**Malicious Clickbait Detection: A Cybersecurity Approach**

**Objective: Clickbait prevention real-time extension- Cybersecurity project integrated with AI/ML**

## Abstract

The rapid growth of cyber threats, including phishing and malware distribution, has leveraged deceptive headlines as a primary vector to lure victims. This study focuses on developing an AI-based detection system to identify malicious clickbait headlines that facilitate malware installation or phishing attacks. By narrowing the scope to cybersecurity applications, this project aims to enhance user protection and reduce the risk of online fraud. We outline the methodology, including data collection, preprocessing, model development, and deployment, and present a robust framework for detecting malicious clickbait headlines with high accuracy and efficiency. We will generate an API Extension which focuses on highlighting the malicious clickbait.

## 1. Introduction

Clickbait headlines have evolved into a potent tool for cybercriminals. Unlike benign clickbait designed to attract attention for entertainment or marketing purposes, malicious clickbait is crafted to deceive users into performing harmful actions such as downloading malware or divulging sensitive information. This paper proposes an AI-driven system that identifies such headlines, focusing exclusively on malicious clickbait while excluding benign examples. This streamlined approach ensures a precise and cybersecurity-oriented solution.

### Objectives

- Develop a system to detect malicious clickbait headlines linked to malware installation or phishing attacks.

- Create a robust dataset for training and testing malicious and non-malicious headlines.

- Implement effective machine learning and natural language processing (NLP) techniques for classification.

-Develop an extension framework to integrate the model with search engines like Google Chrome.

## 2. Related Work

Existing research on clickbait detection primarily focuses on distinguishing benign clickbait from neutral headlines. Studies have explored traditional machine learning models and deep learning approaches for text classification. However, few works target malicious URL aimed at cybersecurity threats.

This research fills the gap by concentrating on detecting deceptive headlines with malicious intent.

## 3. Methodology

### 3.1 Data Collection

#### Malicious Clickbait Dataset

1. \*Phishing Headlines\*:

- Sources: PhishingCorpus, Enron Dataset (phishing emails), PhishTank, Scamwatch.

- Examples: "Your account has been compromised! Click here to verify your details."

2. \*Malware-Linked Headlines\*:

- Sources: Cybersecurity incident reports (e.g., AV-TEST, Symantec Threat Reports), malicious ad campaign archives.

- Examples: "Critical system error detected! Download the patch here."

#### Non-Malicious Dataset

- Headlines from legitimate sources such as news websites (e.g., BBC, Reuters) and trusted cybersecurity blogs.

### 3.2 Data Preprocessing

1. Text Cleaning:

- Convert text to lowercase.

- Remove special characters and URLs.

2. Tokenization:

- Split headlines into individual words.

3. Vectorization:

- Utilize TF-IDF or word embeddings (e.g., Word2Vec, GloVe) to represent text numerically.

### 3.3 Feature Engineering

- Extract meaningful features such as:

- Presence of urgency words (e.g., "urgent," "immediate").

- Action-oriented verbs (e.g., "click," "install," "verify").

- Suspicious phrases (e.g., "limited time offer," "you’ve won").

- Length and capitalization patterns.

### 3.4 Model Development

#### Models

1. \*Traditional Machine Learning\*:

- Logistic Regression, Random Forest, XGBoost.

- Naive Bayes for text classification.

2. \*Deep Learning\*:

- LSTM or GRU for sequential patterns.

- Transformer-based models (e.g., BERT) for contextual understanding.

#### Handling Imbalanced Datasets

- Techniques such as SMOTE (Synthetic Minority Oversampling) and weighted loss functions to address class imbalance.

### 3.5 Evaluation

Metrics:

- \*Precision\*: Accuracy of flagged malicious headlines.

- \*Recall\*: Ability to detect all malicious headlines.

- \*F1-Score\*: Balance between precision and recall.

## 4. Results and Discussion

The proposed system achieves high accuracy in detecting malicious clickbait headlines. Transformer-based models like BERT outperformed traditional ML models in capturing nuanced linguistic patterns. The use of SMOTE and weighted loss functions effectively addressed dataset imbalance. However, challenges remain in detecting subtle malicious intent, which requires further advancements in contextual NLP.

## 5. Deployment

### API Development

- Built using Flask to classify headlines as "Malicious" or "Not Malicious" in real-time.

### UI Development

- A HTML, CSS and JAVASCRIPT based web interface was implemented to allow users to test the system interactively.

### Integration

- Potential integration with browser extensions for real-time malicious headline detection.

## 6. Challenges and Future Work

### Challenges

1. \*Subtle Malicious Intent\*: Some malicious headlines mimic legitimate ones closely.

2. \*Dataset Limitations\*: Scarcity of publicly available labelled datasets.

### Future Work

1. \*Advanced NLP Models\*: Explore larger pre-trained models (e.g., GPT-based architectures).

2. \*Dynamic Contextual Understanding\*: Incorporate metadata such as sender information and source credibility.

3. \*Scalability\*: Deploy the system for large-scale cybersecurity tools.

## 7. Conclusion

This research presents a focused and effective framework for detecting malicious clickbait headlines, addressing a critical cybersecurity challenge. By leveraging advanced NLP techniques and deploying a functional detection system, this work contributes significantly to user protection against phishing and malware attacks.

## References

1. PhishTank: https://www.phishtank.com/

2. AV-TEST Reports: https://www.av-test.org/

3. Enron Email Dataset: https://www.cs.cmu.edu/~enron/

Websites for data acquisition:

Mp3 songs website

https://www.zscaler.com/blogs/security-research/look-top-websites-blocked

https://dynamicbusiness.com/locked/norton-reveals-100-most-dangerous-websites4168.html

https://otx.alienvault.com/pulse/64de49881646bdef1ba1cf2f