AdSafe: AI Based Facebook Fake Ad Verification

Case Study Overview

▼ **Background**

Social media platforms like Facebook, Instagram, and LinkedIn etc. are pivotal in digital advertising due to their massive user bases and sophisticated targeting capabilities. However, these platforms face significant challenges from fraudulent ads, which undermine user trust and platform integrity such as the ones listed below :

- **Fraudulent Financial Products**: Ads promoting fake investment opportunities or loan schemes can lead to substantial financial losses, especially for vulnerable populations who may lack financial literacy.
- **Misleading Ads and Consumer Trust**: Businesses face increased return rates and loss of sales from disillusioned customers, damaging their reputation and financial stability.
- **Data Breaches and Identity Theft**: The Penamon Institute estimates that the average cost of a data breach in 2023 was $4.45 million, highlighting the financial impact on organizations affected by such scams
- **Regulatory Penalties**: Platforms must invest in compliance measures to avoid hefty fines and legal action.
- **Content Moderation**: Developing and maintaining effective ad verification systems involves significant costs in technology and human resources.

▼ **Problem Statement**

Social media platforms need to implement an AI-driven ad verification system to detect and mitigate fraudulent ads effectively. The challenge is to create a solution that balances security with user experience, while handling the massive volume of ads and evolving tactics used by fraudsters.

▼ **Objectives**

Keeping all the core problem areas in mind we intend to develop a solution that covers the below areas:

- **Ad Verification**: Develop AI to validate the authenticity of ads in real-time.
- **User Protection**: Safeguard users from fraudulent ads with timely warnings.
- **Scalability**: Ensure the system handles a high volume of ads efficiently.
- **User Trust**: Enhance trust by reducing the incidence of fake ads.

Discovery

▼ What is Social Media Advertising

Social media advertising is when you pay to reach your target demographics on any social media platform. It's a form of digital advertising done on social media. The ad comes up organically in your target audience's social media feeds as a post or Story. But it has a "sponsored," "promoted," "boosted," or any similar label attached to disclose the post is an ad.

- 26% of users who click on Facebook ads end up buying the advertised product
- Over 10 million+ businesses use Facebook for advertising

▼ Types of Social Media Advertising

Each social platform has a unique set of ad products. Here's a list of some of the many ad types offered on each platform

▼ Extent of Social Media Usage

A.T Forbes report on May 18, 2023

The most used social media platform in the world is Facebook, with 2.9 billion monthly active users across the world

- In 2023, an estimated 4.9 billion people use social media across the world
- The social media app market in 2022 was valued at $49.09 billion
- 84% of people aged 18 to 29 use at least one social media site

▼ Scale of Advertising on Social Media

- The average CTR of ads across social media was 1.21% in 2022
- 77% of businesses use social media to reach customers

- 90% of users follow at least one brand on social media
- 76% of social media users have purchased something they saw on social media
- 3.8 million posts on Instagram had the hashtag “ad” in 2021
- Influencer spending hit $4.14 billion in 2022
- The minimum average cost of a sponsored YouTube video with 1 million views is $2,500
- The minimum average cost of an Instagram post with 1 million followers is $1,200

▼ **Fake Ad Challenges on Social Media**

A fake ad on social media is a misleading or deceptive advertisement that often promotes false or exaggerated claims, scams, or fraudulent products/services. These ads can be used to trick users into providing personal information, making fraudulent purchases, or participating in scams. They typically exploit trust and can appear convincing by mimicking legitimate brands or using persuasive tactics.

Below are few key types of challenges and statistics associated with Face Ads:

1.Digital Ad Fraud:

- **33%** of digital ad spend is lost to fraud.
- **7%** of advertisers are willing to pay an **11%** premium for authentic ad placement.
- **31%** of Android and **25%** of iOS traffic is fraudulent.
- In-app advertising has **25%** fewer fraud instances than web advertising.
- **36%** of display ad clicks, **17%** of CTV impressions, and **11%** of search ad clicks are fraudulent.
- Only **12%** of ad impressions are accurately tracked.
- **43%** of web traffic is invalid.

2.Online Scams:

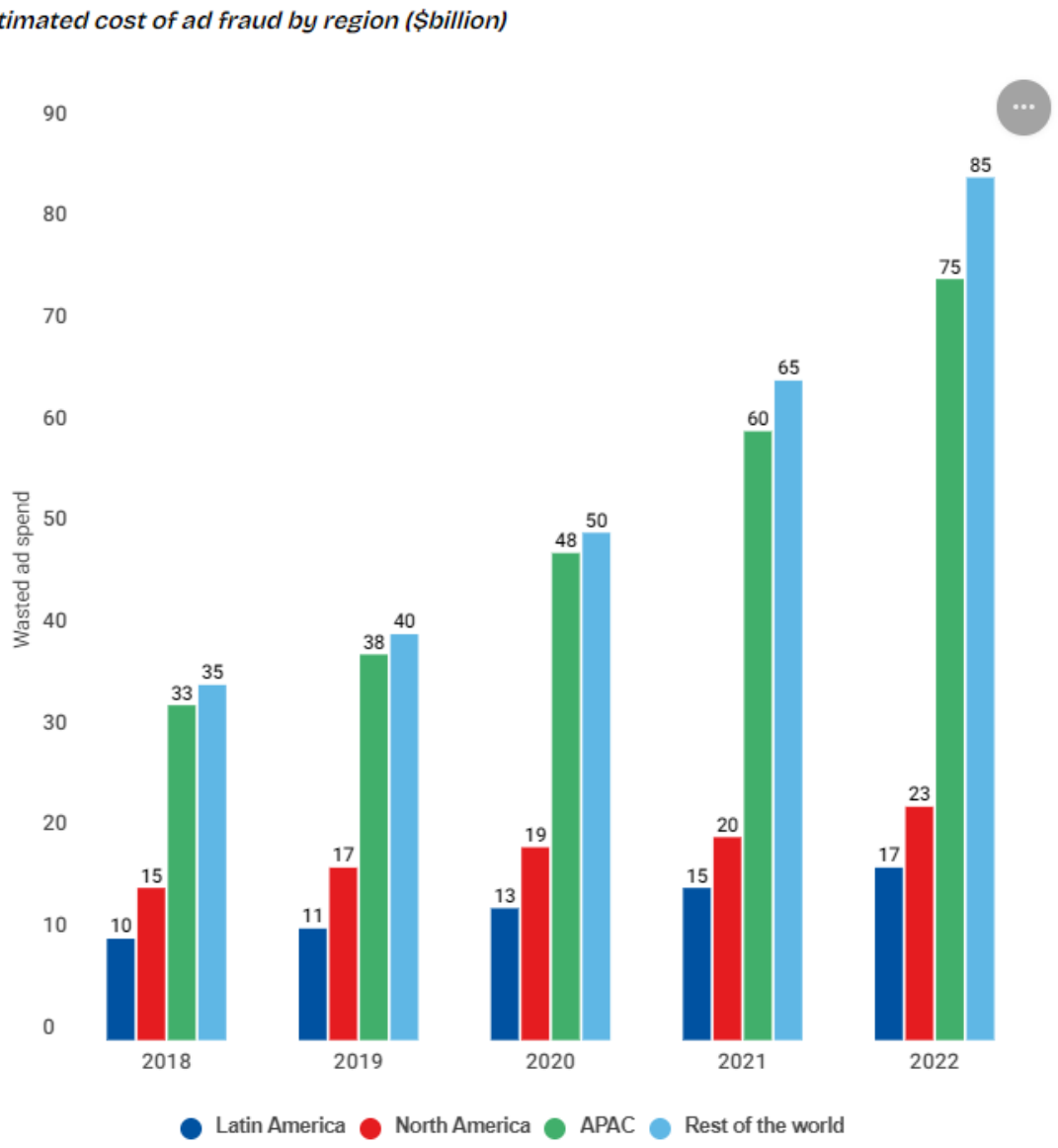
- **90%** of data breaches are caused by phishing scams.
- **30%** of people encountered job scams on social media.
- **12%** of individuals click on phishing URLs on social media.
- **85%** of teenagers and young adults have fallen victim to shopping scams.
- Inheritance scams caused over **$200,000** in losses in 2022.
- **16%** of social media users have experienced imposter scams.
- **75%** of lottery scam victims are aged 50+.
- **50%** of investment scams occur on platforms like Instagram and Facebook.
- Losses from charity scams increased by **44%** in 2022.

- **25%** of social media scams are romance scams.
- **$1.5 billion** lost due to influencer scams.

▼ **Economic Impact Due to Fake Ads**

Year	Consumer Finance (Billion $)	E-commerce (Billion $)	Cybersecurity (Billion $)	Legal and Compliance (Billion $)	Operational Costs (Billion $)
2019	1.9	0.8	2.1	0.6	1.5
2020	3.3	1.2	2.8	0.75	1.8
2021	4.5	1.5	3.2	0.9	2
2022	4.2	1.4	3.5	1	2.1
2023	4.8	1.6	3.8	1.1	2.3

Region wise Impact on Economy



▼ Why Target Facebook Ad Management

Facebook uses algorithms and artificial intelligence (AI) to detect fraudulent content. However, these systems are not perfect and can easily miss sophisticated or novel fake ads.

While Facebook sees significant usage across various countries, India has the greatest number of users. Latest reports show that there are 314.6 million Facebook users in India. The United States comes next, with 175 million users.

And Facebook is one of the leading platforms with more cybercrime concerns among social media.

Platform	Cybercrime Cases (2023)	Financial Loss (INR Crores)	Mental Trauma Cases	Users in India (2024)	Ads reach
Facebook	13,900	126	4,600	366.9 million	362 million

Platform	Cybercrime Cases (2023)	Financial Loss (INR Crores)	Mental Trauma Cases	Users in India (2024)	Ads reach
Instagram	9,500	79	2,300	362.9 million	358 million
WhatsApp	5,200	54	1,300	540 million	
YouTube	2,700	38	800	462 million	456 million
Telegram	1,600	16	450	104 million	
X (Twitter)	1,100	12	250	27 million	25.5 million
LinkedIn	500	8	100	120 million	
Snapchat	300	3	80	201 million	198 million
Total	34,800	336	9,880	-	

▼ **Current State - Facebook Ad Management**

Kindly refer the supporting document for details :
https://drive.google.com/file/d/1F5CqKntzYU-e6VCL8P-8Cx2InIWajq9e/view?usp=drive_link

▼ **Competitor Analysis**

Feature	ClickGUARD	ClickCease	TrafficGuard
Bot Detection	Advanced bot mitigation software helps identify and block suspicious bot behavior based on various factors like time on site and post-click interaction.	Bot detection capabilities, but may not have the same level of sophistication or data-driven approach as ClickGUARD.	Leverages machine learning to detect and prevent sophisticated bot-driven ad fraud
Automated Blocking	Real-time click processing platform takes immediate protective actions upon threat detection, all exclusions documented for review.	Delayed bulk exclusions, potentially allowing some fraudulent clicks to slip through before being blocked.	Real-time blocking of fraudulent traffic to prevent wasted ad spend
Real-Time Blocking	Enables real-time blocking of fraudulent clicks.	Does not offer real-time blocking.	Real-time blocking of invalid traffic to prevent wasted ad spend
IP Threat Mitigation	Identifies and blocks threats based on IP addresses.	Offers IP threat mitigation features.	Blocks invalid traffic based on IP addresses and other indicators
Behavior Analysis Mitigation	Analyzes user behavior to identify and block suspicious click patterns.	Includes behavior analysis for fraud detection.	Employs behavioral analysis to identify and block invalid traffic patterns
Device Tracking	Tracks devices used for clicks to identify patterns and potential fraud.	Tracks devices used for clicks.	Tracks device IDs and other identifiers to identify fraudulent behavior

Advanced Reporting	Advanced, scheduled, and on-demand reporting with comprehensive insights into campaign performance.	Advanced reporting features available, but may not be as user-friendly or customizable as ClickGUARD's reports.	Reporting capabilities not explicitly mentioned

Additionally even individuals today have option to safe guard from fake ads on Facebook with tools like **AdGuard** (blocks malicious ads), **Trend Micro Check** (detects scams), **McAfee WebAdvisor** (website safety ratings), and **Scamadviser** (trustworthiness checks).

Solution Coverage

▼ Key User Personas

User Personas for Facebook Ad AI Model

Sarah Thompson

Role: Data Analyst in the Ad Integrity Team

Age: 32

Experience: 5 years at Facebook, 8 years in data analysis

Goals:

- Monitor the AI model's performance in detecting fake ads
- Identify trends and patterns in fake ad attempts
- Improve the model's accuracy and efficiency

Pain Points:

- Needs to quickly identify and respond to new types of fake ads
- Struggles with balancing false positives and false negatives
- Requires clear visualizations to communicate model performance to non-technical stakeholders

Alex Rodriguez

Role: Regular Facebook user

Age: 28

Occupation: Marketing Coordinator

Goals:

- Engage with genuine and relevant ads
- Protect personal information from scams or fraudulent ads
- Have a seamless browsing experience on Facebook

Pain Points:

- Concerned about clicking on potentially harmful or deceptive ads
- Frustrated by irrelevant or spammy content in their feed
- Wants transparency about ad authenticity without disrupting their user experience

Emma Chen

Role: Small Business Owner

Age: 39

Business: Online boutique selling handmade jewelry

Goals:

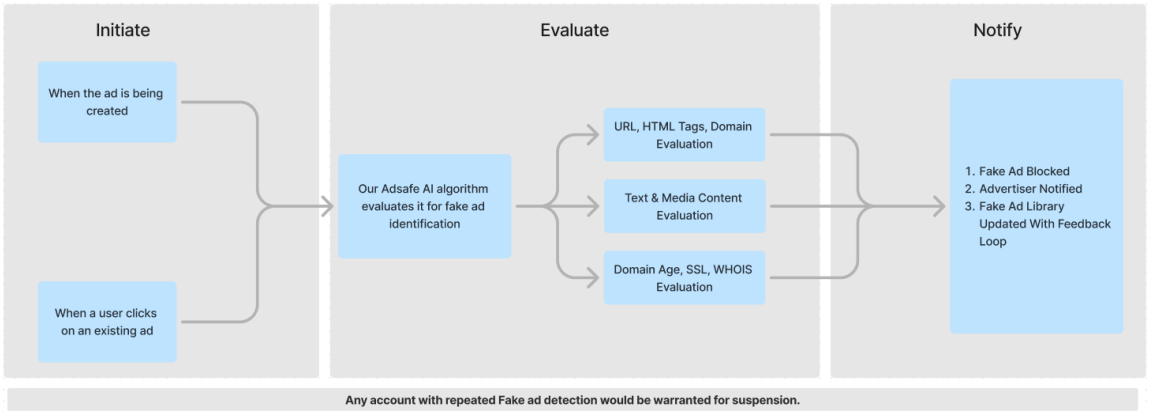
- Reach potential customers through Facebook ads
- Ensure her ads comply with Facebook's policies
- Maximize return on ad spend

Pain Points:

- Concerned about her legitimate ads being mistakenly flagged as fake
- Needs clear guidelines on what constitutes an authentic ad
- Wants to understand why an ad might be rejected or flagged

▼ Solution Outline

▼ High Level Workflow



▼ Real-Time Detection and Monitoring

Let's walk through an example to understand how the system detects fake ads using AI and ML.

1. **User Posts an Ad**:
    - **Content**: A user posts an ad with the text: "Buy Rolex brand watches at 90% discount! Click now: fake-url.com. Great offer".
    - **Action**: The user submits the ad on the platform.
2. **Collecting Text Data**:
    - The system collects the text of the ad: "Buy Rolex brand watches at 90% discount! Click now: fake-url.com". Great offer.
3. **Processing Text with AI**:
    - **Preprocessing**:
        - **Clean the Text**: Remove punctuation, convert to lowercase: "buy Rolex brand watches at 90% discounts click now fake-url com". Great offer.
        - **Remove Stop Words**: Eliminate common words that don't carry much meaning: "click, Great, offer".
    - **Tokenization**:
        - Split the text into individual words (tokens): ["buy", "Rolex", "brand", "watches", "90", "discount", "fake-url", "com"].
    - **Feature Extraction**:
        - Identify key patterns and features such as the presence of certain keywords (e.g., "discount", “Rolex”, “90%” "click", "fake-url") that are often found in fake ads.
        - Generate n-grams (combinations of words) to find common phrases used in fake ads.
    - **Website Scanner system:** Send API request with URL information, we will get elements like where website is registered, registration date etc.
4. **Analyzing with ML Models**:
    - **Training**:
        - The system has pre-trained models that have learned from past data what fake ads typically look like.
    - **Classification**:
        - The model analyzes the extracted features from the ad text.
        - The presence of words like “Rolex brand”, "90% discount" and "fake-url" are strong indicators of a potential fake ad.

    - The model calculates a score or probability that the ad is fake based on these indicators.
  - **Flagging**:
    - If the model's score exceeds a certain threshold, the ad is flagged as potentially fake.
5. **Monitoring and Verification**:
   - The flagged ad is sent for further review by the system or human moderators (optional). Moderators review the flagged ad to confirm if it's fake.
6. **Taking Action**:
   - If confirmed fake, the ad is removed from the platform.
   - The user is notified about the violation of ad policies.
7. **Improving the System**:
   - Feedback from the moderators on why the ad was confirmed as fake helps improve the AI and ML models.
   - The models are continuously updated to better detect similar fake ads in the future.

▼ **Existing Ad User Reporting system**

▼ **User Interaction**

1. **Report Initiation:**
   - **User Action:** User encounters a problematic ad and clicks the "Report Ad" button.
   - **UI:** A reporting form appears, allowing users to select a reason, provide a description, and upload evidence (e.g., screenshots).
2. **Report Submission:**
   - **User Action:** User submits the report.
   - **System Response:** The system confirms receipt and stores the report data, including ad details and user feedback.

▼ **Report Processing**

1. **Data Ingestion:**
   - **System Action:** The report is stored in the report database with associated ad metadata.
2. **Initial Analysis:**
   - **Automated Processing:** AI/ML models classify the report based on the issue type and severity.

- **Algorithms Used:**
  - **Text Analysis:** NLP models analyze report descriptions.
  - **Image Analysis:** Computer vision models evaluate evidence images.
  - **Anomaly Detection:** Identifies patterns indicating policy violations.

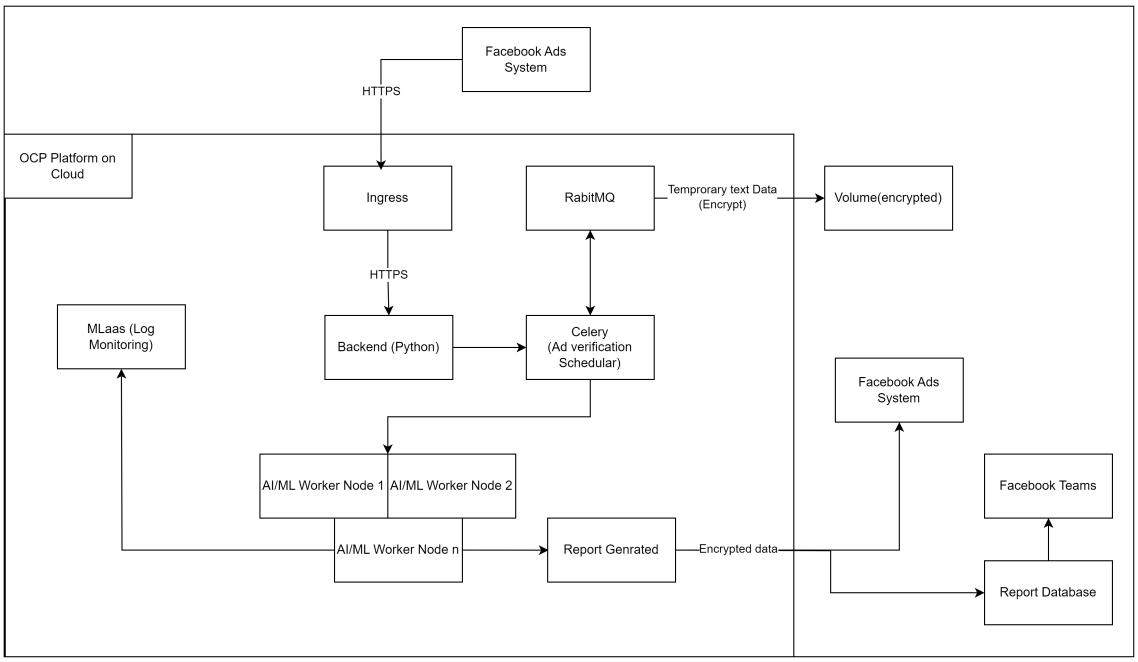
3. **Prioritization:**
   - **System Action:** Reports are prioritized for review based on classification results and severity.

▼ **Review and Action**

1. **Moderation Review:**
   - **Human Moderators (Optional):** Review high-priority reports, examine the ad, and consider user-provided evidence.
2. **Decision Making:**
   - **Actions:** Based on the review, decisions are made to remove the ad, issue a warning to the advertiser, or take no action.
3. **Report Status Update:**
   - **System Action:** Update the report status to reflect the outcome and notify the user of the resolution.

▼ **Technical Solution Overview : System Design**

To integrate this AI/ML tool with Facebook ad system for identifying fake ads, the system must handle large volumes of data efficiently, ensure high availability, and be scalable. Here's a detailed explanation of each component and the technologies used, including load balancing, global data handling, temporary data storage, and deployment using OpenShift Container Platform (OCP) and Docker.



▼ OCP Platform

To implement the AI/ML tool for identifying fake ads with Facebook using OpenShift Container Platform (OCP), we need to consider how each component integrates within OCP. This involves containerization, orchestration, monitoring, and scaling. Here's how each component fits into the OCP architecture.

▼ Facebook Ad

Source of ad data that needs to be analyzed for fake ads.

▼ Authentication

Create a system user for Facebook ad system and register it in OCP role and give editor role. Whenever any advertiser publishes any ad, facebook ad system will trigger a verification request to our AI/ML system. OCP will validate the request, if that request authentic, it will be sent further to Ingress.

▼ Ingress

Entry point for incoming API ad data from Facebook Ads System to the OCP platform via HTTPS and manages authentication.

▼ RabbitMQ

- Message broker to handle communication between services and manage temporary data storage.
- Ensures that data is queued and processed in an orderly fashion.

▼ Volume (Encrypted)

Temporary storage for test data like text, extracted data from images, urls information etc. Ensuring this data is encrypted for security purposes.

▼ **Backend**

- Handles business logic, data processing, and coordination between components.
- Receives data from Ingress and forwards it to Celery for processing.

▼ **Celery**

Distributed task queue to manage asynchronous tasks and schedule ad verification processes.

▼ **AI/ML Worker Nodes**

- Nodes dedicated to running AI/ML algorithms to analyze ad data and identify fake ads.
- Scalable to handle varying loads.

▼ **Report Generation**

- Generates reports based on the analysis performed by AI/ML worker nodes.
- These reports are then sent back to the Facebook Ads System.

▼ **MLaaS (Log Monitoring)**

Machine Learning as a Service for monitoring logs, ensuring system health and performance. This will help in case of any bug reported and manual verification need to perform what our system did for particular ad.

▼ **Report Database**

Data like how many ads scanned, is that ad passed the screening, if not what's the reason etc. this kind of data will be stored on a database. This can be used by Facebook teams via our portal.

▼ **High Level Performance of each Component**

Component	Performance Metrics
Data Collection and Ingestion	100,000 ad records/second, sub-second latency, horizontal scalability
AI/ML Model Development	95%+ accuracy, 200 ms inference time, scalable on GPUs
Load Balancing and High Availability	50 ms response time, 99.9%+ uptime, handles thousands of concurrent connections
Global Data Handling	<10 ms read/write latency, strong consistency, petabyte-scale, millions of operations/second
Temporary Data Storage	Sub-millisecond access times, 95%+ cache hit ratio, scalable to gigabytes

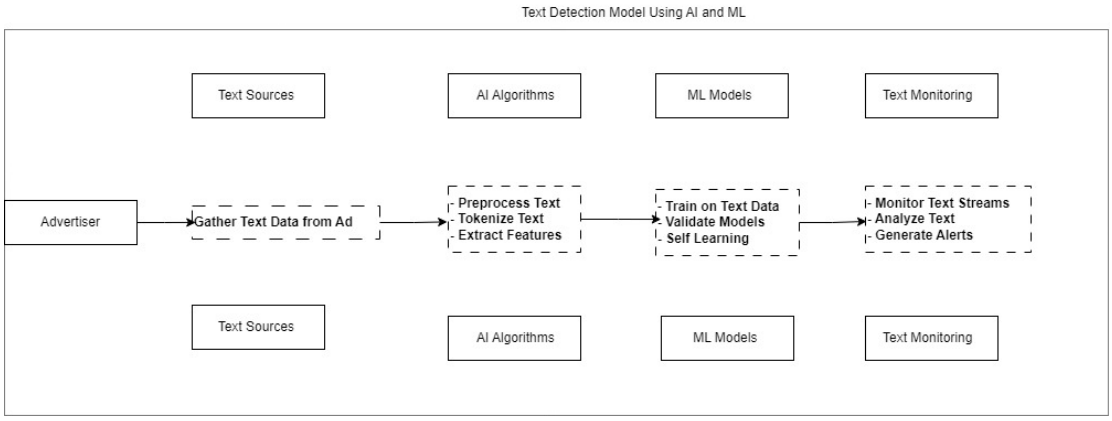
Component	Performance Metrics
Deployment using OCP and Docker	1-2 minutes deployment time, scalable to thousands of containers, efficient resource utilization
Monitoring and Maintenance	<1 second monitoring latency, <5 minutes alert response time, handles thousands of metrics/logs per second
Security and Compliance	<5% encryption overhead, sub-second access control latency, supports thousands of secure connections

▼ **Technical Solution Overview : Fake Ad Review Components**

- **Textual Features:**
  - **NLP Analysis:** Use NLP to analyze the text on the website for suspicious keywords, phrasing, and language patterns indicative of scams.
  - **Content Similarity:** Compare the content with known legitimate sites to detect plagiarism or content similarity.
- **Visual Features:**
  - **Image Analysis:** Detect the use of stolen logos, stock images, and other visual indicators of fake websites.
  - **Layout Analysis:** Analyze the layout and design quality.
- **Structural Features:**
  - **URL Analysis:** Look for anomalies in the URL structure, such as unusual subdomains or use of non-standard TLDs.
  - **HTML Tags:** Identify suspicious use of certain HTML tags and attributes that are common in fake websites.
  - Open Shorten URL with web drivers to check what kind of website it is actualluy
  - Then check that website with website check database
- **Reputation Features:**
  - **Domain Age:** Check the age of the domain; newer domains are often suspicious.
  - **SSL Certificate:** Verify the presence and validity of an SSL certificate.
  - **WHOIS Information:** Analyze WHOIS data for patterns common in fake websites.

▼ **Technical Solution Overview : Fake Ad Evaluation Models**

▼ **Text Detection**



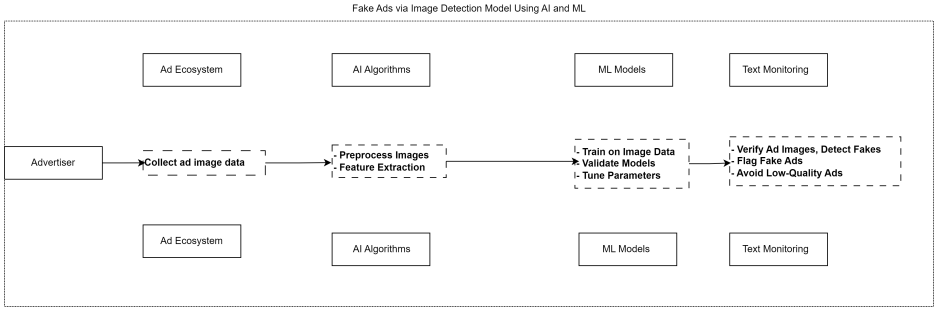
1. **Text Sources**:
    - **Gather Text Data**: Collect text data from various sources such as ad heading, description, user's comments from ad, etc.
    - **Provide Text Data**: Text data is sent to the AI algorithms for preprocessing and analysis.
2. **AI Algorithms**:
    - **Preprocess Text**: Clean and prepare the raw text data.
        - Remove punctuation and special characters.
        - Convert text to lowercase.
        - Remove stop words.
        - Apply stemming or lemmatization.
        - Stemming: This simple form of word reduction focuses on removing word endings (suffixes) to obtain a base form, often resulting in non-dictionary words.
        - lemmatization: Lemmatization goes beyond truncating words and analyzes the context of the sentence, considering the word's use in the larger text and its inflected form. After determining the word's context, the lemmatization algorithm returns the word's base form (lemma) from a dictionary reference.
    - **Tokenize Text**: Break down the preprocessed text into smaller units (tokens).
        - Word tokenization.
        - Sentence tokenization.
    - **Extract Features**: Extract meaningful information from the tokens.
        - Generate n-grams.
        - Identify part-of-speech tags.

    - Recognize named entities.
- **Send Features**: Send the extracted features to the ML models for training.

3. **ML Models**:
    - **Train on Text Data**: Use the features extracted from the text data to train the machine learning models.
        - Implement supervised, unsupervised, or semi-supervised learning as needed.
    - **Validate Models**: Assess model performance using validation datasets.
        - Use metrics like accuracy, precision, recall, and F1 score.
    - **Tune Parameters**: Adjust hyperparameters to improve model performance.
        - Apply techniques like grid search or random search.
    - **Deploy Models**: Deploy the trained models to the text monitoring system.
4. **Text Monitoring**:
    - **Monitor Text Streams**: Continuously observe incoming text data in real-time or near-real-time.
        - Detect and flag suspicious or anomalous text patterns.
    - **Analyze Text**: Use deployed ML models to analyze the monitored text streams.
        - Identify potential risks, frauds, or other concerns.
    - **Generate Alerts**: Create alerts or notifications when suspicious text is detected.
        - Ensure timely intervention and action.
5. **User/Publisher**:
    - **Review Flagged Texts**: Human experts or automated systems review the texts flagged as suspicious.
        - Validate the flagged texts for authenticity and relevance.
    - **Take Actions**: Decide on the necessary actions based on the review outcomes.
        - Block or moderate content.
        - Alert authorities if needed.
        - Update system rules and models based on review feedback.
6. **Feedback Loop**:
    - **Refine Models**: Use feedback from user reviews and actions to refine AI algorithms and ML models.
        - Incorporate new insights and patterns.

- Improve accuracy and reliability of text detection over time.

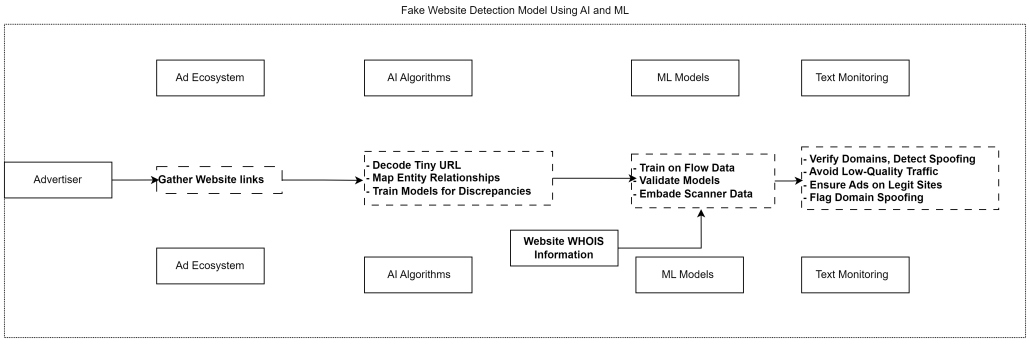
▼ **Images Detection**



- **Ad Ecosystem**:
  - **Gather Ad Image Data**: Collect ad image data from various sources such as ad networks, publishers, and traffic logs.
  - **Provide Ad Image Data**: Send ad image data to the AI algorithms for preprocessing and analysis.
- **AI Algorithms**:
  - **Preprocess Images**: Clean and prepare the raw image data.
    - **Image Resizing**: Standardize image sizes.
    - **Normalization**: Adjust pixel values for consistent scaling.
    - **Augmentation**: Apply transformations like rotation, flipping, etc., to create diverse training data.
  - **Feature Extraction**: Extract meaningful features from images.
    - **Edge Detection**: Identify and highlight edges within the image.
    - **Color Histogram**: Analyze color distribution.
    - **Texture Analysis**: Evaluate the surface texture of the images using NLP.
  - **Send Features for Analysis**: Send the extracted features to the ML models for further analysis.
- **ML Models**:
  - **Train on Image Data**: Use the features extracted from the image data to train machine learning models.
    - Implement supervised, unsupervised, or semi-supervised learning as needed.
  - **Validate Models**: Assess model performance using validation datasets.

        - Use metrics like accuracy, precision, recall, and F1 score.
    - **Tune Parameters**: Adjust hyperparameters to improve model performance, employing techniques like grid search or random search.
    - **Deploy Models**: Deploy the trained models to the image monitoring system.
- **Image Monitoring**:
    - **Verify Ad Images, Detect Fakes**: Continuously monitor ad images and use ML models to verify and detect fake or misleading images.
    - **Flag Fake Ads**: Identify and flag ads that exhibit suspicious behavior or appear to be fake.
    - **Avoid Low-Quality Ads**: Ensure that low-quality or fraudulent ads are avoided.
    - **Ensure Legit Ads**: Confirm that ads are legitimate and meet the required standards.
- **Advertiser**:
    - **Review Flagged Ads**: Advertisers review the ads flagged as suspicious.
    - **Take Actions**: Based on the review, necessary actions are taken, such as blocking fraudulent ads or updating targeting rules.
    - **Continue Image Monitoring**: Maintain ongoing monitoring to ensure continued protection against fake ads.
- **Feedback Loop**:
    - **Refine Models**: Use feedback from the advertiser's actions to refine AI algorithms and ML models.
    - **Improve Accuracy**: Continuously enhance the accuracy and reliability of the fake ad detection system.

▼ **Fake websites Detection**



1. **Ad Ecosystem**:

- **Gather Website Data**: Collect website data from various sources such as ad network logs, URL's, SEO words, HTML Format, Website scanner system etc.
- **Website WHOIS Information**: Check the age of the domain; newer domains are often suspicious, verify the presence and validity of an SSL certificate, analyze data for patterns common in fake websites.
- **Provide Website Data**: Send website data, **WHOIS information** to the AI algorithms for preprocessing and analysis.

2. **AI Algorithms**:
   - **Map Entity Relationships**: Analyze relationships between different entities in the ad ecosystem.
   - **Train Models for Flow Discrepancies**: Identify patterns and train models to detect discrepancies in web traffic flows.
   - **Send Flow Data for Analysis**: Processed data is sent to the ML models for further analysis.
3. **ML Models**:
   - **Train on Flow Data**: Use the flow data to train machine learning models to detect anomalies and potential spoofing.
   - **Validate Models**: Assess model performance using validation datasets, focusing on metrics like precision and recall.
   - **Tune Parameters**: Adjust hyperparameters to improve model performance, employing techniques like grid search or random search.
   - **Deploy Models**: Deploy the trained models to the traffic monitoring system.
4. **Traffic Monitoring**:
   - **Verify Domains, Detect Spoofing**: Continuously monitor web traffic and use ML models to verify domains and detect spoofing.
   - **Flag Domain Spoofing**: Identify and flag domains that exhibit suspicious behavior or spoofing activities.
   - **Avoid Low-Quality Traffic**: Ensure that low-quality or fraudulent traffic is avoided.
   - **Ensure Ads on Legit Sites**: Confirm that ads are placed on legitimate sites.
5. **Advertiser**:
   - **Review Flagged Domains**: Advertisers review the domains flagged as suspicious.
   - **Take Actions**: Based on the review, necessary actions are taken, such as blocking fraudulent sites or updating targeting rules.

- **Continue Flow Monitoring**: Maintain ongoing monitoring to ensure continued protection against fake websites.

6. **Feedback Loop**:
   - **Refine Models**: Use feedback from the advertiser's actions to refine AI algorithms and ML models.
   - **Improve Accuracy**: Continuously enhance the accuracy and reliability of the fake website detection system.

▼ **Reporting, Compliance & Continuous Improvement**

1. **Trend Analysis:**
   - **System Action:** Aggregate and analyze report data to identify trends and patterns.
   - **Algorithms Used:** Statistical and machine learning models for trend analysis and anomaly detection.
2. **Insights and Alerts:**
   - **Dashboard:** Provides visualizations of report metrics and alerts on emerging issues.
3. **Policy Improvement:**
   - **Action:** Use insights to refine ad policies and improve the reporting process.

Compliance and Security

1. **Data Protection:**
   - **Encryption:** Encrypt report data both in transit and at rest.
   - **Access Control:** Implement role-based access to ensure data security.
2. **Audit Trails:**
   - **Logging:** Maintain logs of all actions taken on reports for accountability and compliance.

Feedback Loop

1. **Continuous Improvement:**
   - **System Updates:** Regularly update AI/ML models and reporting processes based on user feedback and new insights and continues to update the Facebook Fake Ad Library for enhanced coverage in subsequent runs.

▼ **Future Enhancements : Video Detection**

In our upcoming Version 2 (V2) of the Fake Ad Detection System, we plan to extend our capabilities to include video advertisements. Recognizing the high computational demands of video analysis, we will focus on optimizing both performance and accuracy through advanced techniques and strategic approaches.

Key Features of V2:

1. **Frame-by-Frame Analysis:** We will decompose video ads into individual frames, allowing us to process and analyze each frame independently.
2. **Efficient Pixel Inspection:** Advanced algorithms will be employed to inspect each pixel and detect anomalies that may indicate the presence of fake content. We will leverage GPU acceleration to handle the high computational load, ensuring faster processing times.
3. **AI and ML Integration:** Our system will integrate state-of-the-art artificial intelligence (AI) and machine learning (ML) models specifically trained for video content.

By incorporating these features, V2 of our Fake Ad Detection System will offer a comprehensive and powerful solution for identifying fake advertisements in video content. This advancement will significantly enhance our ability to protect users from deceptive ads and maintain the integrity of advertising platforms.

▼ **User Experience & Trust Management**

We will introduce a new Facebook page dedicated to keeping users informed about the latest and most popular scams circulating online and in everyday life. Our mission is to educate and empower our community by providing regular updates, insightful tips, and expert advice on how to recognize and avoid fraudulent schemes.

Key Features:

- **Timely Updates:** Stay informed with real-time alerts on new and trending scams, ensuring you're always one step ahead of potential threats.
- **Educational Content:** Access a wealth of information on various types of scams, including phishing, fake ads, financial fraud, and more, with detailed explanations on how they operate and how to avoid them.
- **Community Engagement:** Join discussions with fellow community members, share your experiences, and learn from others' encounters with scams.
- **Expert Advice:** Benefit from expert insights and recommendations on best practices for staying safe online and offline.
- **Resource Links:** Find helpful resources and links to official websites for reporting scams and seeking assistance if you fall victim to fraud.

Follow our page to stay vigilant and protect yourself and your loved ones from the ever-evolving landscape of scams and frauds. Together, we can build a safer and more informed community.

▼ **Technical Consideration For System Success**

Kindly refer the supporting document for details : https://drive.google.com/file/d/1F5CqKntzYU-e6VCL8P-8Cx2InlWajq9e/view?usp=drive_link

▼ **Non Functional Considerations**

Category	Requirement
Performance	Scalability to handle, 200 ms response time, 100 verifications/sec throughput
Reliability	99.9% uptime, fault tolerance with zero data loss
Security	AES-256 and TLS 1.2+ encryption, RBAC, comprehensive audit logging
Usability	User-friendly interface, user feedback mechanisms
Maintainability	High code quality, comprehensive automated testing
Interoperability	RESTful APIs, support for standard data formats (JSON, XML)
Compliance	GDPR and CCPA compliance, detailed audit trails
Extensibility	Modular architecture, support for plugins and extensions

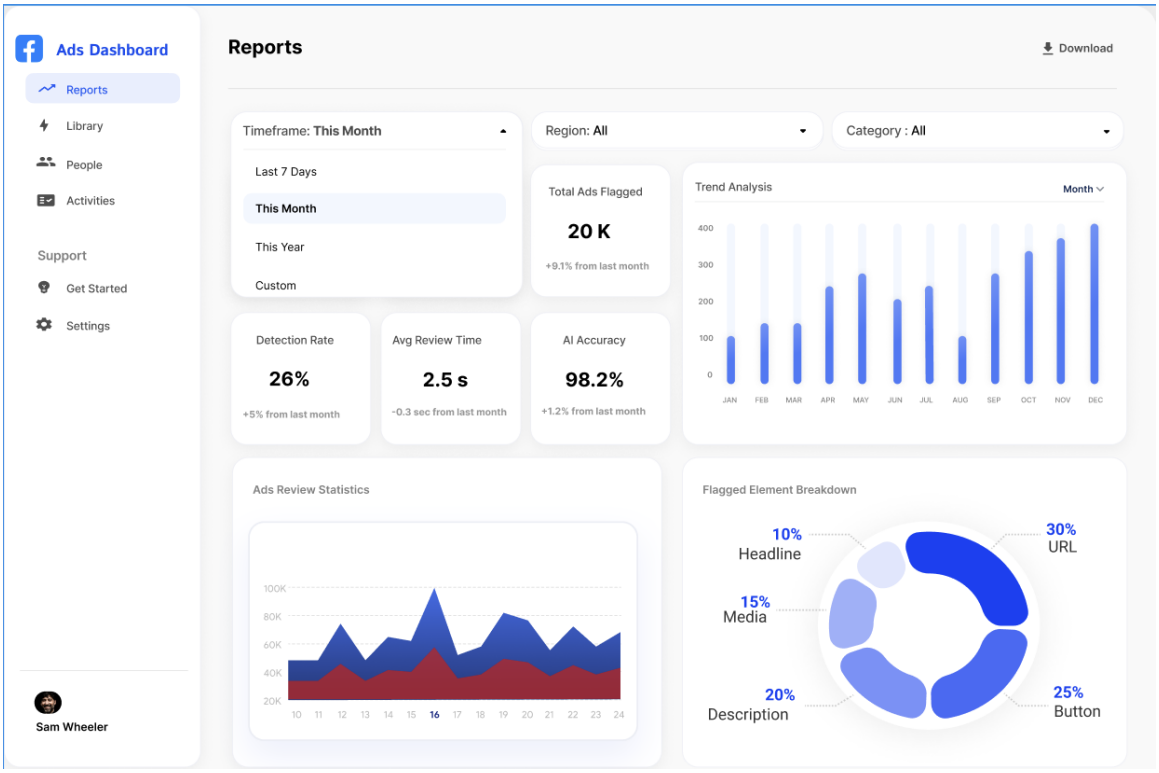
▼ Mockups and Prototypes

▼ Design Mockups

User-side Mockup (Facebook app) - https://www.figma.com/design/oWrRloEhJgOqzSCvtMZQTN/Untitled?node-id=0-1&t=99m1J7QM1xDLyJWj-1-

Ad-portal Prototype (Facebook ad manager) - https://www.figma.com/design/8YG7dCNDZLaH1e3mxEA4jv/Untitled?node-id=0-1&t=xQ7MCOxZL4NgrHTo-1--

AdSafe Facebook Manager Dashboard



▼ **Prototypes**

User-side Prototype (Facebook app) - https://www.figma.com/proto/oWrRloEhJgOqzSCvtMZQTN/Untitled?node-id=11-3961&t=eHlHiqLgQeJcTaiu-1&scaling=scale-down&content-scaling=fixed&page-id=0%3A1&starting-point-node-id=1%3A296

Ad-portal Prototype (Facebook ad manager) - https://www.figma.com/proto/8YG7dCNDZLaH1e3mxEA4jv/Untitled?node-id=103-5&t=CnNUDDGy1jKRm7Fm-1&scaling=scale-down&content-scaling=fixed&page-id=0%3A1&starting-point-node-id=29%3A2

▼ **Risks & Mitigation Approach**

Risk	Potential Impact	Mitigation Strategies
Revenue Loss	Reduced ad revenue, decreased advertiser trust	Enhance detection accuracy, provide detailed reports, attract new advertisers, implement penalties
High Computational Needs	Increased costs, system performance degradation	Optimize algorithms, utilize GPU acceleration, leverage cloud services, implement cost management
Need for New Infrastructure	Increased costs, deployment delays	Incremental deployment, cloud infrastructure, automated deployment, partner with cloud providers
System Downtime and Reliability Issues	Revenue loss, user dissatisfaction	Redundancy and failover, regular maintenance, disaster recovery plan
Compliance and Privacy Concerns	Legal issues, reputational damage	Data encryption, compliance audits, clear consent mechanisms, privacy policy updates
High False Positives	Advertiser dissatisfaction, reduced ad revenue	Refine detection algorithms, continuous model training, user feedback mechanisms
Impact on User Experience	Decreased user satisfaction, churn	Optimize system performance, enhance user interfaces, gather user feedback
Evolving Regulations	Legal and financial risks	Regulatory updates, legal expertise, flexible policies
Collaboration Challenges	Delayed development, suboptimal solutions	Strong partnerships, clear communication, joint efforts

▼ **Performance Metrics**

- **North Star Metric**

The North Star Metric for a fake ad detection product should be the one that most closely aligns with Facebook's core business objective of delivering value to users. In this case, that metric is:

Decrease in fake ad impressions

This metric directly measures the product's effectiveness in reducing the number of fake ads that users see on Facebook. A significant decrease in fake ad impressions indicates that the product is successfully protecting users from exposure to harmful or misleading content.

By focusing on the decrease in fake ad impressions, the fake ad detection product team can ensure that they are developing a product that is aligned with Facebook's core business objectives and that is delivering value to users.

Metric	Description	Formula	Good Performance	Example Value
Detection Performance	False Positive Rate (FPR)	Rate of legitimate ads flagged as fake (lower is better)	< 1%	Not Applicable
Detection Performance	False Negative Rate (FNR)	Rate of fake ads missed (lower is better)	< 5%	Not Applicable
Detection Performance	True Positive Rate (TPR) or Sensitivity	Rate of fake ads correctly identified (higher is better)	> 95%	Not Applicable
Detection Performance	True Negative Rate (TNR) or Specificity	Rate of legitimate ads correctly identified (higher is better)	> 99%	Not Applicable
Product Effectiveness	Number of fake ads detected	Total number of fake ads identified		10000
Product Effectiveness	Number of fake ads removed	Total number of fake ads removed from Facebook		8000
Product Effectiveness	Time to detect	Average time to identify a fake ad (lower is better)		30 minutes
Product Effectiveness	Time to remove	Average time to remove a flagged ad (lower is better)		1 hour
Impact Metrics	Decrease in fake ad impressions		> 80%	Not Applicable
Impact Metrics	Decrease in fake ad clicks	Percentage reduction in fake ad clicks (higher is better)	> 70%	Not Applicable
Operational Metrics	System uptime	Percentage of time the product is operational (higher is better)	> 99.5%	Not Applicable
Operational Metrics	Processing time per ad	Average time to process an ad (lower is better)		50 milliseconds

Detection System	Performance Metrics	Efficiency Metrics	Resource Utilization Metrics	Other Key Metrics
URL	Accuracy, Precision, Recall, F1 Score, AUC-ROC	Response Time, Throughput, Scalability	CPU Usage, Memory Usage	False Positive Rate, False Negative Rate
Image	Accuracy, Precision, Recall, F1 Score, IoU, mAP	Inference Time, Throughput, Scalability	GPU Usage, Memory Usage	False Positive Rate, False Negative Rate
Text	Accuracy, Precision, Recall, F1 Score, BLEU Score, ROUGE Score	Processing Time, Throughput, Scalability	CPU Usage, Memory Usage	False Positive Rate, False Negative Rate

▼ **Launch Roadmap**

Below is the Phase wise Road Map plan and Success Metrics:

▼ Phase 1: Research and Planning (Q1 2025)

Milestone	Details	Version
Market Research	Conduct market research to understand types of ads and common scam patterns.	V1.0
User Feedback Collection	Survey users to gather feedback on ad-related issues and expectations.	V1.0
System Design	Design the AI system architecture, including data sources, algorithms, and feedback loops.	V1.0
Scalability Planning	Assess infrastructure requirements for scaling the system.	V1.0

▼ Phase 2: Development and Testing (Q2 2025)

Milestone	Details	Version
AI Algorithm Development	Implement AI algorithms for ad verification using ML and NLP techniques.	V1.1
User Feedback and Alert Systems	Develop systems for real-time data collection and notifications.	V1.1
Manual Moderation Workflow	Develop workflows for manual moderation to handle edge cases.	V1.1
Initial Testing	Conduct unit testing, integration testing, and system testing.	V1.2
User Acceptance Testing (UAT)	Perform UAT with a small group of users.	V1.2

▼ Phase 3: Pilot Launch (Q3 2025)

Milestone	Details	Version
Limited Rollout	Launch the system to a subset of users or specific region.	V1.3
Performance Monitoring	Track key metrics: detection accuracy, false positive rate, user engagement, scam reduction.	V1.3
Issue Resolution	Identify and resolve issues during the pilot phase.	V1.3
User Training and Support	Provide training materials and support for users.	V1.3

▼ Phase 4: Full Launch and Expansion (Q4 2025 - Q1 2026)

Milestone	Details	Version
Full System Rollout	Launch the system to the entire user base.	V2.0
Infrastructure Scaling	Ensure infrastructure can handle increased load.	V2.0
Continuous Improvement	Gather user feedback and refine AI models and system features.	V2.1
Regular System Updates	Regularly update system to adapt to new scam patterns and improve verification accuracy.	V2.1
Scalability Enhancements	Optimize system performance for efficiency and reliability.	V2.1

▼ Phase 5: Feature Expansions and Enhancements (Q2 2026 - Q4 2026)

Milestone	Details	Version
Advanced AI Features	Develop and integrate advanced AI features such as deeper learning models and enhanced NLP.	V3.0
Enhanced User Interface	Improve the user interface based on feedback and usability testing.	V3.0
Additional Ad Verification Methods	Integrate additional verification methods (e.g., image recognition, video analysis).	V3.1
Multi-Language Support	Expand the system to support multiple languages for global reach.	V3.2
Mobile App Development	Develop mobile app versions for ad verification on-the-go.	V3.3

▼ Phase 6: Global Expansion and Ecosystem Integration (Q1 2027 - Q4 2027)

Milestone	Details	Version
Global Market Entry	Launch the system in new international markets.	V4.0
Integration with Ad Platforms	Partner with major ad platforms for seamless integration.	V4.1
API Development	Develop APIs for third-party integrations.	V4.2
Ecosystem Expansion	Expand the ecosystem with additional tools and features for advertisers and users.	V4.3

▼ **Growth Strategy**

Foundation Building

- **Identify and onboard early adopter advertisers:** We shall select top 20% advertisers based on ad spend and engagement.
- **Develop comprehensive educational resources:** Then create a dedicated training module and supporting materials.
- **Conduct a pilot program:** We will test the feature with 50 early adopters for two months to gather data on performance and user experience.

Scaling and Adoption

- **Gradually expand feature availability:** We will then Increase user base by 20% every month until full rollout.
- **Implement incentive programs:** Additionally we will offer a 10% ad credit for advertisers reducing fake ad impressions by 50% within three months.
- **Showcase successful advertiser campaigns:** To increase advocacy we will highlight case studies of advertisers achieving a 30% increase in ROI after implementing the feature.
- **Ensure seamless integration:** For smooth functioning we will integrate the feature into the ad manager dashboard for a 25% reduction in user onboarding time.

Consumer Focus

- **Launch a public awareness campaign:** Our would be to invest 20% of the marketing budget in consumer-facing campaigns.
- **Publish regular transparency reports:** We will also release reports quarterly detailing a percentage reduction in fake ads.
- **Encourage user-generated content:** In addition to that we may offer incentives for users sharing positive experiences (e.g., $10 gift cards for 100+ shares).
- **Actively engage with users:** Also we will respond to user inquiries and complaints within 24 hours.

Continuous Improvement

- **Gather and implement feedback:** We will continually conduct user surveys every quarter to identify improvement areas.
- **Develop new features:** As we progress we shall introduce advanced features based on top 3 user requests.
- **Expand fake ad management:** To increase value we shall extend the feature to Instagram and Messenger within six months.

Monetization and Revenue Generation (Future Enhancements)

In lieu of positive adoption & traction around ad safety model's execution success we would go ahead with a monetization approach with below considerations:

- **Premium Verification Service:** Offer advertisers a premium verification service for an additional fee of 10% of ad spend. For instance, an advertiser spending $100,000 would pay an extra $10,000 for premium verification.
- **Performance-Based Pricing:** Charge advertisers a fee of $500 per 1000 fake ads removed. For example, if an advertiser has 5000 fake ads removed, they would pay $2500.
- **Data Insights and Analytics:** Offer a monthly subscription-based service starting at $500 for basic analytics and increasing to $2,000 for advanced insights.
- **Ad Quality Score:** Implement a tiered pricing model where advertisers with higher ad quality scores receive a 10% discount on ad costs. For a $100,000 ad campaign, an advertiser with a high quality score would pay $90,000.

Pricing Model	Price Points	Features
Freemium Model	Free, Premium: $50/month per user	Free: Limited verifications, basic reporting, community support; Premium: Unlimited verifications, advanced reporting, priority support
Subscription-Based	Basic: $100/month, Pro: $250/month, Enterprise: Custom	Basic: Up to 1,000 verifications, standard reporting, email support; Pro: Up to 10,000 verifications, advanced analytics, priority support; Enterprise: Unlimited verifications, full analytics, dedicated support
Usage-Based	$0.01/ad (up to 10,000), $0.008/ad (10,001-50,000), $0.005/ad (50,000+)	Pay-as-you-go, scalable capacity, basic analytics, email support, real-time usage dashboard
Value-Based	Custom pricing based on ROI	Custom limits, comprehensive ROI analysis, dedicated support, custom integration and feature development
Tiered Pricing	Starter: $50/month, Growth: $200/month, Scale: $500/month, Enterprise: Custom	Starter: Up to 500 verifications, basic reporting; Growth: Up to 5,000 verifications, advanced analytics; Scale: Up to 20,000 verifications, comprehensive reporting; Enterprise: Unlimited verifications, full analytics, dedicated support
Enterprise Licensing	Starting at $10,000/year	Unlimited verifications, full analytics, dedicated account manager, 24/7 support, custom integration, on-site training

By offering these value-added services, we can generate additional revenue while strengthening our position as a trusted partner for advertisers.

Reference

▼ **Articles referred**

https://fastercapital.com/content/Social-media-advertising--Ad-Revenue--Understanding-Ad-Revenue-in-the-Context-of-Social-Media-Advertising.html

https://malwaretips.com/blogs/facebook-scams/

https://www.hootsuite.com/platform/engagement

https://buffer.com/resources/social-media-advertising-guide/

https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2023/10/social-media-golden-goose-scammers

https://www.aura.com/learn/facebook-scams

https://cheq.ai/blog/all-about-facebook-ad-fraud/