiple layers of intermediate and basic events. These events are connected using logical gates such as AND and OR, which define how basic events interact to cause intermediate events and ultimately the top event.

In the context of this analysis, the top event is the spread of misinformation and false claims, particularly in the realm of COVID-19-related information. By using FTA, we can assess the contributing factors that lead to the spread of misinformation, such as social media amplification, sensationalism, Over-generalized Conclusions, and Exaggerated Claims in Headline. This helps us quantify the risk of misinformation propagation and identify key areas that need attention in combating false information.

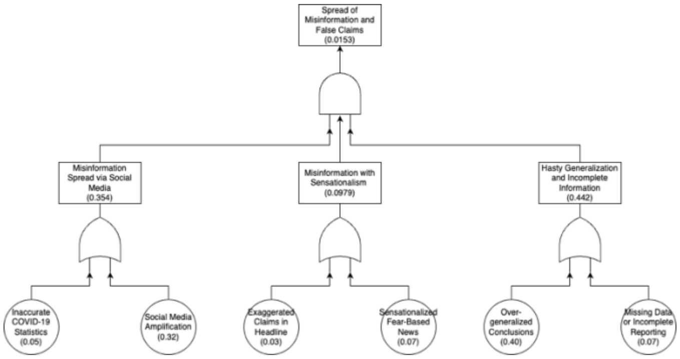

Spread of Misinformation and False Claims (0.0153)
Misinformation Spread via Social Media (0.354)
Misinformation with Sensationalism (0.0979)
Hasty Generalization and Incomplete Information (0.442)
Inaccurate COVID-19 Statistics (0.05)
Social Media Amplification (0.32)
Exaggerated Claims in Headline (0.03)
Sensationalized Fear-Based News (0.07)
Over-generalized Conclusions (0.40)
Missing Data or Incomplete Reporting (0.07)

Fig 11: Fault Tree

Calculating Probabilities:

To find the probability of each basic event, we used keywords (also known as stopping words) that are associated with each event. These keywords were searched within the content and title columns of the dataset. The number of occurrences of these keywords in each article was counted, which was then used to estimate the probability of each event.

P(Basic Event)=Number of Articles Containing Keywords/ Total Number of Articles

P(Inaccurate COVID-19 Statistics) = 0.05

P(Social Media Amplification) = 0.32

P(Exaggerated Claims in Headline) = 0.03

P(Sensationalized Fear-Based News) = 0.07

P(Over-generalized Conclusions) = 0.40

P(Missing Data or Incomplete Reporting) = 0.07

Misinformation Spread via Social Media: By using the data we collected, we can calculate the probability of misinformation spread via social media i.e. p(misinformation spread via social media) = (1 - (1- p(inaccurate COVID-19 series)) x (1 - p(social media amplification))) = (1 – (1- 0.05) x (1- 0.32)) = 0.354

Misinformation with Sensationalism: Similarly, the probability for misinformation with sensationalism can be calculated by p(misinformation with sensationalism) = (1 – (1 – p(exaggerated claims in headline)) x (1 - p(sensationalized fear-based news))) = (1 – (1 - 0.03) x (1 - 0.07)) = 0.0979

Hasty Generalization and Incomplete Information: similarly for probability for hasty generalization and incomplete information can be calculated by p(hasty generalization and incomplete information) = (1 – (1 - p(over-generalized conclusions)) x (1 - p(missing data or incomplete reporting))) = (1 – (1 - 0.40) x (1 - 0.07)) = 0.442

Spread of Misinformation and False Claims: For the top event to occur all the intermediate event should happen so the probability for the top event i.e. p(spread of misinformation and false claims) = p(misinformation spread via social media) x p(misinformation with sensationalism) x p(hasty generalization and incomplete information) = 0.37 x 0.10 x 0.47 = 0.0153

Therefore, that the probability of occurring of top event i.e. Spread of Misinformation and False Claims is 1.53% and the reliability of the system is (1 – probability of spread of misinformation and false claims) = 0.9847

According to this our system is 98.47% reliable

Event Tree Analysis:

Event Tree Analysis (ETA) is a forward-looking, inductive safety analysis technique used to identify and evaluate the possible outcomes following an initiating event. It models all possible event sequences and calculates the probability of each outcome based on the likelihood of intermediate events. In the context of cybersecurity and information risk, ETA can be effectively applied to analyze the propagation of misinformation and its consequences across digital platforms.

In this report, we use Event Tree Analysis to model the propagation of COVID-19-related misinformation using the CoAID (COVID-19 Healthcare Misinformation Dataset). The initiating event for our tree is the spread of misinformation, with branches representing the failure or success of key mitigating actions such as fact-checking, platform detection, and containment measures.

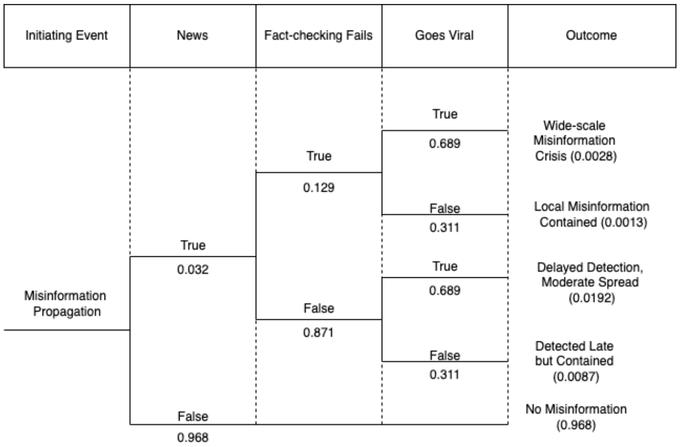

Initiating Event
News
Fact-checking Fails
Goes Viral
Outcome
Misinformation Propagation
True
0.032
False
0.968
True
0.129
False
0.871
True
0.689
False
0.311
True
0.689
False
0.311
Wide-scale Misinformation Crisis (0.0028)
Local Misinformation Contained (0.0013)
Delayed Detection, Moderate Spread (0.0192)
Detected Late but Contained (0.0087)
No Misinformation (0.968)

Fig 12: Event Tree

The event tree begins with an initiating event: Misinformation Propagation. From this point, two primary branches represent whether misinformation occurs (TRUE) or not (FALSE). If misinformation does occur, there is a 3.2% probability (0.032).

This branch further divides based on the probability of fact-checking failure. If fact-checking fails, which occurs with a probability of 12.9% (0.129), the tree continues to evaluate whether the misinformation goes viral. Based on analysis of tweet replies as a proxy for engagement, the probability that misinformation goes viral is 68.9% (0.689).

If fact-checking fails and the content goes viral, this leads to a Wide-scale Misinformation Crisis, with an overall probability of 0.0028. If the content does not go viral after a fact-checking failure, the misinformation impact is Locally Contained, with a probability of 0.0013.

If fact-checking succeeds (87.1% probability), the misinformation may still go viral. In that case, we observe Delayed Detection and Moderate Spread with a probability of 0.0192. If the misinformation does not go viral despite successful fact-checking, it results in Delayed but Contained Misinformation with a probability of 0.0087.

If misinformation propagation itself does not occur (FALSE branch, with a probability of 96.8%), there is No Misinformation impact on the system.

VI. Discussion

This risk analysis employed a variety of techniques, including Bayesian analysis**,** Poisson distribution**,** machine learning classifiers (Decision Tree, Random Forest, XGBoost), and sensitivity analysis, to quantify the risks associated with the propagation of AI-generated misinformation on social media. These methodologies allowed for the identification of key drivers of misinformation and the assessment of the effectiveness of different risk mitigation strategies.

The results of this analysis indicate that machine learning classifiers, particularly XGBoost, were able to provide high accuracy in detecting fake news articles, while the Bayesian and Poisson models offered valuable insights into the probability and frequency of misinformation events. Furthermore, the sensitivity analysis using payoff tables demonstrated how small changes in model assumptions (such as the probability of real news) can significantly impact the classification decisions.

Limitations and Threats to Validity

While the methodologies provided valuable insights, there are several limitations and threats to the validity of the results:

Data Limitations:

- The analysis relied heavily on the CoAID dataset, which focuses on COVID-19-related misinformation. This may not fully represent other types of misinformation or the broader landscape of fake news on social media. The findings might not generalize well to other domains (e.g., political misinformation, financial scams, etc.).
- The dataset also contains textual data only (titles and content), neglecting potentially influential non-textual factors such as images, videos, or user interactions. These factors could contribute to the spread of misinformation and would need to be incorporated for a more holistic analysis.

Key Considerations

Several key considerations were considered during the analysis that could affect the robustness of the results:

Binary Classification: The binary classification of news articles as either fake or real may oversimplify the complex nature of misinformation. In reality, news content may contain elements of both real and fake information, making this approach somewhat limiting.

Feature Extraction: The feature extraction approach used in the analysis relied solely on text data (i.e., titles and content). This approach may have overlooked other important factors that could enhance model performance, such as user sentiment, timing of posts, or non-textual elements like images and videos, which are often central to the spread of misinformation.

Simplified Models: The Bayesian analysis and Poisson distribution models used in the study were simplified to provide initial insights. However, these models may not fully capture the complex dynamics of misinformation spread, such as its virality or the impact of social media algorithms in amplifying certain types of content.

Future Directions

To improve the approach and address the limitations mentioned above, the following steps are suggested:

Expanding the Dataset: A broader dataset that includes a variety of topics beyond COVID-19 misinformation would help increase the generalizability of the model's findings. Incorporating multi-lingual datasets could also address biases in the current model, as misinformation occurs across different languages an**d regions.**

Incorporating Multimodal Features: Future models should incorporate multimodal features, such as images and videos, which play a crucial role in modern misinformation campaigns on social media. Integrating sentiment analysis and user engagement metrics (e.g., likes, shares, comments) could also improve the model's accuracy in detecting **misinformation.**

Advanced Models: Moving beyond traditional machine learning models, deep learning techniques like transformers (e.g., BERT) could be explored for better feature extraction and context understanding from textual data. These models can capture more nuanced patterns and contextual meaning, which could further improve classification accuracy.

Refining Bayesian and Poisson Models: Further refinement of the Bayesian network and Poisson distribution models could incorporate more complex variables such as social media dynamics, including viral spread, and timing of posts. This would help capture the stochastic nature of misinformation propagation.

Real-time Misinformation Detection: Moving towards real-time misinformation detection would be crucial for combating fake news before it spreads widely. Implementing online learning or active learning techniques could allow models to adapt to new misinformation patterns and remain accurate as the nature of fake news evolves**.**

VII. Conclusion

This paper investigates the risks associated with AI-generated misinformation propagation on social media platforms. Given the widespread impact of misinformation, especially in critical areas like public health, politics, and finance, it is crucial to understand its dynamics and develop strategies to mitigate its harmful effects.

To assess the risks, a combination of quantitative risk analysis techniques was employed, including Bayesian analysis, Poisson distribution, machine learning classifiers (Decision Tree, Random Forest, XGBoost), and sensitivity analysis. These methods enabled us to quantify the likelihood and impact of misinformation events, as well as evaluate various risk mitigation strategies. The results highlighted that machine learning classifiers, particularly XGBoost, were effective in classifying fake news with high accuracy, while the Bayesian and Poisson models provided valuable insights into the frequency and probabilities of misinformation.

The key findings of the study indicate that AI-generated misinformation poses significant risks to social media platforms, with XGBoost providing the highest accuracy at 91.24% for fake news classification. Additionally, the analysis revealed that small changes in model assumptions, such as the probability of real news, can greatly influence the classification results. The Poisson distribution modeled the daily occurrence of misinformation events, and sensitivity analysis using payoff tables helped in understanding the decision-making process in misinformation classification.

While the study provided valuable insights, there were several limitations, such as the reliance on a COVID-19-focused dataset and the use of only textual data. Future work should include more diverse datasets, incorporate multimodal features (e.g., images and videos), and apply more advanced models to improve the analysis.

In conclusion, this study contributes to the growing body of knowledge on the risks posed by AI-generated misinformation and provides a framework for developing more effective content moderation systems and risk mitigation strategies.

References

[1] M. V. Mamchenko and A. S. Rey, “Algorithm of Risk Assessment in the Task of Identifying Destructive Content in Social Networks,” *2021 International Conference on Engineering and Telecommunication (En&T)*, pp. 1–7, 2021. doi: 10.1109/ENT50460.2021.9681790.

[2] M. R. Shoaib, Z. Wang, M. T. Ahvanooey, and J. Zhao, “Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models,” *2023 International Conference on Computer and Applications (ICCA)*, pp. 1–10, 2023. doi: 10.1109/ICCA59364.2023.10401723.

[3] Squicciarini, W. McGill, G. Petracca, and S. Huang, “Early Detection of Policies Violations in a Social Media Site: A Bayesian Belief Network Approach,” *2012 IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY)*, pp. 45–52, 2012. doi: 10.1109/POLICY.2012.19.

[4] H. Alquran and S. Banitaan, “Fake News Detection in Social Networks Using Data Mining Techniques,” *2022 IEEE World AI IoT Congress (AIIoT)*, pp. 1–6, 2022. doi: 10.1109/AIIOT54504.2022.9817287.

[5] M. S. Al-Rakhami and A. M. Al-Amri, “Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter,” *IEEE Access*, vol. 8, pp. 155961–155970, Aug. 2020. doi: 10.1109/ACCESS.2020.3019600.

[6] Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan, “Misinformation Propagation in the Age of Twitter,” *Computer*, vol. 47, no. 12, pp. 90–94, Dec. 2014. doi: 10.1109/MC.2014.361.

[7] T. Wu, P. J. Phillips, and R. Chellappa, “Propagation of Facial Identities in a Social Network,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2011.

[8] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “The Spreading of Misinformation Online,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 113, no. 3, pp. 554–559, Jan. 2016. doi: 10.1073/pnas.1517441113.

[9] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “The Spreading of Misinformation Online,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 113, no. 3, pp. 554–559, Jan. 2016. doi: 10.1073/pnas.1517441113.

[10] P. L. Kharvi, “Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media,” *IEEE Security & Privacy*, vol. 22, no. 4, pp. 115–122, Jul./Aug. 2024. doi: 10.1109/MSEC.2024.3425121.