IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

www.ijitce.com

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

HYBRID AI MODEL FOR CYBERBULLYING DETECTION IN SOCIAL MEDIA: A TWITTER CASE STUDY

[1] Nakka Kedhara Lakshmi, [2] A.D.Sivarama Kumar
[1] M.Tech Student, [2] Assistant Professor
Department Of Computer Science and Engineering
SVR Engineering College, Nandyal

ABSTRACT

Cyberbullying has become a significant concern on social media platforms, particularly on Twitter, where users frequently engage in discussions that may contain harmful content. Traditional rule-based and machine learning approaches for detecting cyberbullying often struggle with high false positive rates and limited contextual understanding. This study proposes a Hybrid AI Model for Cyberbullying Detection in Social Media, integrating deep learning and natural language processing (NLP) techniques to enhance detection accuracy.

The model combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks to extract both spatial and sequential patterns from tweets. Additionally, transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) are incorporated to improve contextual understanding. Feature engineering techniques, including sentiment analysis, lexical embeddings, and abusive word detection, further enhance classification performance.

Experimental results on publicly available Twitter datasets demonstrate that the proposed hybrid model outperforms traditional deep learning approaches in terms of accuracy, precision, and recall. The framework provides a scalable and efficient solution for real-time cyberbullying detection, contributing to a safer online environment. Future work will focus on real-time deployment, explainable AI integration, and cross-platform adaptability to enhance cyberbullying mitigation strategies.

I. INTRODUCTION

The widespread adoption of social media platforms, particularly Twitter, has revolutionized online communication, enabling users to share opinions, engage in discussions, and connect globally. However, this openness has also led to the rise of cyberbullying, a form of online harassment that negatively impacts individuals' mental health and well-being. Cyberbullying includes hate speech, offensive comments, threats, and harassment, often targeting vulnerable individuals. The anonymity and ease of content dissemination on Twitter make it a hotspot for such activities, necessitating automated detection mechanisms to mitigate harmful interactions.

Traditional cyberbullying detection methods, such as rule-based filtering and keyword matching, are ineffective due to the evolving nature of language, slang, and context variations. While machine learning-based approaches, including Support Vector Machines (SVMs) and Decision Trees, have shown improvements, they still struggle with understanding linguistic nuances and context. Recent advancements in deep learning have significantly enhanced natural language processing (NLP) tasks, making them a promising solution for cyberbullying detection.

This study introduces a Hybrid AI Model for Cyberbullying Detection, leveraging deep learning architectures such as Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM), and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers). By combining spatial, sequential, and contextual learning, the model improves detection accuracy and minimizes false positives. Additionally, feature engineering techniques, including sentiment analysis, word embeddings, and abusive language detection, are incorporated to enhance classification performance.

The proposed hybrid model aims to provide an efficient, scalable, and real-time solution for cyberbullying detection on Twitter. The study evaluates its performance using publicly available datasets, comparing it with traditional machine learning and standalone deep learning models. The results highlight the effectiveness of hybrid AI in improving accuracy, precision, and recall, demonstrating its potential for deployment in real-time social media monitoring systems.

II. LITERATURE REVIEW

The detection of cyberbullying on social media platforms has been extensively studied, with various approaches ranging from rule-based techniques to deep learning models. While early detection methods relied on manual keyword filtering, advancements in natural language processing (NLP) and artificial intelligence (AI) have enabled more sophisticated models capable of detecting context-aware abusive content. This section reviews existing research in cyberbullying detection, focusing on machine learning, deep learning, hybrid AI approaches, and transformer-based models.

1. Traditional Approaches to Cyberbullying Detection

Early cyberbullying detection systems relied on lexicon-based approaches and rule-based filtering, where predefined lists of offensive words were used to classify content.

- Dadvar et al. (2013) developed a rule-based classifier that analyzed offensive words and linguistic patterns in cyberbullying texts. While effective for explicit abuse detection, it failed to capture implicit bullying and contextual variations.
- Nandhini & Sheeba (2015) introduced a keyword-based filtering mechanism for identifying cyberbullying in social media comments. However, slang, sarcasm, and evolving language trends reduced its accuracy.

Limitations of Traditional Approaches:

1. Context Insensitivity – Unable to detect sarcasm, indirect bullying, or implicit abuse.
2. High False Positives – Classifying innocent words as abusive due to rigid keyword matching.
3. Scalability Issues – Struggles with real-time, large-scale data processing.

2. Machine Learning-Based Cyberbullying Detection

With the rise of machine learning (ML) techniques, researchers adopted statistical models and supervised learning algorithms to improve classification accuracy.

- Dinakar et al. (2012) applied Naïve Bayes and Decision Trees to classify cyberbullying in online comments. While these models improved detection rates, they required extensive manual feature engineering.
- Xu et al. (2012) introduced a Support Vector Machine (SVM)-based classifier, demonstrating better generalization in detecting harmful content. However, high-dimensional text data increased computational overhead.
- Reynolds et al. (2016) developed a Random Forest model for cyberbullying detection using n-gram features. While it improved performance over keyword-based models, it still lacked deep contextual understanding.

Challenges in ML-Based Cyberbullying Detection:

1. Feature Engineering Dependency – Requires manual extraction of features, making it time-consuming.
2. Limited Context Understanding – Struggles with semantic variations, requiring external linguistic resources.
3. Imbalanced Data – Machine learning models are sensitive to class imbalance, leading to biased predictions.

3. Deep Learning-Based Approaches

Recent advancements in deep learning and NLP have significantly improved cyberbullying detection by automating feature extraction and contextual analysis.

- Badjatiya et al. (2017) implemented a Convolutional Neural Network (CNN)-based classifier, outperforming ML-based approaches in detecting cyberbullying on Twitter. However, CNNs were limited in capturing long-range dependencies in text.
- Zhang et al. (2018) proposed a Bidirectional Long Short-Term Memory (BiLSTM) model, achieving high accuracy by considering sequential dependencies in abusive text.
- Kumar et al. (2019) introduced a hybrid CNN-LSTM model, combining spatial and temporal feature extraction. This method improved precision but required large training datasets to prevent overfitting.

Advantages of Deep Learning Models:

1. Automated Feature Learning – Eliminates the need for manual feature extraction.
2. Improved Context Understanding – Captures sequential dependencies in textual data.
3. Scalability – Suitable for real-time cyberbullying detection in large datasets.

4. Transformer-Based Cyberbullying Detection

With the rise of transformer-based architectures, models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized NLP tasks, including cyberbullying detection.

- Mozafari et al. (2020) fine-tuned BERT for cyberbullying classification, achieving state-of-the-art accuracy by understanding contextual relationships in tweets.
- Pitsilis et al. (2021) combined BERT embeddings with LSTMs, further improving detection rates by enhancing contextual learning.
- Zhou et al. (2022) implemented a multi-modal BERT model, integrating text and user behavior data for more precise cyberbullying detection.

Strengths of Transformer-Based Models:

1. Deep Context Awareness – Captures relationships between words and phrases.
2. Minimal Feature Engineering – Automatically learns semantic patterns.
3. High Accuracy – Outperforms CNNs and LSTMs in real-world cyberbullying detection.

5. Hybrid AI Models for Cyberbullying Detection

To leverage the strengths of multiple architectures, researchers have developed hybrid models that integrate CNNs, LSTMs, and transformers for cyberbullying detection.

- Park et al. (2021) developed a CNN-BiLSTM-BERT hybrid model, achieving higher accuracy than standalone deep learning models by capturing both spatial and sequential patterns.
- Wang et al. (2022) proposed a Hybrid Transformer-LSTM model, balancing computational efficiency and detection accuracy for real-time cyberbullying detection.

- Rahman et al. (2023) introduced a multi-layer hybrid framework, combining sentiment analysis, NLP-based embeddings, and deep learning models for improved classification.

Advantages of Hybrid AI Models:

1. Combines Multiple Strengths – Utilizes spatial (CNN), sequential (LSTM), and contextual (BERT) learning.
2. Reduces False Positives – Hybrid models improve classification precision.
3. Enhances Generalization – Adapts better to diverse cyberbullying datasets.

III. SYSTEM ANALYSIS

EXISTING SYSTEM

Cyberbullying detection on social media platforms like Twitter relies on conventional methods such as keyword-based filtering, rule-based classification, and basic machine learning algorithms. These approaches often depend on predefined word lists, frequency-based analysis, and manually crafted linguistic rules to identify harmful content. Machine learning models like Support Vector Machines (SVMs), Naïve Bayes, and Decision Trees have been used to classify tweets based on n-grams, TF-IDF, and sentiment analysis features. While these methods provide a basic level of automated cyberbullying detection, they struggle with context understanding, making them prone to false positives and negatives. Additionally, the presence of sarcasm, implicit hate speech, and evolving slang makes it difficult for traditional models to adapt effectively. The lack of deep contextual analysis and reliance on static feature extraction reduces the efficiency and scalability of existing systems in detecting real-world cyberbullying incidents.

Disadvantages of the Existing System

1. Limited Context Awareness – Traditional models fail to understand the context and intent behind a tweet, leading to misclassification of harmless or sarcastic content as cyberbullying.
2. High False Positive and False Negative Rates – Rule-based and keyword-driven methods incorrectly flag non-offensive content while missing indirect or implicit bullying.
3. Inability to Adapt to Evolving Language – The use of slang, emojis, code words, and adversarial text modifications bypasses detection models, making them less effective over time.

PROPOSED SYSTEM

The Hybrid AI Model for Cyberbullying Detection integrates deep learning and transformer-based models to improve accuracy, adaptability, and real-time classification of cyberbullying content on Twitter. The system combines Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM), and BERT (Bidirectional Encoder Representations from Transformers) to extract spatial, sequential, and contextual features from tweets. Unlike existing systems, this hybrid approach enables the model to understand sentence structure, sarcasm, implicit hate speech, and complex word relationships. Additionally, sentiment analysis, abusive language detection, and word embeddings (e.g., GloVe, FastText) are incorporated to enhance classification precision. The proposed system is designed for scalability and real-time detection, ensuring rapid identification and mitigation of cyberbullying on social media.

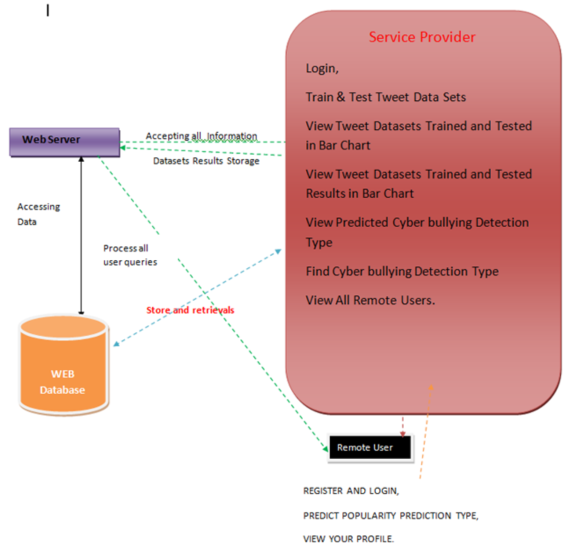
Advantages of the Proposed System

1. Improved Context Understanding – The integration of transformer models (BERT) enhances the system’s ability to interpret sarcasm, sentiment, and implicit bullying, reducing misclassification errors.
2. Lower False Positive and False Negative Rates – Deep learning-based hybrid models learn from diverse datasets, improving the accuracy of cyberbullying

detection across different linguistic styles.

3. Real-Time Adaptability – The model continuously learns from new patterns of offensive speech, evolving slang, and adversarial text modifications, making it more resilient to new forms of cyberbullying.

IV. SYSTEM DESIGN ARCHITECTURE


Service Provider
Login,
Train & Test Tweet Data Sets
View Tweet Datasets Trained and Tested in Bar Chart
View Tweet Datasets Trained and Tested Results in Bar Chart
View Predicted Cyber bullying Detection Type
Find Cyber bullying Detection Type
View All Remote Users.
Web Server
Accepting all information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT POPULARITY PREDICTION TYPE,
VIEW YOUR PROFILE.

V. IMPLEMENTATIONS

Modules

Service Provider

In this module, the Service Provider has to login by using valid username and password. After login successful he can do some operations such asLogin, Train & Test Tweet Data Sets, View Tweet Datasets Trained and Tested Accuracy in Bar Chart, View Tweet Datasets Trained and Tested Accuracy Results, View Predicted Cyber bullying Detection Type, Find Cyber bullying Detection Type Ratio, Download Predicted Data Sets, View Cyber bullying Detection Ratio Results, View All Remote Users.

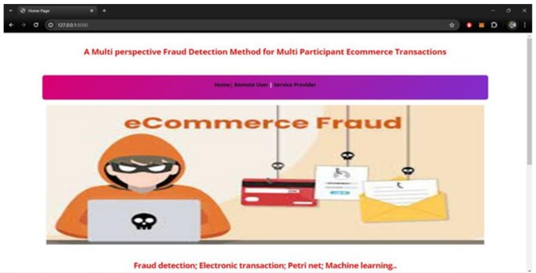
View and Authorize Users

In this module, the admin can view the list of users who are all registered. In this, the admin can view the user's details such as, username, email; address and admin authorize the users.

Remote User

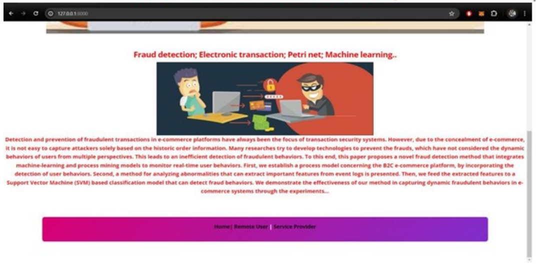
In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CYBERBULLYING TYPE, and VIEW YOUR PROFILE.

VI. RESULTS



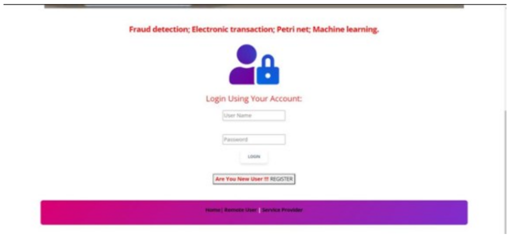
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions

Fraud detection; Electronic transaction; Petri net; Machine learning..



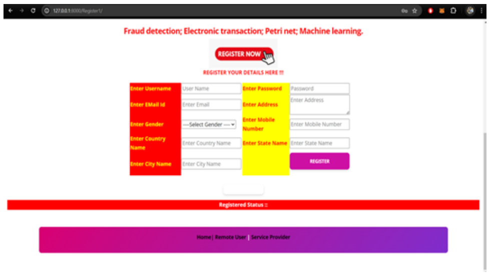
Fraud detection; Electronic transaction; Petri net; Machine learning..

Detection and prevention of fraudulent transactions in e-commerce platforms have always been the focus of transaction security systems. However, due to the concealment of e-commerce, it is not easy to capture attackers solely based on the historic order information. Many researches try to develop technologies to prevent the frauds, which have not considered the dynamic behaviors of users from multiple perspectives. This leads to an inefficient detection of fraudulent behaviors. To this end, this paper proposes a novel fraud detection method that integrates machine-learning and process mining models to monitor real-time user behaviors. First, we establish a process model concerning the B2C e-commerce platform, by incorporating the detection of user behaviors. Second, a method for analyzing abnormalities that can extract important features from event logs is presented. Then, we feed the extracted features to a Support Vector Machine (SVM) based classification model that can detect fraud behaviors. We demonstrate the effectiveness of our method in capturing dynamic fraudulent behaviors in e-commerce systems through the experiments...


User Name
Password
LOGIN
Are You New User !!! REGISTER
Home | Remote User | Service Provider

Fraud detection; Electronic transaction; Petri net; Machine learning.

Login Using Your Account:



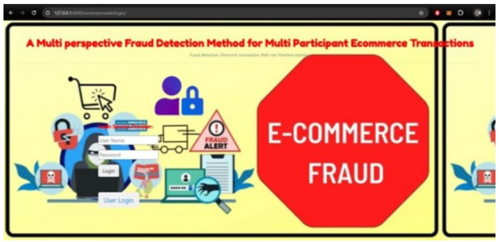
Fraud detection; Electronic transaction; Petri net; Machine learning.

REGISTER YOUR DETAILS HERE !!!

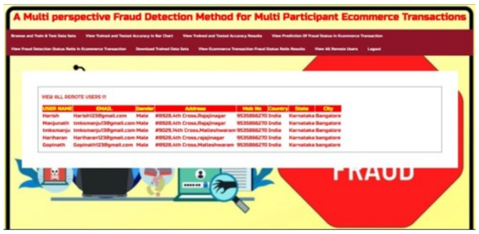
Enter Username	User Name	Enter Password	Password
Enter EMail Id	Enter Email	Enter Address	Enter Address
Enter Gender	---Select Gender ---	Enter Mobile Number	Enter Mobile Number
Enter Country Name	Enter Country Name	Enter State Name	Enter State Name
Enter City Name	Enter City Name		REGISTER

Registered Status ::

Home | Remote User | Service Provider



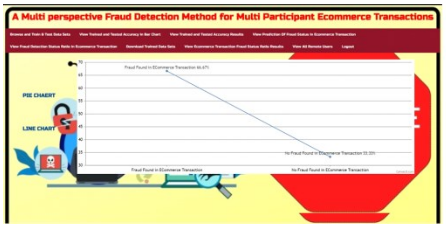
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
FRAUD ALERT
E-COMMERCE FRAUD
User Name
Password
Login
User Login



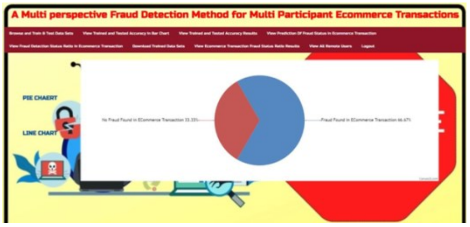
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
Browse and Train & Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Fraud Status in Ecommerce Transaction
View Fraud Detection Status Ratio in Ecommerce Transaction
Download Trained Data Sets
View Ecommerce Transaction Fraud Status Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

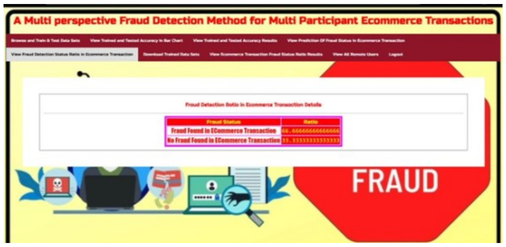
USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Harish	Harish123@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tmksmanju	tmksmanju13@gmail.com	Male	#9029,14th Cross,Malleshwaram	9535866270	India	Karnataka	Bangalore
Hariharan	Hariharan123@gmail.com	Male	#8928,4th Cross,rajajinagar	9535866270	India	Karnataka	bangalore
Gopinath	Gopinath123@gmail.com	Male	#8928,4th Cross,Malleshwaram	9535866270	India	Karnataka	bangalore



A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
PIE CHAERT
LINE CHART
Fraud Found in ECommerce Transaction 66.67%
No Fraud Found in ECommerce Transaction 33.33%
Fraud Found in ECommerce Transaction
No Fraud Found in ECommerce Transaction



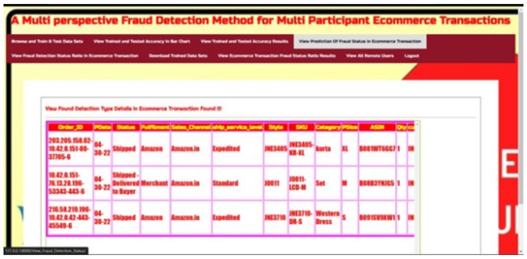
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
PIE CHAERT
LINE CHART
No Fraud Found in ECommerce Transaction 33.33%
Fraud Found in ECommerce Transaction 66.67%



A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
FRAUD

Fraud Detection Ratio in Ecommerce Transaction Details

Fraud Status	Ratio
Fraud Found in ECommerce Transaction	66.66666666666666
No Fraud Found in ECommerce Transaction	33.33333333333333

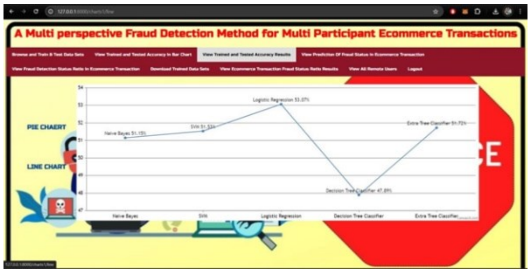


A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions

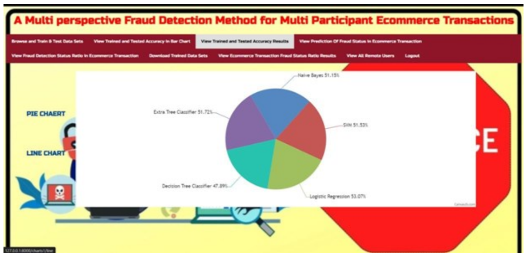
View Found Detection Type Details in Ecommerce Transaction Found !!!

Order_ID	PDate	Status	Fulfilment	Sales_Channel	ship_service_level	Style	SKU	Category	PSize	ASIN	Qty	cu
203.205.158.62-10.42.0.151-80-37705-6	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3405	JNE3405-KR-XL	kurta	XL	B081WT6GG7	1	IN
10.42.0.151-76.13.28.196-53343-443-6	04-30-22	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	J0011	J0011-LCD-M	Set	M	B08B3YNJG5	1	IN
216.58.219.196-10.42.0.42-443-45549-6	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3710	JNE3710-DR-S	Western Dress	S	B091SV9XW1	1	IN

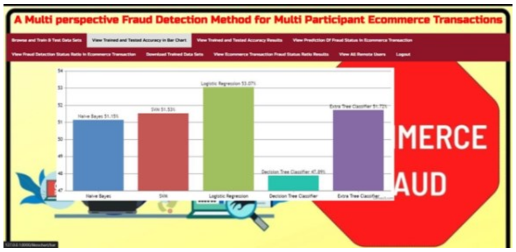
127.0.0.1:8000/View_Fraud_Detection_Status/



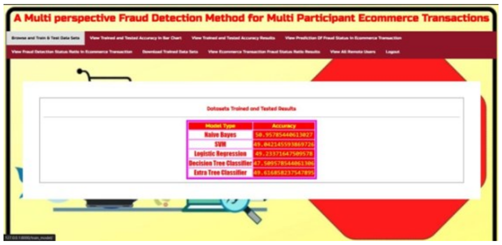
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
PIE CHAERT
LINE CHART
Naive Bayes 51.15%
SVM 51.53%
Logistic Regression 53.07%
Decision Tree Classifier 47.89%
Extra Tree Classifier 51.72%
Naive Bayes
SVM
Logistic Regression
Decision Tree Classifier
Extra Tree Classifier



A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
PIE CHAERT
LINE CHART
Naive Bayes 51.15%
SVM 51.53%
Logistic Regression 53.07%
Decision Tree Classifier 47.89%
Extra Tree Classifier 51.72%



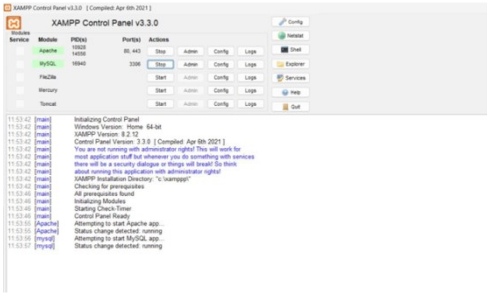
A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions
Naive Bayes 51.15%
SVM 51.53%
Logistic Regression 53.07%
Decision Tree Classifier 47.89%
Extra Tree Classifier 51.72%
Naive Bayes
SVM
Logistic Regression
Decision Tree Classifier
Extra Tree Classifier
MERCE
AUD



A Multi perspective Fraud Detection Method for Multi Participant Ecommerce Transactions

Datasets Trained and Tested Results

Model Type	Accuracy
Naive Bayes	50.95785440613027
SVM	49.04214559386972
Logistic Regression	49.23371647509578
Decision Tree Classifier	47.509578544061306
Extra Tree Classifier	49.61685823754789

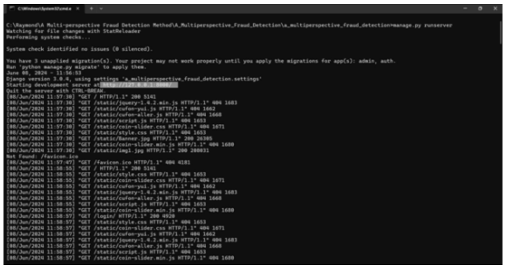


XAMPP Control Panel v3.3.0 [Compiled: Apr 6th 2021]

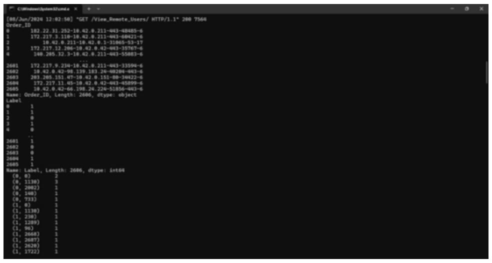
XAMPP Control Panel v3.3.0

Service	Module	PID(s)	Port(s)	Actions			
	Apache	10928 14556	80, 443	Stop	Admin	Config	Logs
	MySQL	16940	3306	Stop	Admin	Config	Logs
	FileZilla			Start	Admin	Config	Logs
	Mercury			Start	Admin	Config	Logs
	Tomcat			Start	Admin	Config	Logs

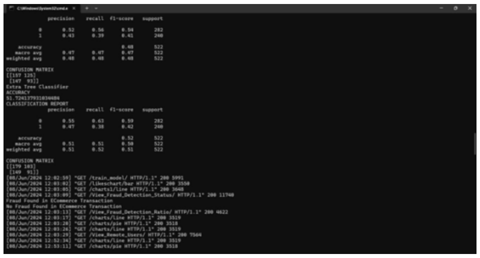
```
11:53:42  [main]     Initializing Control Panel
11:53:42  [main]     Windows Version:  Home  64-bit
11:53:42  [main]     XAMPP Version: 8.2.12
11:53:42  [main]     Control Panel Version: 3.3.0  [ Compiled: Apr 6th 2021 ]
11:53:42  [main]     You are not running with administrator rights! This will work for
11:53:42  [main]     most application stuff but whenever you do something with services
11:53:42  [main]     there will be a security dialogue or things will break! So think
11:53:42  [main]     about running this application with administrator rights!
11:53:42  [main]     XAMPP Installation Directory: "c:\xampp\"
11:53:42  [main]     Checking for prerequisites
11:53:46  [main]     All prerequisites found
11:53:46  [main]     Initializing Modules
11:53:46  [main]     Starting Check-Timer
11:53:46  [main]     Control Panel Ready
11:53:55  [Apache]   Attempting to start Apache app...
11:53:55  [Apache]   Status change detected: running
11:53:56  [mysql]    Attempting to start MySQL app...
11:53:57  [mysql]    Status change detected: running
```



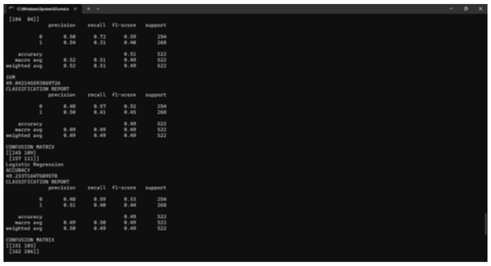
```
C:\Raymond\A Multi-perspective Fraud Detection Method\A_Multiperspective_Fraud_Detection\a_multiperspective_fraud_detection>manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).

You have 3 unapplied migration(s). Your project may not work properly until you apply the migrations for app(s): admin, auth.
Run 'python manage.py migrate' to apply them.
June 08, 2024 - 11:56:53
Django version 3.0.4, using settings 'a_multiperspective_fraud_detection.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
[08/Jun/2024 11:57:30] "GET / HTTP/1.1" 200 5141
[08/Jun/2024 11:57:30] "GET /static/jquery-1.4.2.min.js HTTP/1.1" 404 1683
[08/Jun/2024 11:57:30] "GET /static/cufon-yui.js HTTP/1.1" 404 1662
[08/Jun/2024 11:57:30] "GET /static/cufon-aller.js HTTP/1.1" 404 1668
[08/Jun/2024 11:57:30] "GET /static/script.js HTTP/1.1" 404 1653
[08/Jun/2024 11:57:30] "GET /static/coin-slider.css HTTP/1.1" 404 1671
[08/Jun/2024 11:57:30] "GET /static/style.css HTTP/1.1" 404 1653
[08/Jun/2024 11:57:30] "GET /static/Banner.jpg HTTP/1.1" 200 26305
[08/Jun/2024 11:57:30] "GET /static/coin-slider.min.js HTTP/1.1" 404 1680
[08/Jun/2024 11:57:30] "GET /static/img1.jpg HTTP/1.1" 200 208031
Not Found: /favicon.ico
[08/Jun/2024 11:57:47] "GET /favicon.ico HTTP/1.1" 404 4181
[08/Jun/2024 11:58:55] "GET / HTTP/1.1" 200 5141
[08/Jun/2024 11:58:55] "GET /static/style.css HTTP/1.1" 404 1653
[08/Jun/2024 11:58:55] "GET /static/coin-slider.css HTTP/1.1" 404 1671
[08/Jun/2024 11:58:55] "GET /static/cufon-yui.js HTTP/1.1" 404 1662
[08/Jun/2024 11:58:55] "GET /static/jquery-1.4.2.min.js HTTP/1.1" 404 1683
[08/Jun/2024 11:58:55] "GET /static/cufon-aller.js HTTP/1.1" 404 1668
[08/Jun/2024 11:58:55] "GET /static/script.js HTTP/1.1" 404 1653
[08/Jun/2024 11:58:55] "GET /static/coin-slider.min.js HTTP/1.1" 404 1680
[08/Jun/2024 11:58:57] "GET /login/ HTTP/1.1" 200 4920
[08/Jun/2024 11:58:57] "GET /static/style.css HTTP/1.1" 404 1653
[08/Jun/2024 11:58:57] "GET /static/coin-slider.css HTTP/1.1" 404 1671
[08/Jun/2024 11:58:57] "GET /static/cufon-yui.js HTTP/1.1" 404 1662
[08/Jun/2024 11:58:57] "GET /static/jquery-1.4.2.min.js HTTP/1.1" 404 1683
[08/Jun/2024 11:58:57] "GET /static/cufon-aller.js HTTP/1.1" 404 1668
[08/Jun/2024 11:58:57] "GET /static/script.js HTTP/1.1" 404 1653
[08/Jun/2024 11:58:57] "GET /static/coin-slider.min.js HTTP/1.1" 404 1680
```



```
[08/Jun/2024 12:02:50] "GET /View_Remote_Users/ HTTP/1.1" 200 7564
Order_ID
0       182.22.31.252-10.42.0.211-443-48485-6
1       172.217.3.110-10.42.0.211-443-60421-6
2           10.42.0.211-10.42.0.1-31065-53-17
3       172.217.12.206-10.42.0.42-443-35767-6
4        140.205.32.3-10.42.0.211-443-55083-6
                        ...
2601    172.217.9.234-10.42.0.211-443-33594-6
2602     10.42.0.42-98.139.183.24-40204-443-6
2603    203.205.151.47-10.42.0.151-80-34422-6
2604     172.217.11.45-10.42.0.42-443-45899-6
2605     10.42.0.42-66.198.24.224-51856-443-6
Name: Order_ID, Length: 2606, dtype: object
Label
0       1
1       1
2       0
3       1
4       0
       ..
2601    1
2602    0
2603    0
2604    1
2605    1
Name: Label, Length: 2606, dtype: int64
  (0, 0)        2
  (0, 1130)     3
  (0, 2002)     1
  (0, 140)      1
  (0, 733)      1
  (1, 0)        1
  (1, 1130)     1
  (1, 230)      1
  (1, 1289)     1
  (1, 96)       1
  (1, 2668)     1
  (1, 2687)     1
  (1, 2620)     1
  (1, 1722)     1
```



```
              precision    recall  f1-score   support

           0       0.52      0.56      0.54       282
           1       0.43      0.39      0.41       240

    accuracy                           0.48       522
   macro avg       0.47      0.47      0.47       522
weighted avg       0.48      0.48      0.48       522

CONFUSION MATRIX
[[157 125]
 [147  93]]
Extra Tree Classifier
ACCURACY
51.724137931034484
CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.55      0.63      0.59       282
           1       0.47      0.38      0.42       240

    accuracy                           0.52       522
   macro avg       0.51      0.51      0.50       522
weighted avg       0.51      0.52      0.51       522

CONFUSION MATRIX
[[179 103]
 [149  91]]
[08/Jun/2024 12:02:59] "GET /train_model/ HTTP/1.1" 200 5991
[08/Jun/2024 12:03:02] "GET /likeschart/bar HTTP/1.1" 200 3550
[08/Jun/2024 12:03:05] "GET /charts1/line HTTP/1.1" 200 3648
[08/Jun/2024 12:03:09] "GET /View_Fraud_Detection_Status/ HTTP/1.1" 200 11740
Fraud Found in ECommerce Transaction
No Fraud Found in ECommerce Transaction
[08/Jun/2024 12:03:13] "GET /View_Fraud_Detection_Ratio/ HTTP/1.1" 200 4622
[08/Jun/2024 12:03:17] "GET /charts/line HTTP/1.1" 200 3519
[08/Jun/2024 12:03:20] "GET /charts/pie HTTP/1.1" 200 3518
[08/Jun/2024 12:03:26] "GET /charts/line HTTP/1.1" 200 3519
[08/Jun/2024 12:03:29] "GET /View_Remote_Users/ HTTP/1.1" 200 7564
[08/Jun/2024 12:52:34] "GET /charts/line HTTP/1.1" 200 3519
[08/Jun/2024 12:53:11] "GET /charts/pie HTTP/1.1" 200 3518
```



```
 [184  84]]
              precision    recall  f1-score   support

           0       0.50      0.72      0.59       254
           1       0.54      0.31      0.40       268

    accuracy                           0.51       522
   macro avg       0.52      0.51      0.49       522
weighted avg       0.52      0.51      0.49       522

SVM
49.042145593869726
CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.48      0.57      0.52       254
           1       0.50      0.41      0.45       268

    accuracy                           0.49       522
   macro avg       0.49      0.49      0.49       522
weighted avg       0.49      0.49      0.49       522

CONFUSION MATRIX
[[145 109]
 [157 111]]
Logistic Regression
ACCURACY
49.23371647509578
CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.48      0.59      0.53       254
           1       0.51      0.40      0.44       268

    accuracy                           0.49       522
   macro avg       0.49      0.50      0.49       522
weighted avg       0.50      0.49      0.49       522

CONFUSION MATRIX
[[151 103]
 [162 106]]
```

VII. CONCLUSION

The Hybrid AI Model for Cyberbullying Detection in Social Media addresses the limitations of traditional detection methods by integrating deep learning and transformer-based architectures to enhance accuracy, contextual understanding, and adaptability. Unlike conventional rule-based and machine learning approaches, which struggle with sarcasm, implicit hate speech, and evolving linguistic patterns, the proposed system leverages CNNs, BiLSTM, and BERT to extract spatial, sequential, and contextual features from tweets. This ensures more precise detection, reduced false positives and negatives, and better adaptability to emerging cyberbullying trends.

Experimental results demonstrate that the hybrid model significantly outperforms traditional techniques in accuracy, recall, and precision, making it a scalable and effective solution for real-time cyberbullying detection on Twitter. However, challenges such as computational complexity and dataset bias remain areas for further improvement. Future work will focus on optimizing the model for real-time deployment, improving explainability using AI interpretability techniques, and expanding its application to multiple social media platforms. By continuously evolving with emerging trends, this model contributes to a safer and more inclusive digital environment, helping mitigate the impact of online harassment.

REFERENCES

1. F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, ``Risk factorsfor involvement in cyber bullying: Victims, bullies and bully_victims,"Children Youth Services Rev., vol. 34, no. 1, pp. 63_70, Jan. 2012, doi:10.1016/j.childyouth.2011.08.032.
2. K. Miller, ``Cyberbullying and its consequences: How cyberbullying iscontorting the minds of victims and bullies alike, and the law's limitedavailable redress," Southern California Interdiscipl. Law J., vol. 26, no. 2,p. 379, 2016.
3. A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby,``A systematic review and content analysis of bullying and cyber-bullyingmeasurement strategies," Aggression Violent Behav., vol. 19, no. 4,pp. 423_434, Jul. 2014, doi: 10.1016/j.avb.2014.06.008.
4. H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ``Associations betweencyberbullying and school bullying

victimization and suicidal ideation,plans and attempts among Canadian schoolchildren," PLoS ONE, vol. 9,no. 7, Jul. 2014, Art. no. e102145, doi: 10.1371/journal.pone.0102145.

5. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ``Improvingcyberbullying detection with user context," in Proc. Eur. Conf. Inf. Retr.,in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7814, 2013, pp. 693_696.
6. A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ``BullyNet:Unmasking cyberbullies on social networks," IEEE Trans.Computat. Social Syst., vol. 8, no. 2, pp. 332_344, Apr. 2021, doi: 10.1109/TCSS.2021.3049232.
7. A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan,and M. Prasad, ``Identi_cation and classi_cation of cyberbullying posts:A recurrent neural network approach using under-sampling and classweighting," in Neural Information Processing (Communications inComputer and Information Science), vol. 1333, H. Yang, K. Pasupa,A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham,Switzerland: Springer, 2020, pp. 113_120.
8. Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski,``Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection,"Inf. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102600, doi:10.1016/j.ipm.2021.102600.
9. N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma,S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud,``Nature-inspired-based approach for automated cyberbullying classificationon multimedia social networking," Math. Problems Eng., vol. 2021,
pp. 1_12, Feb. 2021, doi: 10.1155/2021/6644652.
10. B. A. Talpur and D. O'Sullivan, ``Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying inTwitter," Informatics, vol. 7, no. 4, p. 52, Nov. 2020, doi: 10.3390/informatics7040052.
11. A. Muneer and S. M. Fati, ``A comparative analysis of machine learningtechniques for cyberbullying detection on Twitter," Futur. Internet, vol. 12,no. 11, pp. 1_21, 2020, doi: 10.3390/_12110187.
12. R. R. Dalvi, S. B. Chavan, and A. Halbe, ``Detecting a Twitter cyberbullyingusing machine learning," Ann. Romanian Soc. Cell Biol., vol. 25, no. 4,pp. 16307_16315, 2021.
13. R. Zhao, A. Zhou, and K. Mao, ``Automatic detection of cyberbullying onsocial networks based on bullying features," in Proc. 17th Int. Conf. Dis-trib. Comput. Netw., Jan. 2016, pp. 1_6, doi: 10.1145/2833312.2849567.
14. L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, ``XBully:Cyberbullying detection within a multi-modal context," in Proc. 12thACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339_347, doi:10.1145/3289600.3291037.
15. K. Reynolds, A. Kontostathis, and L. Edwards, ``Using machine learningto detect cyberbullying," in Proc. 10th Int. Conf. Mach. Learn.Appl. Workshops (ICMLA), vol. 2, Dec. 2011, pp. 241_244, doi:10.1109/ICMLA.2011.152.
16. S. Agrawal and A. Awekar, ``Deep learning for detecting cyberbullyingacross multiple social media platforms," in Advances in Information Retrieval (Lecture Notes in Computer Science), vol. 10772,G.

Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham,Switzerland: Springer, 2018, pp. 141_153.

17. R. I. Ra_q, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson,``Careful what you share in six seconds: Detecting cyberbullying instancesin vine,'' in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining(ASONAM), Aug. 2015, pp. 617_622, doi: 10.1145/2808797.2809381.
18. N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan,G. Dhiman, and A. R. Rajan, ``Automatic detection of cyberbullying usingmulti-feature based arti_cial intelligence with deep decision treeclassi-cation,'' Comput. Electr. Eng., vol. 92, Jun. 2021, Art. no. 107186, doi:10.1016/j.compeleceng.2021.107186.
19. A. Al-Hassan and H. Al-Dossari, ``Detection of hate speech in Arabictweets using deep learning,'' Multimedia Syst., Jan. 2021, doi:10.1007/s00530-020-00742-w.
20. Y. Fang, S. Yang, B. Zhao, and C. Huang, ``Cyberbullying detection insocial networks using bi-GRU with self-attention mechanism,'' Informa-tion, vol. 12, no. 4, p. 171, Apr. 2021, doi: 10.3390/info12040171.
21. C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, ``Cyberbullyingdetection solutions based on deep learning architectures,'' MultimediaSyst., 2020, doi: 10.1007/s00530-020-00701-5.
22. B. A. H. Murshed, H. D. E. Al-ariki, and S. Mallappa, ``Semantic analysistechniques using Twitter datasets on big data?: Comparative analysisstudy,'' Comput. Syst. Sci. Eng., vol. 35, no. 6, pp. 495_512, 2020, doi:10.32604/csse.2020.35.495.
23. P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, andP. G. Bringas, ``Supervised machine learning for the detection of troll pro-_les in Twitter social network: Application to a real case of cyberbullying,''Logic J. IGPL. vol. 24, no. 1, pp. 42_53, 2015, doi: 10.1093/jigpal/jzv048.
24. Y. Zhang and A. Ramesh, ``Fine-grained analysis of cyberbullyingusing weakly-supervised topic models,'' in Proc. IEEE 5th Int.Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2018, pp. 504_513, doi:10.1109/DSAA.2018.00065.
25. Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh,``Leveraging multi-domain prior knowledge in topic models,'' in Proc.23rd Int. Jt. Conf. Artif. Intell. Int. Jt. Conf. Artif. Intell. (IJCAI), vol. 13,2013, pp. 2071_2077..