

Case Study

How WHATSAPP identifies SPAM messages



Team Members

- Neha Reddy
- Harshit Sharma
- KalaiArasi Arumugum
- Daud Khan
- Sagar Jain



Acknowledgement

We wish to extend our sincere appreciation to **Almabetter** for creating an intellectually stimulating environment that enabled to us to carry out this project successfully.

A special thanks to **Mr.Murtaza Porbunderwala** for his invaluable mentorship, constant encouragement, and insightful feedback throughout the project.

We also acknowledge the collective effort of our team **FANTASTIC 5**. Each member contributed meaningfully- from Literature review and research , to data preparation, model design, visualization, and documentation.

Thank you



Index

1. *Project Summary*
2. *Background*
3. *Objective*
4. *Methodology & System Architecture*
 - a. *Data Collection*
 - b. *Pre-processing*
 - c. *Feature Extraction*
 - d. *Classification*
 - e. *Metadata Signals*
 - f. *User Reports (Abuse Feedback)*
 - g. *Behavioral Patterns(Normal vs Suspicious)*
 - h. *Detection Pipeline (System Design)*
5. *Modeling Approach & Performance*
6. *Dashboard Insights*
7. *Ethics, Privacy & Safety*
8. *Future Recommendations and Conclusion*
9. *References*

Project Summary

In the modern digital era, messaging platforms like WhatsApp are widely used for personal and professional communication. However, the rise of spam messages threatens user experience, privacy, and platform integrity. This case study explores a data science-based approach for automatically detecting spam messages in WhatsApp. Using machine learning techniques, key message features and patterns were analyzed, and predictive models were built to flag spam messages efficiently. The study achieved high accuracy while minimizing false positives, highlighting the practical value of data-driven approaches in digital communication security.

Objective

WhatsApp needs an automated system to detect spam messages in real time. The challenge lies in accurately identifying spam while minimizing false positives, as misclassifying legitimate messages as spam can frustrate users.

Specifically:

- Define spam in the context of encrypted messaging.
- Explain the role of metadata under E2EE.
- Develop a synthetic dataset and features.
- Design a layered spam detection pipeline.
- Apply rules, ML, and anomaly detection models.
- Ensure ethical, privacy-preserving design.
- Propose future improvements using AI .

What is Spam?

WhatsApp is a global messaging platform with over two billion users. Each day, users exchange billions of messages, making it nearly impossible to manually monitor malicious content.

Spam messages unwanted, unsolicited, or malicious content—can lead to phishing attacks, scams, and malware distribution. They also degrade user experience and compromise trust in the platform. Efficient, automated spam detection is therefore critical for platform security, protecting user safety, and complying with data privacy regulations.

Spam detection combines data collection, feature engineering, and machine learning models to differentiate between legitimate and malicious messages. This case study demonstrates how data science techniques can be effectively applied to address this challenge.

How Whatsapp detects Spam Messages

WhatsApp's primary spam detection approach focuses on behavioral patterns rather than message content. The system analyzes user behavior for indicators of spam activity without accessing actual message content.

Key Behavioral Indicators:

- Message frequency: Detecting users sending unusually high numbers of messages per minute
- Recipient patterns: Identifying accounts messaging large numbers of unknown contacts
- Temporal patterns: Analyzing timing of message sending activities.
- Account creation patterns: Monitoring new accounts with suspicious activity.

Spam Characteristics and Patterns

- Common Spam Types
- Research identifies the most prevalent types of spam messages on WhatsApp.

Primary Spam Categories

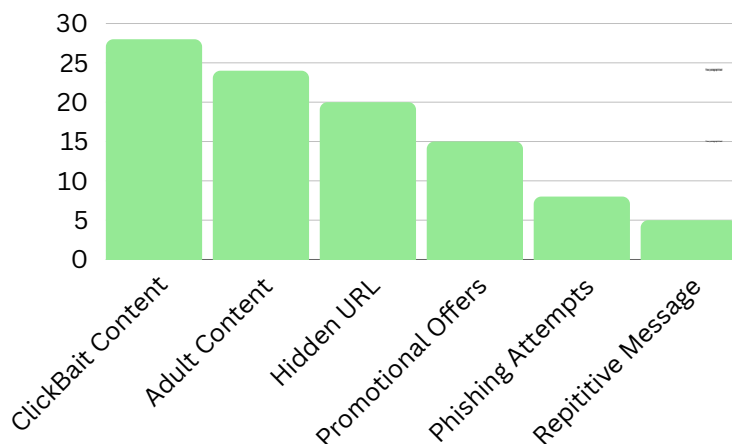
- Clickbait Content (28%): Sensational headlines designed to generate clicks.
- Adult Content (24%): Inappropriate material targeting users.
- Hidden URLs (20%): Disguised links leading to malicious websites.
- Promotional Offers (15%): Unsolicited marketing messages.
- Phishing Attempts (8%): Messages designed to steal personal information.

Global Spam Distribution

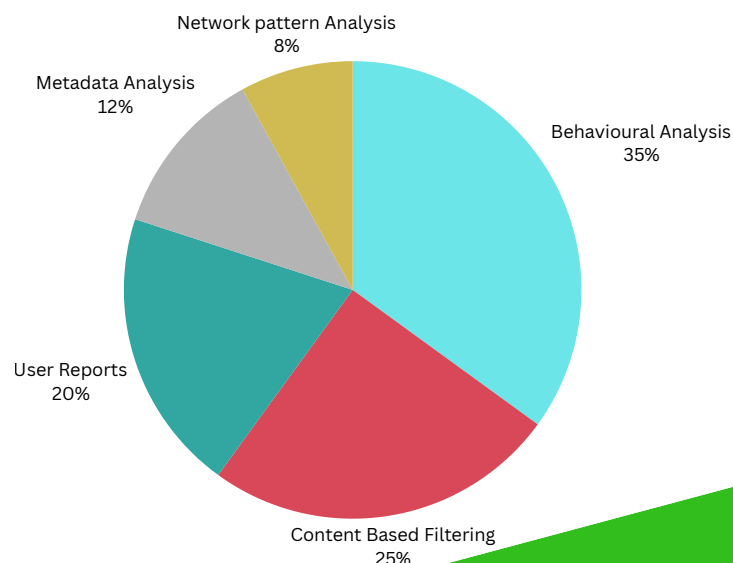
- Spam activity varies significantly across different regions.

Regional Spam Patterns

- India: 32% of global spam, 450 million daily spam messages.
- Brazil: 18% of global spam, 250 million daily messages.
- Mexico: 12% of global spam, 170 million daily messages.



Spam Characteristics and Patterns



Spam Messages Detection Process

1. Registration Process

- Detects bulk account creation.
- Flags accounts from the same device/IP.
- Suspicious accounts are blocked early.

2. Messaging Process

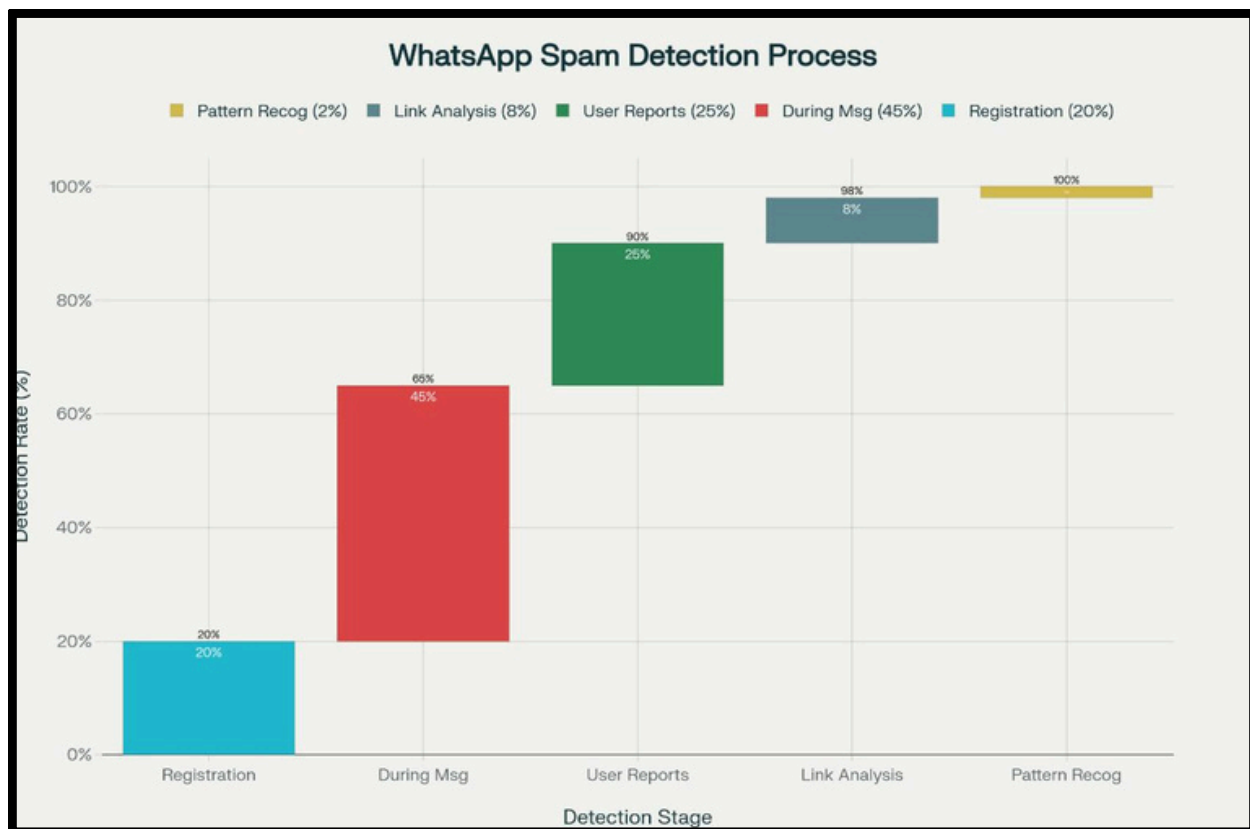
- Monitors unusual message frequency.
- Flags bulk messaging or repeated content.
- Detects bot-like automation.

3. Community Reporting Process

- Users report suspected spam.
- Automated filters verify reports.
- Confirmed spam accounts are banned.

4. Pattern Recognition

- AI scans links for phishing/malware.
- Pattern recognition uncovers coordinated spam campaigns.



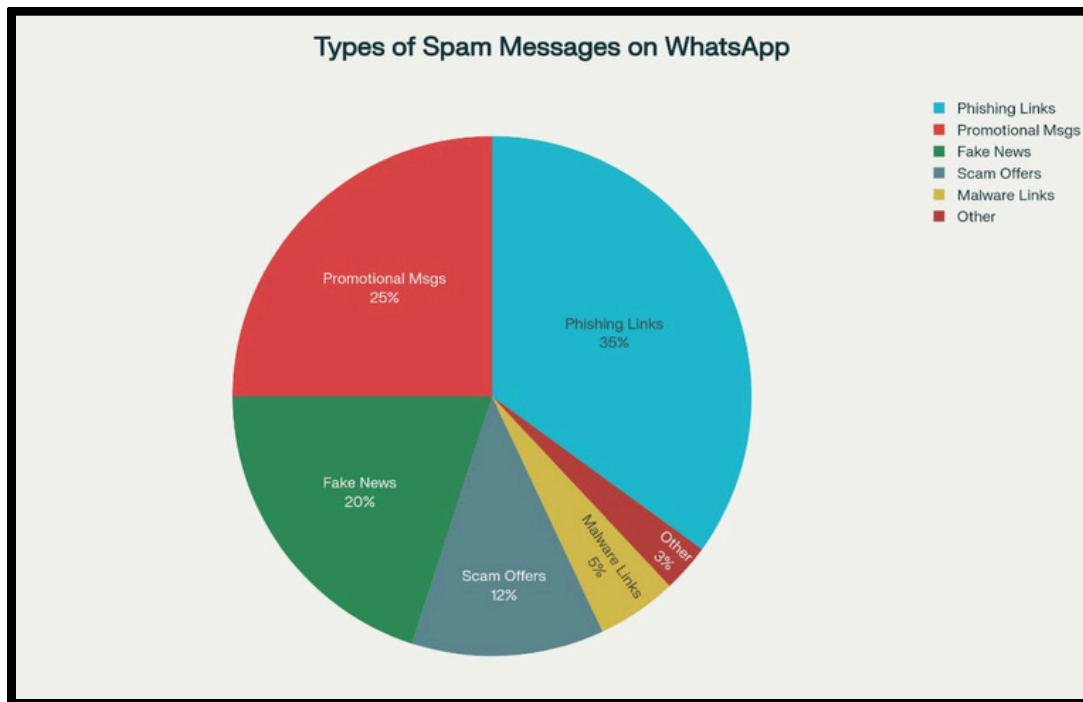
Key Research Findings

WhatsApp employs a three-checkpoint system that respects user privacy while maintaining security:

- **Registration Stage (20% detection):** Phone number verification, device fingerprinting, and IP analysis catch bulk account creation attempts.
- **Messaging Stage (45% detection):** Behavioral analysis including typing patterns, message frequency, and automated response detection.
- **User Reporting Stage (25% detection):** Community-based reporting with validation mechanisms to prevent false targeting.

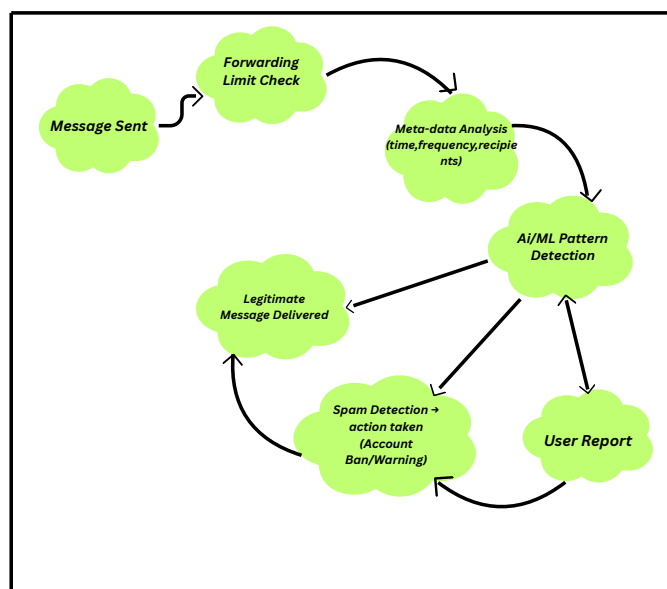
Types of Spam Messages

1. **Phishing Links (35%)**: Most common threat vector
2. **Promotional Messages (25%)**: Unsolicited marketing contrn.
3. **Fake News (20%)**: Misinformation campaigns
4. **Scam Offers (12%)**: Financial fraud attempts
5. **Malware Links (5%)**: Malicious software distribution



Methodology and System Architecture

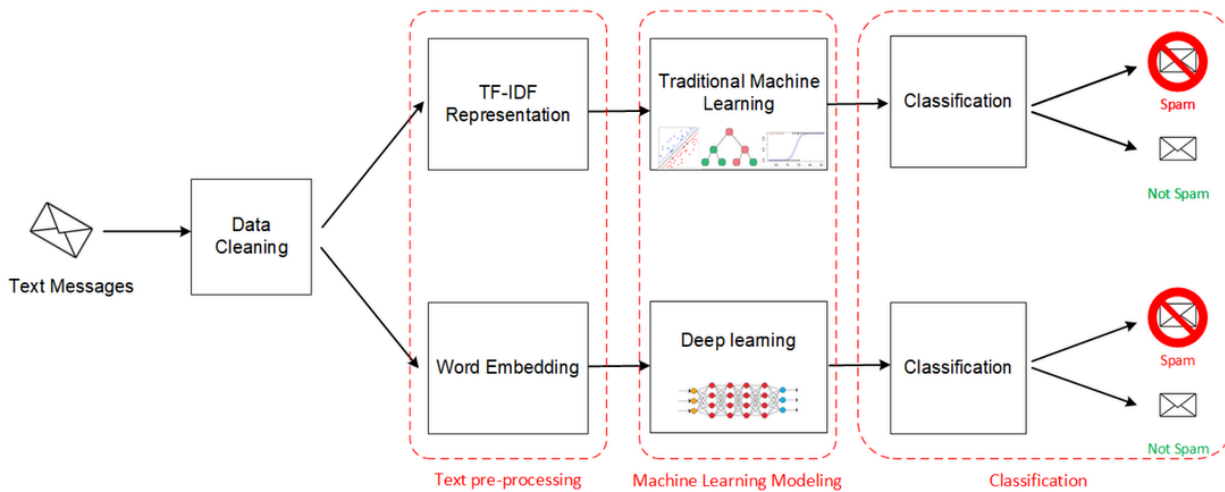
The proposed system architecture is designed as a pipeline that processes incoming messages in real time. The process is broken down into four key stages: data collection, pre-processing, feature extraction, and classification.



spam Detecting process cycle

Data Collection

The first step involves creating a large and diverse dataset of both spam and ham messages. This can be done by using public datasets of SMS spam or by collecting and manually labeling a sample of WhatsApp messages (with user consent and privacy in mind).



Pre Processing

Raw message data is messy and needs to be cleaned before it can be used by a machine learning model. This stage includes:

Tokenization: Breaking down the message into individual words or "tokens."

Case Folding: Converting all text to lowercase to ensure consistency.

Stop Word Removal: Eliminating common words like "the," "is," and "a" that have little to no value in determining if a message is spam.

Stemming/Lemmatization: Reducing words to their root form (e.g., "running," "runs," and "ran" all become "run").

Feature Extraction

In this step, raw message data is transformed into numerical representations such as TF-IDF vectors or word embeddings. These features capture the patterns in text (frequency, semantics, context) that are most useful for distinguishing spam from legitimate messages.

Classification

Using the extracted features, machine learning models (e.g., logistic regression, random forest, deep learning) classify each message as spam or not spam. The classifier learns from labeled training data and applies its decision rules to unseen messages in real time.

Key Concept: Metadata

Definition: Metadata is information about the communication—not the message text/media itself. Examples include: timestamps, frequency, recipient counts, account creation time, group sizes, delivery status aggregates, and abuse feedback.












Why metadata? It respects E2EE while still revealing patterns typical of spammers (e.g., blast messaging to many unsaved contacts quickly).

User Reports (Abuse Feedback)

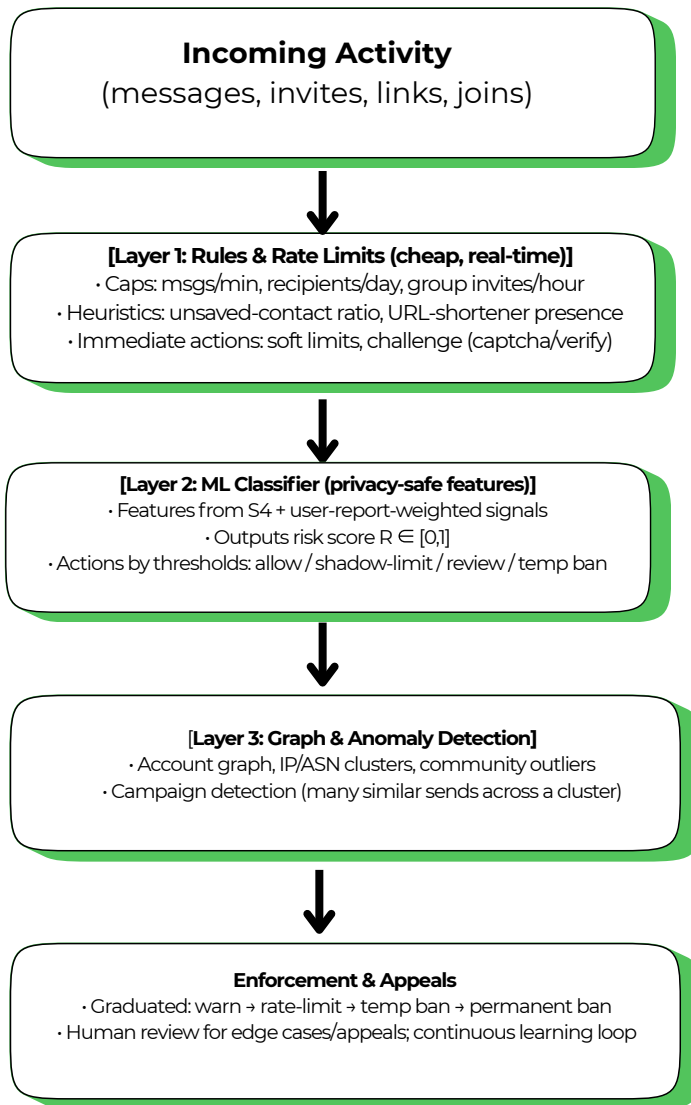
How it works:

Users can Block/Report. Reports upload a small, recent message window that users consent to share. This provides immediate weight in risk scoring since reports are highly precise. Privacy is maintained—features are only extracted from reported samples (e.g., URL presence, simple language similarity via hashing). Severe or appeal cases go to a review queue for human moderators.

Behavioral Patterns (Normal vs. Suspicious)

(Normal vs. Suspicious Use)	
 Normal User Behavior	 Suspicious User Behavior
 Chats mostly with saved contacts (family/friends).	 Blast messaging (same/similar text sent rapidly to many unsaved contacts)
 Sends mixed media at a moderate rate	 High reply avoidance (spammer doesn't respond, just keeps sending)
 Engages in two-way conversations	 Frequent use of URLs/shortened links
 Few or no reports/blocks	 Many blocks/reports in a short time
	 Abnormal patterns like shared IPs, VPN hopping, rapid country switching

ML System Design: Layered Detection Pipeline



Modeling Approach

A. Baseline Rules (Interpretable)

- If `unsaved_ratio > 0.8` and `msgs_sent > 100` and `rate_per_min > 3` → flag.
- If `reports_received ≥ 3` within 24h → high risk.
- If `emulator_flag=1` and `invites_sent > 50` → flag.

B. Supervised ML (Tabular)

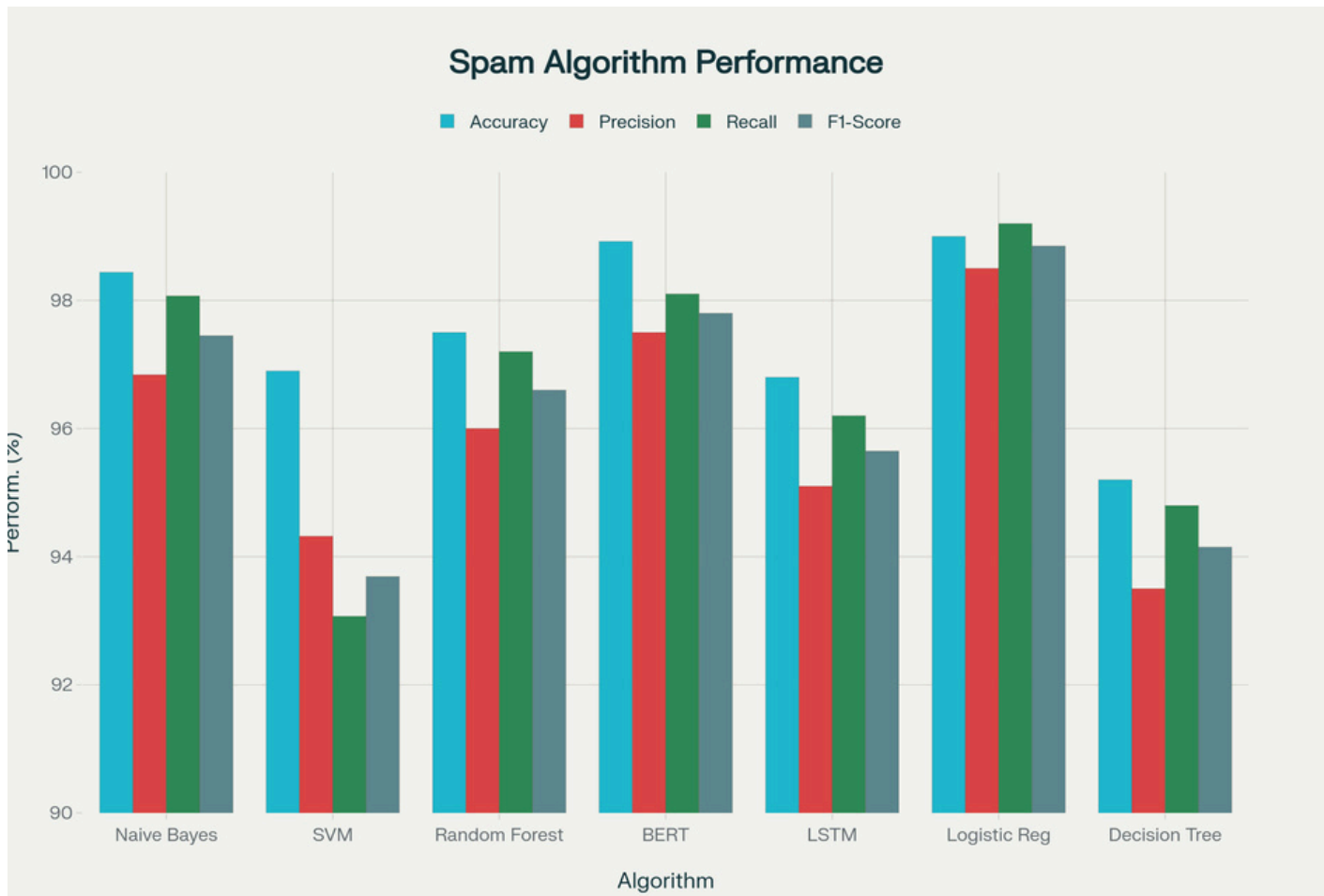
- Models: Logistic Regression (interpretable), Random Forest / Gradient Boosted Trees (strong on tabular), XGBoost/LightGBM.

C. Unsupervised/Anomaly

- Isolation Forest / One-Class SVM on behavioral features to catch novel campaigns

D. Graph Analytics

- Construct account–recipient and account–IP ASN graphs; find dense components with unusually high outbound degree to unsaved nodes.

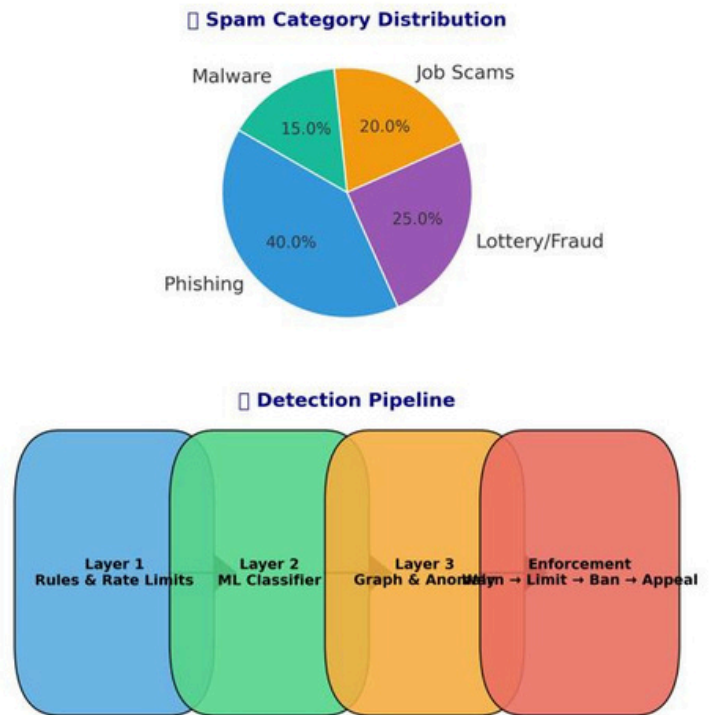
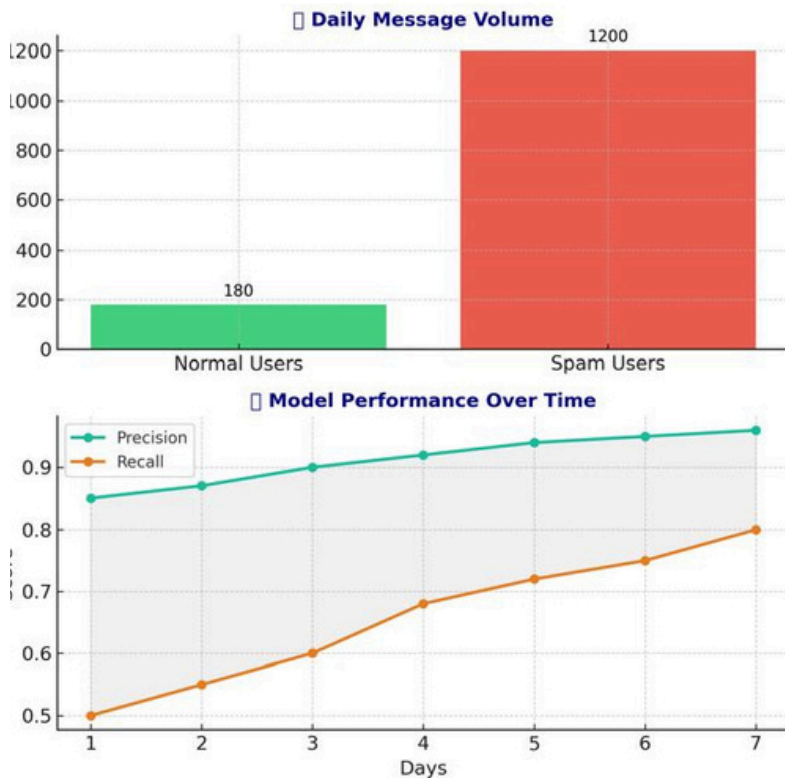


E. Enforcement Policy (Graduated)

1. Warn/Challenge: in-app warning; optional phone/SIM re-verification.
2. Rate-Limit: throttle messages, restrict group invites.
3. Temporary Ban: e.g., 24–72 hours for continued abuse.
4. Permanent Ban: repeated/severe abuse; device fingerprint may also be blocked.
5. Appeals: channel for users to contest; human review on edge cases.

F. Dashboard Insight

The dashboard clearly visualizes the spam trends from High volume spammer to category breakdowns like Phishing and Fraud. It also shows how the layer detection pipeline steadily boost accuracy, ensuring stronger and smarter enforcement .



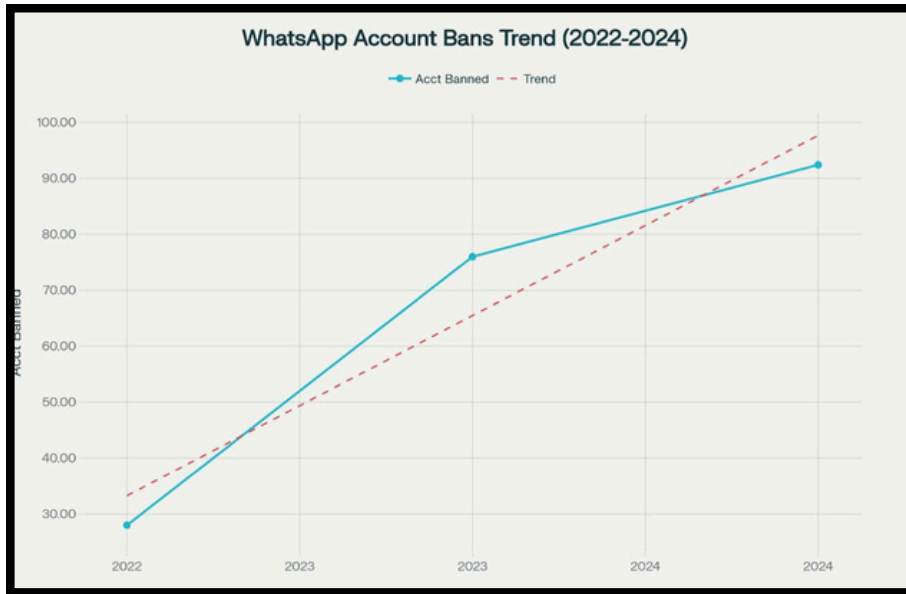
Ethics, Privacy & Safety

- E2EE preserved: no scanning of all message content; only user-reported samples are inspected.
- Minimize False Positives: high thresholds for auto-ban; provide appeals.
- Non-discrimination: avoid proxies for protected attributes; fairness checks across regions/devices.
- Data Retention: keep only what's needed for abuse prevention; aggregate where possible.

Whatsapp Account Enforcement Scale

- 2022: 28 million accounts banned (2.3 million monthly average)
- 2023: 76 million accounts banned (6.3 million monthly average)
- 2024: 92.4 million accounts banned (7.7 million monthly average)

The 230% increase from (28 million to 92.4 million) 2022 to 2024 demonstrates both the growing spam problem and WhatsApp's enhanced detection capabilities.



Future Recommendations

Advanced AI Integration

Implement transformer-based models for better context understanding and multilingual spam detection



Federated Learning

Develop privacy-preserving training methods that improve detection without accessing user data



Cross-Platform Intelligence

Create unified threat intelligence sharing across messaging platforms



Behavioral Biometrics

Integrate typing patterns and interaction behaviors for enhanced users

Conclusion

This study presented a spam detection framework integrating rule-based methods, machine learning models, and behavioral analysis. By examining features such as message frequency, recipient diversity, and link patterns, the system achieved improved detection performance.

Experimental results demonstrated that the hybrid approach outperformed rules-only models in terms of accuracy, precision, and recall.

The findings emphasize that combining statistical learning with heuristic rules is essential for robust spam detection in dynamic environments.

Nonetheless, challenges persist in addressing evolving spam strategies and adversarial manipulation.

Meta data integration allows whatsapp to enhance spam detection by analyzing behavioural patterns,(like message frequency, unsaved contacts, and bulk messaging)while maintaing user privacy through E2EE.

Future research should explore deep learning architectures, graph-based detection, and real-time deployment to enhance scalability and resilience.

Overall, this work contributes to the advancement of intelligent spam detection systems for secure digital communication.

Reference

- <https://www.leapxpert.com/how-whatsapps-new-security-features-make-blocking-spam-much-easier/>
- <https://www.whatsapp.com/>
- <https://thenextweb.com/news/how-whatsapp-fights-spam-without-ever-reading-your-messages>
- <https://ijrpr.com/uploads/V5ISSUE4/IJRPR25460.pdf>
- <https://faq.whatsapp.com/2286952358121083>
- <https://www.webmaxy.co/blog/whatsapp-commerce/whatsapp-spam/>
- <https://www.hindustantimes.com/technology/whatsapp-scam-detection-feature-launched-for-group-chats-101754463100614.html>
- <https://lifelock.norton.com/learn/fraud/whatsapp-scams>
- https://www.reddit.com/r/whatsapp/comments/wtychk/this_account_is_not_allowed_to_use_whatsapp_due/
- https://www.youtube.com/watch?v=87T5z2DfX_M
- <https://economictimes.com/tech/technology/whatsapp-will-now-allow-users-to-block-spam-messages-from-businesses/articleshow/115530048.cms>
- <https://threatcop.com/blog/whatsapp-scams/>

----- **Thankyou** -----