Unmasking Review Manipulation on E-Commerce Platforms

Sasidhar Reddy Kasa[1], Sruthi Ravuri[2*], Snikitha Polisetty[3], Kavya Akkineni[4] and D. S. Bhupal Naik[5]
{sasidharreddykasa341@gmail.com[1], sruthiravuri3@gmail.com[2], snikitha.polisetty222085@gmail.com[3], kavya.chowdarykks@gmail.com[4], dsbhupal@gmail.com[5]}

Department of Advanced Computer Science and Engineering, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Vadlamudi, Guntur (Dt)-522213, Andhra Pradesh, India[1, 2, 3, 4, 5]

Abstract. The explosion of online reviews has greatly impacted customer choices and company reputations. Then again, there are fake reviews on these sites across the board. The objective of this paper is to Use NLP with supervised machine learning in order to Detect fake reviews. The procedure includes the text pre-processing tokenization, stop words removal, and stemming and the feature extraction based on TF-IDF. Finally, the genuine and fake reviews are classified using a pretrained classification model. Furthermore, a user of the system can access the same through an OTP based user authentication. The experimental results show a good detection performance of fake reviews, and contribute to the trust of online review systems.

Keywords: Fake review detection, machine learning, natural language processing, TF-IDF vectorization, OTP authentication, text classification, online review analysis.

1 Introduction

In today's online era, online reviews serve as the foundation of consumer decisions and shape the reputation of companies. Websites such as Amazon, Yelp and TripAdvisor rely heavily on people giving star ratings to rate everything and anything from products to services, to help consumers make informed decisions. However, the increase in counterfeit reviews are posing a grave problem, casting doubts over the reliability of the services and eroding confidence from the consumers [3]. Maliciously misleading reviews, written to only damage others, are also used to deliberately lower ratings of products, deceive consumers and undermine competition. The fraudulent chain of events not only distorts the market but poses a serious danger to the legitimacy of online providers.

Traditional approaches used to identify fake reviews, manual moderation and rule-based systems, are proving to be inadequate when considering the quantity and the quality of the user generated content available. Human moderation happens on much slower time scales and at higher cost, and rule-based systems cannot account for dynamic changes in strategies by the fraudulent reviewer [9]. These restrictions denote the necessity to adopt advanced automatic systems for fighting this type of reviews. Towards coping with this important topic, this paper presents a machine-learning method for detecting spurious reviews. The system proposed makes use of Natural Language Processing (NLP) for pre-processing and analysing of review text and the extracted features are fed to a supervised learning-based classification. The system

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357932

can provide high throughput processing to learn new forms of malicious activity due to the automated detection. The platform also includes a secure OTP (One-Time Password)-based verification system that assures that only authorised users have access to the system and is more reliable and secure.

The presented method has several merits over former methods. First, it greatly improves the detection accuracy by detecting subtle linguistic information and behavioural clues in fake reviews [5]. Second, its scalability was very high, able to handle millions of reviews in real time, making it suitable for large online platforms. Third, the solution is flexible – the machine learning models can be retrained against fresh data to match new tactics adopted by false reviewers. Fourth, especial user interface helps rapid use e.g., user can input review data to analyse and get an instant feedback on whether the review is fake.

The use of this research goes beyond the technical, and its impact is of major value to consumers, businesses, and internet sites. To consumers, the system offers a sure means of filtering out false reviews, allowing them to make informed purchasing decisions. To businesses, it protects their reputation by detecting and removing false reviews that may damage their reputation. To internet sites, it improves overall trust, with more users willing to register and leave honest feedback.

This paper is structured as follows: Section II is an in- depth literature review of the current methods in detecting fake reviews, their weaknesses and strengths. Section III presents the methodology applied in the proposed system, namely text preprocessing, feature extraction, and classification. Section IV describes the experimental results, illustrating the performance of the system in detecting fake reviews. Section V describes the general implications of the study, and Section VI concludes with directions for future research to improve the system and overcome future challenges. Lastly, this paper is part of the ongoing research in the fight against manipulative reviews through the development of a strong, scalable, and reliable machine learning-based approach. Through the use of sophisticated NLP methods and secure verification procedures, the system not only increases the authenticity of online reviews but also contributes to a safer and more transparent online marketplace.

2 Literature Survey

Prior work on fake review detection can be generally divided into three types: linguistic analysis, machine learning based approach, and hybrid method. In detecting fake reviews, language analysis is crucial, as it exposes certain tendencies in the language used [7]. There are a few linguistic features that fake reviews use that can be exploited to distinguish them from genuine reviews. Rather, the expectation is that their surface form should comply with a number of simple well-known heuristic patterns, such as they often include hyperbolic expressions, like “the best” or “absolutely amazing,” and redundant phrases whose only effect is to artificially emphasize their point [4], [10]. They could also show mixed feelings, (pan hands / Getty Images as in the text of the review doesn’t jibe with the number of stars or the context. Also noticed that fraudulent reviews tend to use more superlatives (e.g., “perfect” and “incredible”) and less pronouns (e.g., “I” and “we”) as they often lack personal experiences [12]. By detecting these linguistic markers and using advanced methods of text analysis we are able to effectively detect fake content and linguistic analysis is one of the main pillars in the architectures of fake review detectors. Fig 1 shows Linguistic Patterns in Fake Reviews.

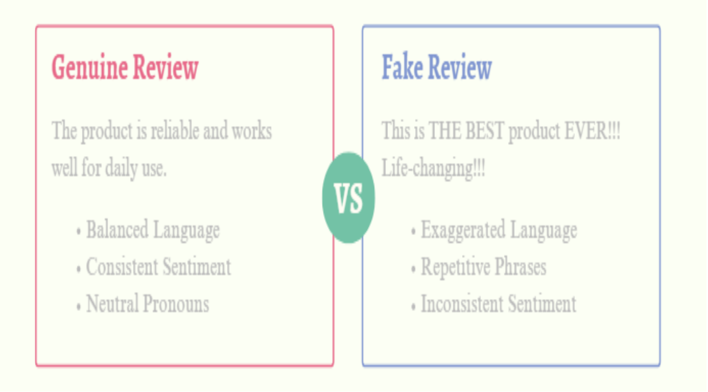


Fig. 1. Linguistic Patterns in Fake Reviews.

Due to their ability to automatically learn patterns in data, machine learning techniques have been used more and more in recent times for detecting fake reviews. SVM, Random Forests and Neural Networks are effective to recognize reviews as genuine or fake by learning weights to lexical and syntactic features. Fig 2 shows Machine Learning Techniques for Fake Review Detection demonstrated that SVM was able to detect fabricated reviews when trained using characteristics, such as word frequency and grammatical patterns [8]. They are more reliable and scalable than classical rule-based systems, a fact that is key for the war against fake opinions on the internet.

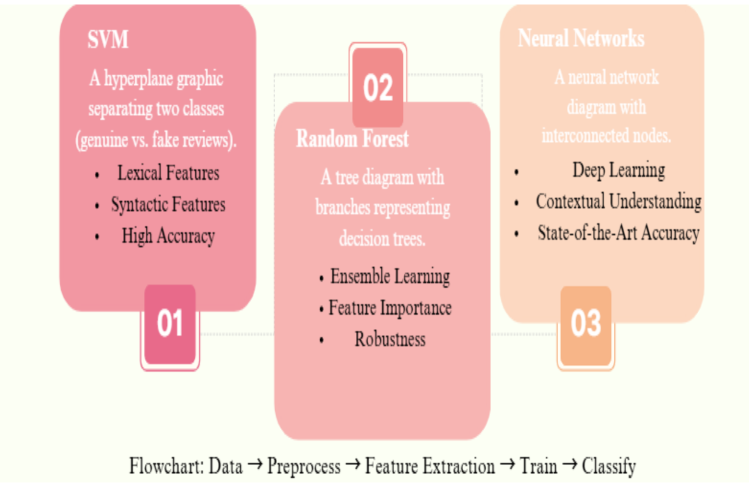

SVM
A hyperplane graphic separating two classes (genuine vs. fake reviews).
• Lexical Features
• Syntactic Features
• High Accuracy
01
02
Random Forest
A tree diagram with branches representing decision trees.
• Ensemble Learning
• Feature Importance
• Robustness
Neural Networks
A neural network diagram with interconnected nodes.
• Deep Learning
• Contextual Understanding
• State-of-the-Art Accuracy
03
Flowchart: Data → Preprocess → Feature Extraction → Train → Classify

Fig. 2. Machine Learning Techniques for Fake Review Detection.

Hybrid models for fake review detection use a combination of linguistic information with machine learning to increase the detection accuracy. They evaluate review text, sentiment, grammar, as well as review metadata such as reviewer identity, timestamp and ratings history. It was developed a text analysis model consolidated with behaviour aspects that can identify overstated language, conflicting views, and unexpected behaviour of a reviewer [11, 13]. Such a joint approach would ensure the best trade-off in terms of dependability of being able to detect fake reviews and to reinforce trust of customers in the digital ecosystems. Fig 3 shows Hybrid Approach: Combining Linguistic and Behavioural Features.

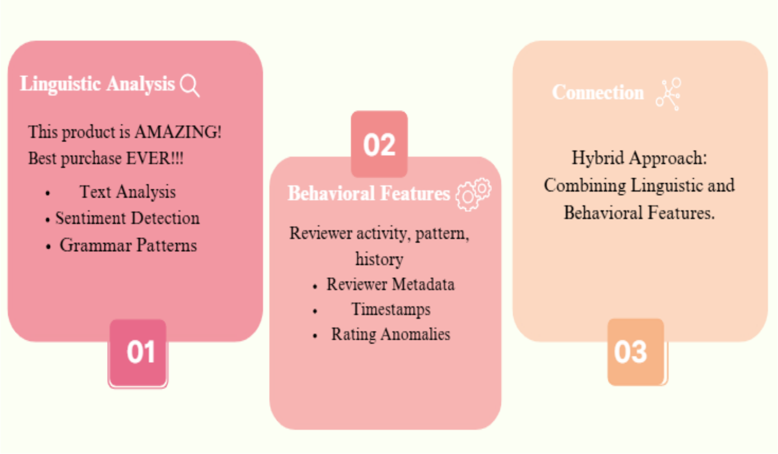


Fig. 3. Hybrid Approach: Combining Linguistic and Behavioral Features.

3 Methodology

The system we advocate will help differentiate real and phony e-reviews more efficiently. It uses secure user authentication and state of the art Natural Language Processing (NLP), in combination with machine learning algorithms. With a robust authentication procedure, the secured access is provided for the users and with the review text analysis, the significant features are extracted, which are important for correct classification. The architecture is completely modular and with all parts making big impact in delivering reliable and accurate result. The system is mainly composed of the following four subsystems:

- User Authentication: Provides secure access with an OTP based authentication.
- Text Pre-processing: Clean and process raw review text to be used for feature extraction.
- Feature Extraction: It transforms the pre-processed text into numeric features with TF-IDF vectorization [6].
- Classification – Uses a pre-trained machine learning model to label the reviews as real or fake.

Each component resolves important issues with respect to the review classification process i.e. protecting the data, reducing noise of the text, and selecting an appropriate feature representation. The combination of these modules leads to a complete (end-to-end) solution for spotting fake reviews in the contexts of e-commerce, hospitality, and social media. Fig 4 shows Proposed System Architecture for Genuine/Fake Review Classification.

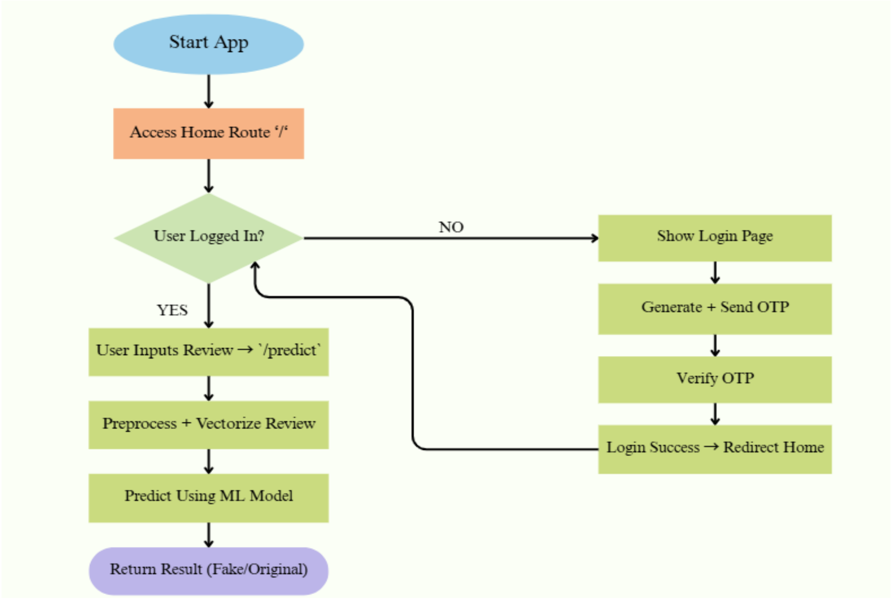

Start App
Access Home Route '/'
User Logged In?
NO
YES
Show Login Page
Generate + Send OTP
Verify OTP
Login Success → Redirect Home
User Inputs Review → '/predict'
Preprocess + Vectorize Review
Predict Using ML Model
Return Result (Fake/Original)

Fig. 4. Proposed System Architecture for Genuine/Fake Review Classification.

3.1 User Authentication

The system uses a secure OTP-based authentication mechanism to ensure that only authenticated users can access functionalities such as review classification.

The process includes user registration, OTP-based login, verification, session management, and email integration. The Fig 5 shows the login process where users input their email credentials to initiate access. Fig 6 demonstrates the OTP generation mechanism, where a six-digit one-time password is created and sent to the user's email. Fig 7 presents the OTP verification step, validating the user's input against the session-stored OTP for secure login. Fig 8 depicts the session management framework that maintains user authentication across protected routes. Finally, Fig 9 consolidates the entire user authentication process, showcasing the interaction between registration, OTP-based login, verification, and session control to ensure only legitimate users access the system.

3.1.1 User Registration

1. Users register by providing their name, email, and password via the /register route.
2. The data is stored in an Excel file (users.xlsx) for persistence.
3. Duplicate email addresses are not allowed.
4. Successful registration records the user's details along with the registration date.

Key Steps:

1. Validate input fields.
2. Check if email is already registered.
3. Save new user details to the Excel file.

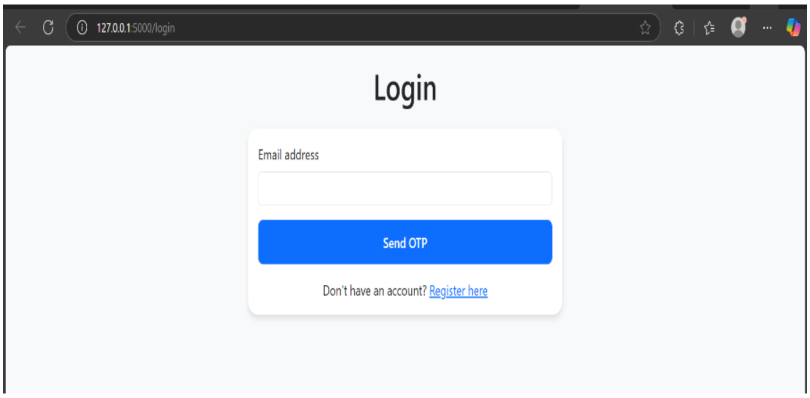


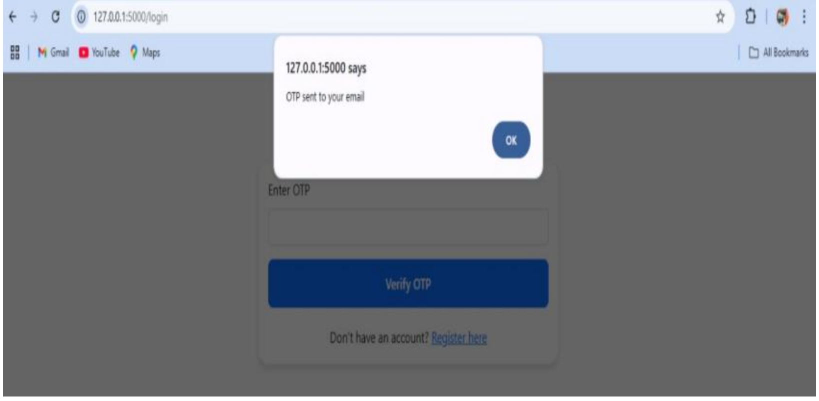
Fig. 5. Login Process.

3.1.2 OTP-Based Login

- Users log in by providing their email through the /login route.
- The system generates a 6-digit OTP and sends it using the send otp email function.
- The OTP is stored in the Flask session.

Key Steps:

1. Validate email.
2. Check registration status.
3. Generate and store OTP.
4. Send OTP via SMTP.


127.0.0.1:5000 says

OTP sent to your email

Fig. 6. OTP Generation.

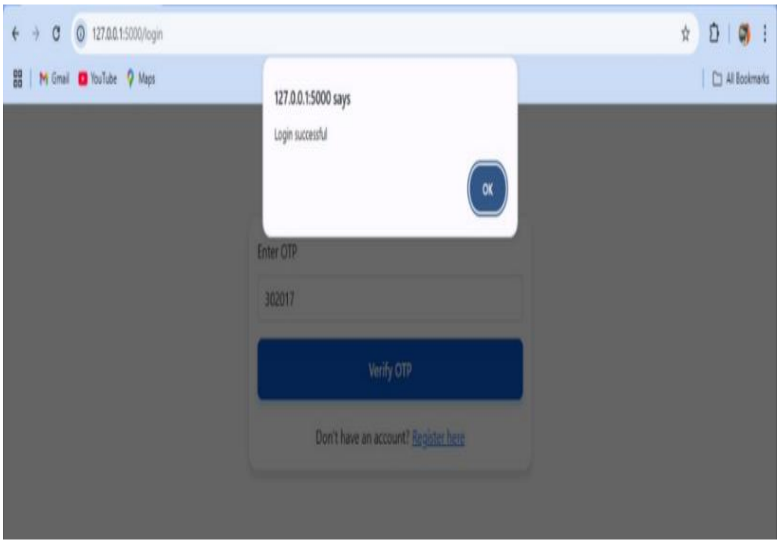
3.1.3 OTP Verification

- Users submit OTP via the /verify-otp route.
- The OTP is compared with the session-stored OTP.

- Valid OTPs allow login and session email storage.

Key Steps:

1. Validate OTP input.
2. Compare OTP with session.
3. Log in and store user email.



127.0.0.1:5000 says

Login successful

Enter OTP

302017

Don't have an account? Register here

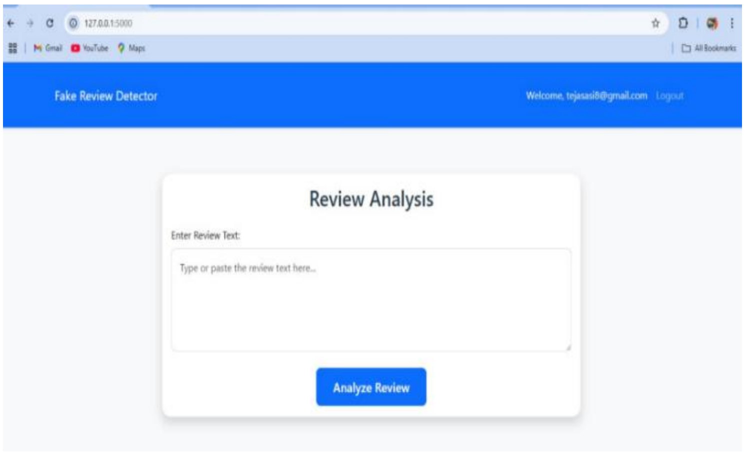
Fig. 7. OTP Verification.

3.1.4 Session Management

- Flask sessions are used to track authenticated users.
- Logged-in emails are saved for protected route access.
- Users can log out via the /logout route.

Key Steps:

1. Store user email after login.
2. Verify session for access control.
3. Clear session on logout.



Fake Review Detector

Welcome, tejasasi8@gmail.com Logout

Review Analysis

Enter Review Text:

Type or paste the review text here...

Analyze Review

Fig. 8. Session Management.

3.1.5 Email Integration

- Uses smtplib for sending OTPs via Gmail’s SMTP server.
- Emails include OTP and usage instructions.

Key Steps:

1. Compose email (subject, body, recipient).
2. Connect to SMTP and authenticate.
3. Send email and handle errors.

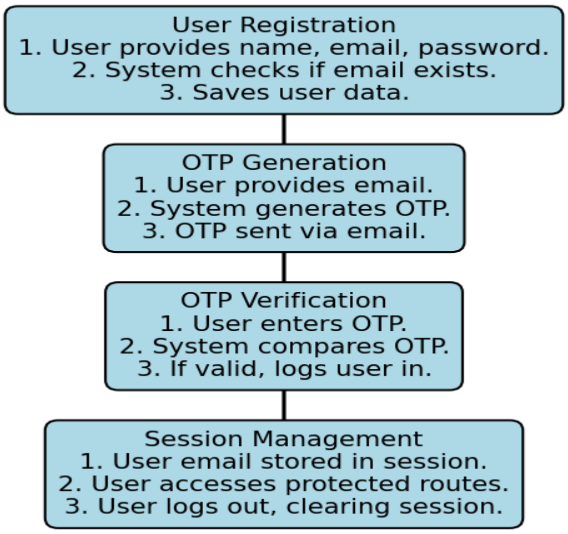

User Registration
1. User provides name, email, password.
2. System checks if email exists.
3. Saves user data.
OTP Generation
1. User provides email.
2. System generates OTP.
3. OTP sent via email.
OTP Verification
1. User enters OTP.
2. System compares OTP.
3. If valid, logs user in.
Session Management
1. User email stored in session.
2. User accesses protected routes.
3. User logs out, clearing session.

Fig. 9. User Authentication Process.

3.2 Text Preprocessing

Text preprocessing is essential for transforming raw reviews into clean input. The process involves:

1. Tokenization: Breaking the text into individual words.
2. Stop word Removal: Removing common, non-informative words.
3. Stemming: Reducing words to their root form (e.g., “running” to “run”).

These steps reduce noise, standardize input, and enhance the quality of data used for feature extraction. Fig 10 shows Text Preprocessing Flowchart.

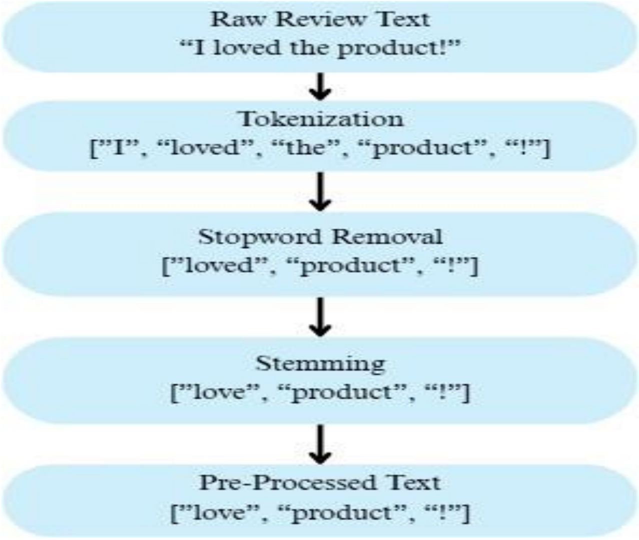


Fig. 10. Text Preprocessing Flowchart.

3.3 Feature Extraction

Feature extraction converts preprocessed reviews into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. This assigns weights based on the importance of terms across the corpus. The process includes:

1. TF (Term Frequency): Frequency of a word in the document.
2. IDF (Inverse Document Frequency): How rare a word is in the corpus.
3. TF-IDF Score: Product of TF and IDF indicating word importance.

TF-IDF vectors are generated using the job lib loaded vectorizer and are suit- able for ML-based classification. Fig 11 shows Feature Extraction Process.

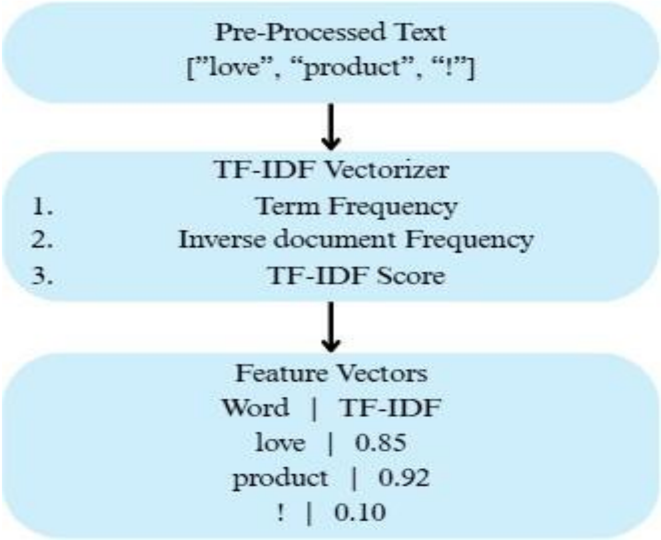


Fig. 11. Feature Extraction Process.

3.4 Classification

The final step classifies reviews as genuine or fake using a pre-trained machine learning model loaded via joblib. Fig 12 shows Classification Process in the System. The model processes TF-IDF feature vectors and outputs a prediction based on learned patterns. Steps:

1. User submits review text.
2. Text is preprocessed (tokenization, stop word removal, stemming).
3. TF-IDF vectorizer transforms the cleaned text.
4. The model predicts whether the review is genuine or fake.

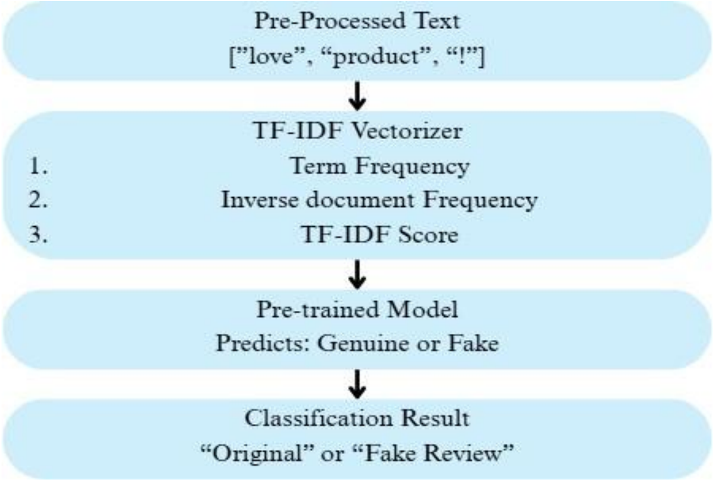


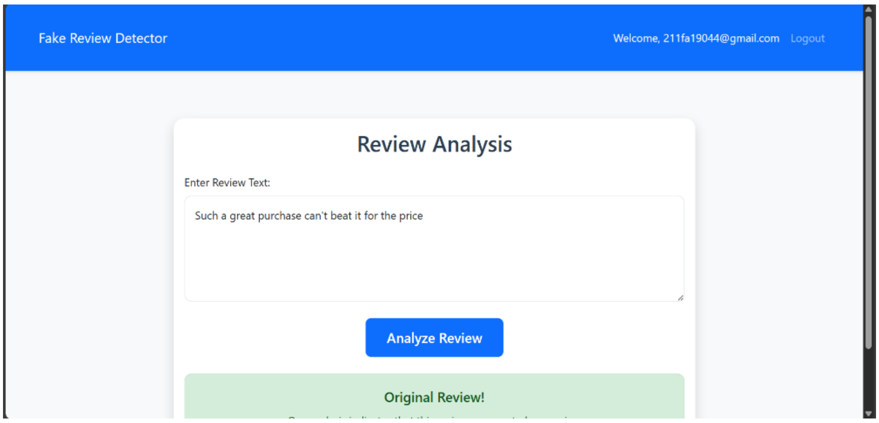
Fig. 12. Classification Process in the System.

4 Results and Evaluation

The proposed system achieves high accuracy in detecting fake reviews, as demonstrated by the following evaluation metrics: Table 1 shows Model Evaluation on Different Datasets. The Fig 13 & Fig. 14 shows detected review and graphical representation.

Table 1. Model Evaluation on Different Datasets.

Dataset	Accuracy	Precision	Recall	F1-Score
Original Review (OR)	92%	91%	93%	92%
Computer Generated (CG)	90%	89%	91%	90%



Review Analysis

Enter Review Text:

Such a great purchase can't beat it for the price

Original Review!

Fig. 13. Detected Review.

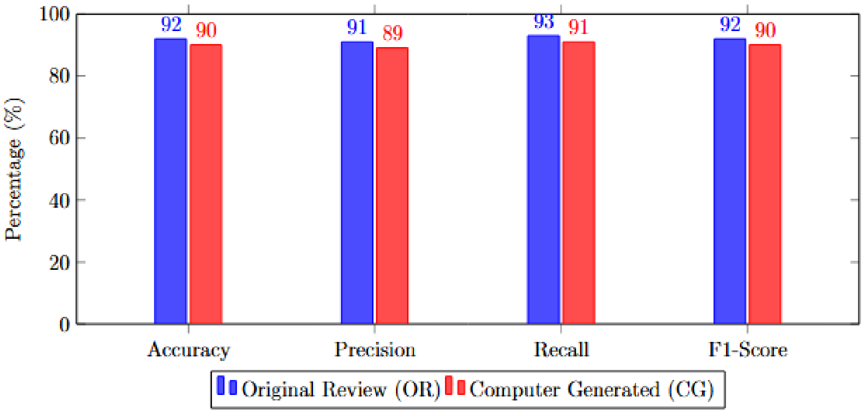


Fig. 14. Graphical Representation.

5 Conclusion

The proposed web based fake review detection system is neatly implementing NLP, ML, and secure user authentication in user-friendly Flask-framework [1], [2]. It provides registration and secure login with OTP based authentication system giving more security Flexibility By using our solution, you can make your system much more secure as it would support implementing OTP checking. Text pre-processing is performed on the user input reviews which includes tokenization, stop word removal, and stemming, and the reviews are further vectorized as numerical vectors through TF-IDF (term frequency-inverse document frequency) vectorization [6]. These features are then entered into a pre-trained machine learning model to determine whether a review is genuine or fake.

The use of email-based OTP verification enables the prediction service to be accessed by only authentic users, and is secure and efficient. With user id and registration data stored in an Excel file format, the system is easy to deploy and maintains important records without using complex databases.

Given that it possesses modular architecture, real-time prediction model, and extendable design, the system backward a sound weapon to combat the growing threat of such content as fake online reviews. It can also be enhanced with deep learning-based machine translation for better accuracy, cloud based infrastructure to scale up, and support for more languages to serve a broader user base. Collectively, this work demonstrates the successful deployment of AI-enabled technologies for real-world applications, retaining user privacy and system usability [15, 14].

References

[1] Ronacher, A.: Flask: Web Development with Python. Flask Documentation (2010)

[2] Pallets Projects: Flask Documentation. https://flask.palletsprojects.com/ (2023)

[3] Jurafsky, D., Martin, J.H.: Speech and Language Processing, 3rd edn. Pearson Education (2021)

[4] Mikolov, T., et al.: Distributed Representations of Words and Phrases and their Compositionality. In: NeurIPS (2013)

[5] Saif, H., et al.: On Stopword Removal and Data Sparsity for Sentiment Analysis of Twitter. In: LREC (2014)

[6] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24(5), 513–523 (1988)

[7] Ott, M., et al.: Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In: ACL, pp. 309–319 (2011)

[8] Li, F., et al.: Learning to Identify Review Spam. In: IJCAI, pp. 2488–2493 (2011)

[9] Kumar, A., Sharma, R.: Fake review detection using supervised learning approach. J. Big Data 8(1), 1–21 (2021)

[10] Resnick, P.: RFC 2821 - SMTP. Internet Engineering Task Force (2001)

[11] Gao, W., et al.: A Hybrid Deep Learning Approach for Fake Review Detection. In: CIKM (2022)

[12] Naik, D.S.B., et al.: Parallel processing of enhanced K-means using OpenMP. In: ICCIC, pp. 1–4 (2013). https://doi.org/10.1109/ICCIC.2013.6724291

[13] S., B.N.D., et al.: Comparative analysis of ML algorithms for IIoT anomaly detection. Int. J. Inf. Retr. Res. 12(1), 1–55 (2022). https://doi.org/10.4018/ijirr.298647

[14] Kogut, A., et al.: mmWave Antenna Based on WG Mode DR for IoT. Wireless Commun. Mob. Comput. (2022). https://doi.org/10.1155/2022/5877263

[15] Naik, D.B., et al.: Detecting Fake Instagram Profiles. In: iQ-CCHESS (2023). https://doi.org/10.1109/iQ-CCHESS56596.2023.10391800