PhishBait

AI-Based In-Game Chat Phishing Detection

AI&NN - 20CYS304 Mini Project

Anagha B Prashanth - CB.SC.U4CYS23002
Devinandha - CB.SC.U4CYS23011
Ishitha Praveen - CB.SC.U4CYS23018

AIM

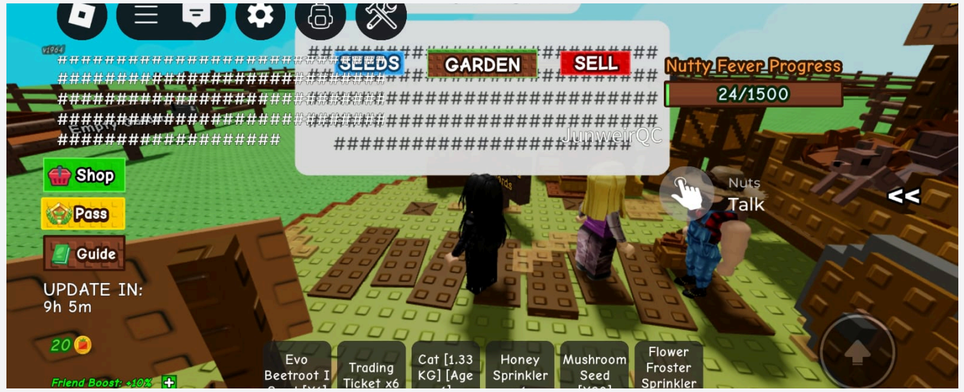
To design and develop an AI-based, real-time phishing detection system for in-game chat environments that automatically identifies phishing or scam messages and notifies users through a browser extension, which will send the chats messages to a locally hosted Flask service that runs the machine-learning model and returns classification results.

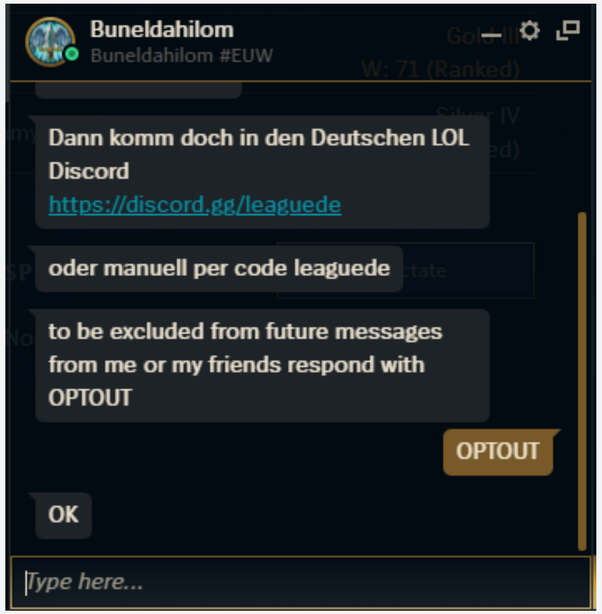
PROBLEM STATEMENT

Online multiplayer games often include open chat systems where players freely communicate. However, these chat environments are increasingly being exploited by attackers who send phishing messages, scam links, or fake promotions to trick users into revealing personal information or visiting malicious sites.

Hence, there is a need for a real-time detection system that can monitor in-game chats locally, detect phishing or scam messages instantly, and alert users without interrupting gameplay.

CURRENT CHALLENGES





- Growing online multiplayer game community
- Players easily tricked by fake offers or links
- Traditional filters often fail due to informal game chat language

PROPOSED SYSTEM

1. In-Game Chat Capture: The browser extension captures live messages from the game chat and monitors all incoming and outgoing text in real time.

2. Pre-processing Module: This module extracts the message text and prepares it for feature extraction, ensuring it is in a format suitable for analysis.

3. Feature Extraction (TF-IDF Vectorization): The extracted text is converted into numerical vectors using TF-IDF, which highlights suspicious terms such as “link,” “free,” or “click.”

PROPOSED SYSTEM

4. Feature vectors get analysed by a Multi-layer Perceptron model, which classifies each message as phishing or safe and provides a probability score indicating detection confidence. Additional parameters are also included to get better game-specific results

5. Alert System: The classification results are sent back to the browser extension, where visual alerts in the form of coloured badges are displayed beside all messages to inform players in real time.

DATASETS USED

To train and evaluate the AI model for phishing and scam message detection in the in-game chat messages, three datasets were used: two publicly available SMS phishing datasets and one custom-generated dataset tailored for game environments.

- SMS Phishing Dataset for Machine Learning and Pattern Recognition
- Kaggle SMS Spam Collection (Text Classification)
- Custom Dataset

ARCHITECTURE OVERVIEW

Browser Extension

- Injected into game pages via manifest.json + content.js
- Observes chat DOM and extracts new messages
- Sends messages to local Flask server: http://127.0.0.1:5000/detect

Flask Server (chat_server.py)

- Exposes /detect (POST) and /status (GET) endpoints
- Background thread trains ML model on startup

Phishing Detector (phishing_detector.py)

- TF-IDF + MLP pipeline
- Combines ML prediction + heuristics (keywords, URLs, punctuation, all-caps)
- Returns JSON: {label, score, heuristics, keywords}

Extension UI

- Displays transient badge next to chat message (~12s) showing label + score

FILE ROLES

Manifest.json	Declares MV3 extension, permissions, injects content.js
content.js	Monitors chat, sends messages to server, shows badges
chat_server.py	Flask app, handles /detect and /status, starts training thread
phishing_detector.py	Loads datasets, trains model, computes predictions

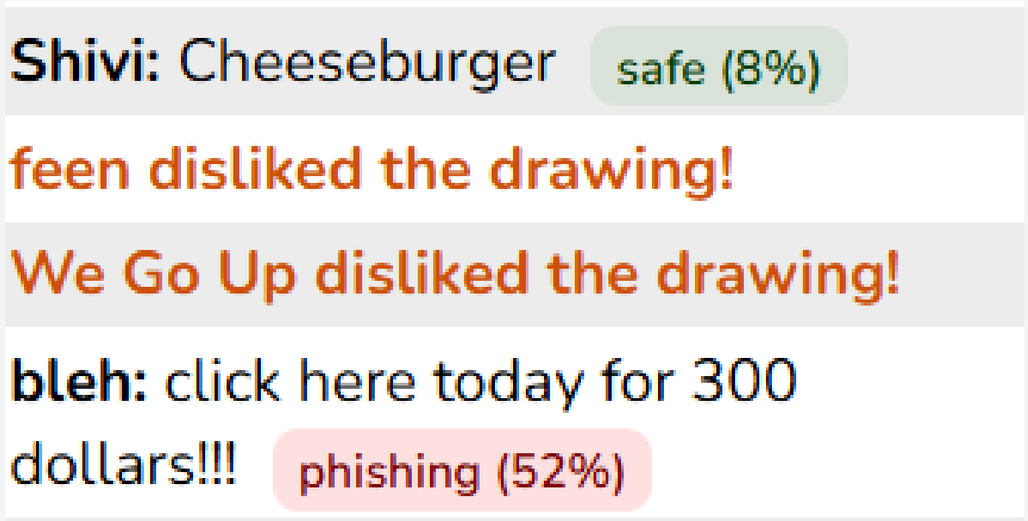
PhishBait - Request/Response Flow

1. User opens chat → `content.js` observes new messages
2. Sends message to Flask server `/detect` via `fetch`
3. Server calls `predict(text)` → ML + heuristics → returns JSON result
4. Extension parses response → shows badge beside message

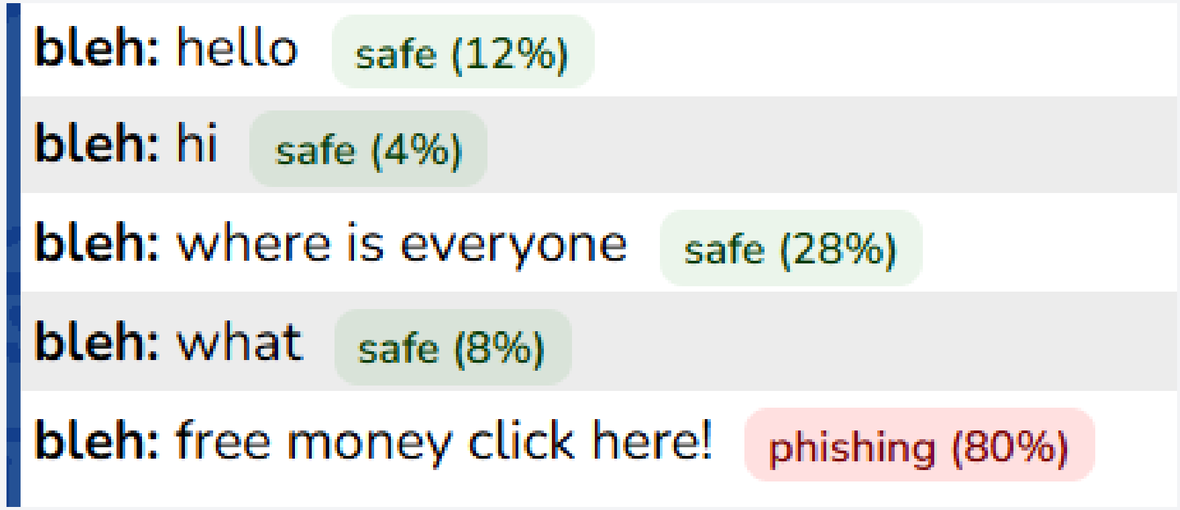
```
PS C:\Users\anagh\OneDrive\Desktop\abp\college\SEM 5\ai&nn\oratio\gamephis
gamephisher\flask_app.py"
127.0.0.1 - - [11/Oct/2025 11:05:43] "OPTIONS /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:05:43] "POST /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:06:04] "OPTIONS /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:06:04] "POST /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:06:06] "POST /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:06:11] "OPTIONS /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:06:11] "POST /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:07:11] "OPTIONS /detect HTTP/1.1" 200 -
127.0.0.1 - - [11/Oct/2025 11:07:11] "POST /detect HTTP/1.1" 200 -
```

RESULTS

System manages to successfully classify game chats as phishing or safe in real-time



Can handle multiple simultaneous chat streams without lag

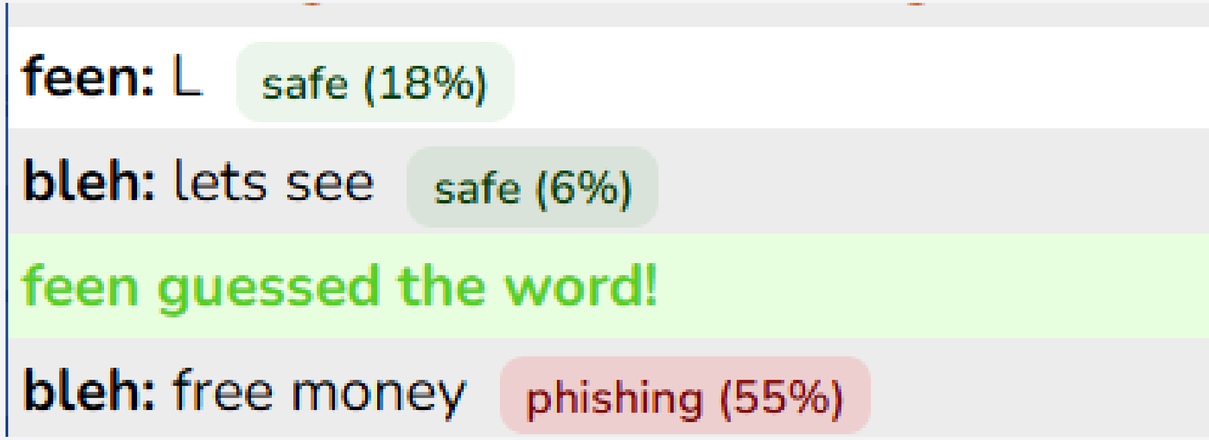


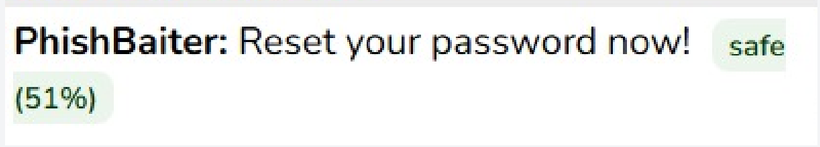
Provides the probability scores, with confidence levels for each message, along with coloured badges near messages to indicate phishing warnings or safe messages

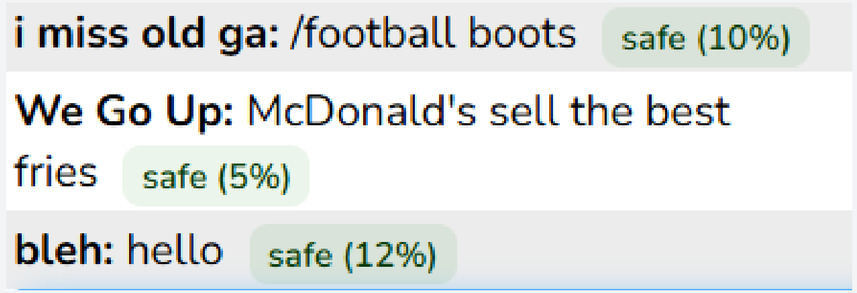
RESULTS

Most phishing messages detected involved:

1. Promises of in-game rewards
2. Suspicious URLs and link-related content
3. Requests for personal info/related texts

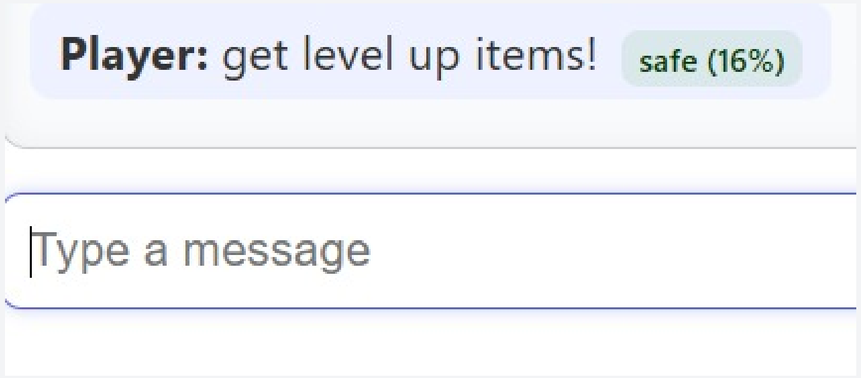






DRAWBACKS

- Language: The model’s accuracy needs to be improved for slang, abbreviations, or non-English messages.
- Limited Dataset Coverage: Training data may not represent all in-game chat styles or emerging phishing patterns.
- Keyword Sensitivity: Over-reliance on specific keywords can cause false positives or miss context-based phishing attempts.



CONCLUSION

PhishBait provides an effective, real-time safeguard against phishing and scams in live game chats by integrating a lightweight browser extension, a local Flask gateway, and MLP. Starting from a keyword and TF-IDF baseline with active learning improvements, it delivers fast, accurate detection while maintaining user-friendly performance. By preventing malicious interactions, securing accounts, and fostering safer communities, PhishBait empowers players to enjoy games confidently and responsibly.

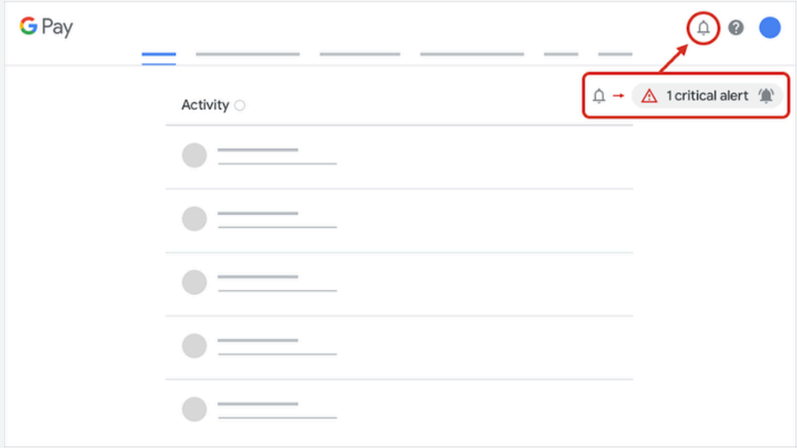
FUTURE SCOPE

Improve dataset + labeling

52	Just finished the main story, wow!	not_spam
53	Get free HeroesFall with this program: http://freeitems.online/claim	spam
54	Verify to claim your reward: http://gamekeys.xyz/claim	spam
55	Click here to get a rare mount instantly: http://loot-claim.biz/claim	spam
56	Free gold for everyone! Click here to claim: gamekeys.xyz	spam

Contextual & user-aware detection

Explainability & human-in-the-loop



THANK YOU