International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Spam Detection and Fake User Identification Using Machine Learning Algorithm

Prof.Pawar S.D[1],Holkar Omkar[2],Waghmare Akash[3].

Assistant Professor[1], Computer Department, SPCOET SomeshwarnagarCollege, Baramati, IndiaStudent[2], Computer Department, SPCOET Someshwarnagar College, Baramati, India Student[3], Computer Department, SPCOET Someshwarnagar College, Baramati, India

---

Abstract - The rise of online platforms and social networks has brought about an exponential increase in the amount of spam content and the proliferation of fake user accounts. These malicious activities pose significant threats to the integrity, security, and user experience of online communities. Consequently, there is a pressing need for effective techniques to detect and mitigate spam and identify fake user accounts.

This project proposes a comprehensive approach to tackle spam detection and fake user identification through the application of advanced machine learning and data analysis techniques. The primary objective is to develop a robust system capable of accurately detecting and flagging spam content, as well as identifying and removing fake user accounts from online platforms.

***Key Words*:**Machine Learning,Linear Regression Algorithm ,Jupiter,Django,vs code.

1. INTRODUCTION*(Size 11, Times New roman)*

The advent of the internet and the proliferation of online platforms have revolutionized the way people connect, communicate, and share information. However, this tremendous growth has also attracted malicious actors who engage in spamming activities and create fake user accounts to exploit and disrupt online communities. Spam content, characterized by unwanted or unsolicited messages, advertisements, or scams, not only inundates platforms but also diminishes user experience and hampers the authenticity of online interactions. Fake user accounts, on the other hand, can be utilized for various nefarious purposes, including spreading misinformation, conducting fraudulent activities, or manipulating online discussions.

Addressing the issue of spam and fake user accounts is of paramount importance to maintain the integrity, security, and trustworthiness of online platforms and social networks. Manual moderation and user reporting systems have limitations in terms of scale, efficiency, and accuracy, making it essential to develop automated techniques that can effectively detect and identify such malicious activities. This project aims to design and implement a comprehensive system that leverages advanced machine learning and data analysis techniques for spam detection and fake user identification.

2. LITERATURE SURVEY

Title: Twitter fake account detection.

Authors: B. Erçahin, Ö. Aktaş, D. Kilinç, and C. Akyol Social networking sites such as Twitter and Facebook attracts millions of users across the world and their interaction with social networking has affected their life. This popularity in social networking has led to different problems including the possibility of exposing incorrect information to their users through fake accounts which results to the spread of malicious content. This situation can result to a huge damage in the real world to the society. In our study, we present a classification method for detecting the fake accounts on Twitter. We have preprocessed our dataset using a supervised discretization technique named Entropy Minimization Discretization (EMD) on numerical features and analyzed the results of the Naïve Bayes algorithm.

Title: Detecting spammer son Twitter.

Author: F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. With millions of users tweeting around the world, real time search systems and different types of mining tools are emerging to allow people tracking the repercussion of events and news on Twitter. However, although appealing as mech-anisms to ease the spread of news and allow users to discuss events and post their status, these services open opportu-nities for new forms of spam.

3. METHODOLOGY

The Used Cars data set was taken and data processing has done to filter the data and to remove some unnecessary data. The model was trained with the processed data using the random forest algorithm to predict the sales of used cars with higher accuracy. Fig 1 shows the structured outline for proposed Methodology.

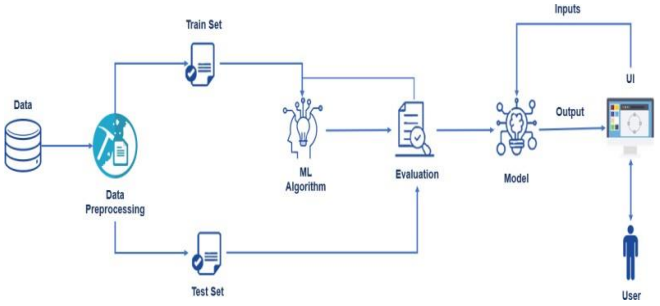

Fig 1: Structured outline of Proposed Methodology

A) Dataset Collection

Gather a diverse dataset comprising both legitimate and spam content, as well as user account data.Acquire data from various online platforms and social networks, ensuring representation from different domains and demographics.Implement appropriate mechanisms to ensure data privacy and compliance with legal requirements.

B) Data Preprocessing

Clean and preprocess the collected data, removing irrelevant information, duplicate entries, and any personally identifiable information (PII).

Normalize and standardize textual content by removing special characters, converting to lowercase, and applying stemming or lemmatization techniques.

Perform data sampling techniques to balance the dataset, especially if the distribution between legitimate and spam content or fake and genuine user accounts is imbalanced.

C) Feature Engineering:

Extract relevant features from the preprocessed data, including textual features (e.g., word frequency, sentiment analysis), metadata features (e.g., timestamps, user interactions), and network features (e.g., user connections, network structure).

Select and engineer features that effectively capture the characteristics and patterns associated with spam content and fake user accounts.

C) Logistic Regression Algorithm

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical .For example ,To predict whether an email is spam (1) or (0) .Whether the tumor is malignant (1) or not (0)Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time .From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

D) SVM Algorithm:

In machine learning, support-vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems. Developed at AT&T Bell Laboratories by Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997), it presents one of the most robust prediction methods, based on the statistical learning framework or VC theory proposed by Vapnik and Chervonekis (1974) and Vapnik (1982, 1995). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

Result :

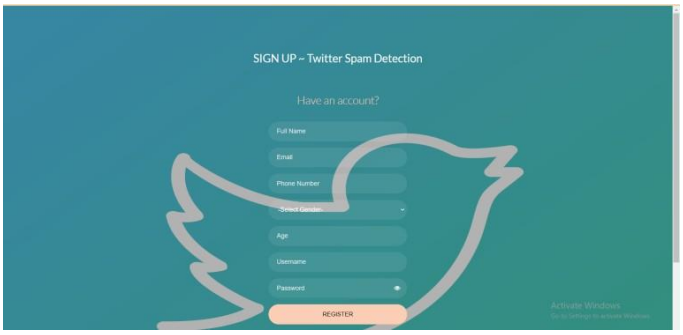

Fig 4: Prediction page

The final prediction system has been incorporated into the HTML CSS GUI application for the car price prediction.

The proposed prediction model has been evaluated on the test subset and model achieved overall accuracy of 92%.

Robust to Overfitting: Random Forest reduces overfitting by aggregating predictions from multiple decision trees, which helps to generalize well on unseen data.

Handling High-Dimensional Data: Random Forest performs well even with high-dimensional data since it randomly selects a subset of features at each node, focusing on relevant features and reducing the impact of irrelevant ones.

Robustness to Outliers: Random Forest is less sensitive to outliers compared to individual decision trees because it considers multiple trees and averages their predictions.

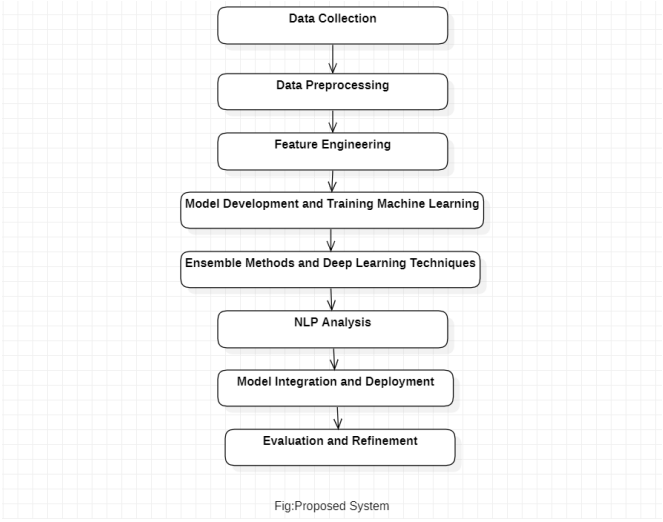

Fig:Proposed System

4. CONCLUSIONS

In this paper, we have trained our model with used cars data set to predict the price. Here we have used the Random forest algorithm and we got the accuracy 92%.The main limitation of this studyis the low number of records that have been used. In future work, we intend to collect more data related to spam messages and links and to use more advanced techniques.

REFERENCES

[1].Yuejun Li, Xiao Feng, and Shuwu Zhang. Detecting fake reviews utilizing semantic and emotion model. In Information Science and Control Engineering (ICISCE), 2016 3rd International Conference on, pages 317– 320. IEEE, 2016.

[2].Nitin Jindal and Bing Liu. Review spam detection. In Proceedings of the 16th international conference on World Wide Web,pages 1189– 1190. ACM, 2007.

[3]. E. I. Elmurngi and A.Gherbi, "Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques," Journal of Computer Science, vol. 14, no. 5, pp. 714– 726, June 2018.

[4]. Alimuddin Melleng, Anna-Jurek Loughrey, Deepak P, Text Representation Based on Sentiment and Emotion for Detection of Fake Reviews.

[5]. Chengai Sun, Qiaolin Du and Gang Tian, Exploiting Product Related Review Features for Fake Review Detection, Mathematical Problems in Engineering, 2016.