A

Mini Project Report

on

Fake Product Review Detection

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF ENGINEERING

IN

Computer Science & Engineering

Artificial Intelligence & Machine Learning

by

Shreyas Mhatre(23106135)

Krish Sharma(23106115)

Parth Shah(23106087)

Varun Panchal(23106050)

Under the guidance of

Prof. Tejashri Kolhe

Department of Computer Science & Engineering

(Artificial Intelligence & Machine Learning)

A. P. Shah Institute of Technology

G. B. Road, Kasarvadavali, Thane (W)-400615

University Of Mumbai

2024-2025

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

CERTIFICATE

This is to certify that the project entitled "**Fake Product Review Detection"** is a Bonafide work of Shreyas Mhatre (23106135), Krish Sharma (23106115), Parth Shah (23106087), Varun Panchal (23105060) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Tejashri Kolhe	Dr. Jaya Gupta
Mini Project Guide	Head of Department

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

PROJECT REPORT APPROVAL

This Mini project report entitled "**Fake Product Review Detection**" by **Shreyas Mhatre, Krish Sharma, Parth Shah, Varun Panchal** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

External Examiner: ___________________

Internal Examiner: ___________________

Place: APSIT, Thane

Date:

DECLARATION

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Shreyas Mhatre	Krish Sharma	Parth Shah	Varun Panchal
(23106135)	(23106115)	(23106087)	(23106050)

ABSTRACT

The **Fake Product Review Detection** mini-project aims to develop a machine learning-based system that can automatically identify fake reviews posted on e-commerce platforms. With the increasing reliance on online reviews for purchasing decisions, the prevalence of fake reviews has become a significant issue, leading to consumer misinformation and brand reputation damage. This project addresses the challenge of detecting fake reviews by utilizing natural language processing (NLP) techniques and machine learning algorithms. The process involves collecting a dataset of product reviews, preprocessing the data to clean and structure it, and extracting key features such as sentiment, review length, and keyword frequency. Several machine learning models, including Logistic Regression, Random Forest, and Support Vector Machines (SVM), are trained to classify reviews as either fake or real based on these features. The models are then evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure their effectiveness. The final output of the project is a classification system capable of flagging suspicious reviews, helping consumers make informed purchasing decisions and protecting businesses from fraudulent activities. This project highlights the practical application of machine learning and NLP in tackling real-world problems associated with online review manipulation.

Index

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

In the digital age, **online reviews** play a critical role in shaping consumer behavior. Be it for products on e-commerce platforms, restaurants on food delivery apps, or services offered on freelancing websites—users often depend on the **opinions and experiences shared by others** to make informed decisions. However, the growing reliance on online reviews has also given rise to a significant problem: the proliferation of **fake or manipulated reviews**. These reviews are often generated to **falsely promote** or **unjustly discredit** products and services, leading to misinformation and distrust.

With the increasing sophistication of **AI-generated content**, it has become even more challenging to distinguish between **genuine user feedback** and **synthetically produced text**. Traditional rule-based systems fail to detect such reviews reliably due to their inability to adapt to the dynamic nature of language and manipulation tactics.

This mini-project aims to address this issue by implementing a **Fake Review Detection System** using the **Ollama API** and the **Mistral-Nemo large language model (LLM)**. The goal is to create an intelligent, explainable solution that can **analyze a review's tone, structure, content, and style** to provide a probability-based judgment on its authenticity.

Fake reviews are often **indistinguishable from genuine ones** due to their realistic phrasing and emotional tone. Manual moderation of reviews is time-consuming and not scalable. Furthermore, generic detection systems that rely solely on keyword spotting or review length cannot keep up with **modern language models** that produce highly nuanced fake content.

Therefore, the core problem this project aims to solve is:
How can we accurately and efficiently determine the authenticity of a review using AI-driven natural language analysis?

The objective of this mini project is to design and develop a **web-based application** that allows users to **input any product or service review** and instantly receive feedback about its authenticity. The system will:

- Send the review text to a **locally hosted LLM (Mistral-Nemo)** using the **Ollama API**.
- Use a custom prompt to instruct the LLM to **analyze linguistic and emotional patterns**, **specificity**, **overuse of praise**, **errors**, and other markers of fakeness.
- Return a **probability score (0–100%)** and a **biased classification** (Fake / Cannot Determine / Real).
- Provide a **structured explanation** that can be viewed optionally through a dropdown in the UI.

This mini project is designed to be **lightweight**, **user-friendly**, and **entirely offline-capable** (since Ollama runs locally). It uses:

- **HTML/CSS/JavaScript** for the front-end interface.
- **Python Flask** as a minimal backend to handle API routing.
- **Ollama** to host and serve the **Mistral-Nemo model**, which is responsible for review analysis.

The system is not trained on a dataset; instead, it leverages the **generalized reasoning capability** of the LLM to perform subjective analysis. The user interface supports **live feedback**, **loading states**, and **structured result rendering** for clarity.

This project showcases how powerful open-source tools and large language models can be used to address real-world challenges such as misinformation and content authenticity.

By offering **real-time fake review detection** without relying on a large dataset or server-dependent infrastructure, it becomes a **practical solution** for individual users, small businesses, or educational demos.

Moreover, it demonstrates a **simple yet effective integration of modern AI** in web applications using basic frameworks like Flask and JavaScript—making it ideal for a mini project that balances technical learning and social relevance.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1-HISTORY

History of Fake Review Detection

The rise of online marketplaces in the early 2000s brought about a dramatic shift in how consumers made purchasing decisions. As websites like Amazon, TripAdvisor, and Yelp gained popularity, so did their user review sections, which allowed customers to share their experiences. While these platforms initially relied on user trust and transparency, it soon became apparent that fake reviews were being used to manipulate ratings, promote products, or discredit competitors. This problem laid the foundation for a new area of research: fake review detection.

Early Methods and Manual Detection

In the beginning, review platforms implemented basic filters and manual moderation to detect spam. Simple rules such as detecting duplicate reviews, monitoring IP addresses, or flagging overly promotional language were used. These systems were often static and easy to bypass, relying heavily on human intervention and heuristics.

By the mid-2000s, researchers began exploring supervised machine learning techniques. The release of early labelled datasets, such as the Ott et al. (2011) corpus using Amazon Mechanical Turk (AMT), enabled the development of classifiers such as Naïve Bayes, Decision Trees, and Support Vector Machines (SVM). These models focused primarily on textual features like word frequency, sentiment polarity, and review length. However, such datasets were small and often crowdsourced, raising concerns about real-world applicability.
Rise of Behavioural and Metadata Analysis

Around 2012–2015, researchers began expanding the feature space to include behavioural data, such as reviewer history, posting time, and rating deviation. This shift was inspired by the idea that spam detection should not rely on text alone, but also examine how and when reviews are posted. This led to more robust classifiers that combined textual and metadata features, often achieving improved accuracy across domains.

Introduction of Unsupervised and Graph-Based Techniques

Due to the scarcity of labelled data, researchers explored unsupervised methods like clustering, outlier detection, and graph-based algorithms. These methods aimed to identify suspicious reviewer groups or detect anomalies in rating behaviour. This phase of research also introduced semi-supervised learning, which allowed systems to learn from limited labelled data and large unlabelled corpora, increasing scalability.

Deep Learning and Transformer Models

From 2017 onward, the field saw a major evolution with the introduction of deep learning and transformer architectures. Models like CNNs, LSTMs, and later BERT, RoBERTa, and DistilBERT revolutionized fake review detection by understanding context and semantics better than traditional models. These models required minimal feature engineering and could detect subtle patterns in writing style, tone, and emotion.

Transformer-based models, especially fine-tuned versions like RoBERTa, achieved state-of-the-art accuracy but came at the cost of computational complexity and lack of transparency. Researchers also began exploring explainability and domain transferability, trying to make these models more practical and generalizable.

Current Trends

Today, fake review detection continues to evolve. The rise of LLMs (Large Language Models) like Mistral-Nemo, combined with local APIs like Ollama, has made it possible to build lightweight, explainable systems without needing massive datasets or online dependencies. These models can reason about a review in a zero-shot setting, marking a new era of AI-assisted, user-friendly, and privacy-preserving detection systems—the direction our current project embraces.

2.2-LITERATURE REVIEW

The task of detecting fake reviews has received increasing attention in recent years due to the growing impact of online reviews on consumer behaviour. As platforms like Amazon, Yelp, and TripAdvisor became integral to purchasing decisions, malicious actors began exploiting these systems by posting fake reviews—either to unjustly promote products or to damage the reputation of competitors. The body of research around fake review detection has since evolved through multiple stages, ranging from simple rule-based filtering systems to sophisticated machine learning and deep learning models.

In early studies, researchers focused primarily on textual features derived from the review content. These included bag-of-words representations, n-grams, part-of-speech tags, and sentiment polarity. The underlying assumption was that fake reviews often contained unusual or extreme language patterns. However, it soon became clear that relying on linguistic features alone was insufficient, especially as spammers adopted more natural and contextually sound writing styles. This limitation encouraged researchers to explore behavioural features, which consider metadata about the reviewer and their interactions with the platform. These features include rating deviation, review burstiness, frequency of reviews, and reviewer history, offering critical insights into user behaviour patterns that could indicate deception.

As the field progressed, machine learning techniques became the core of fake review detection systems. Supervised learning models such as Support Vector Machines (SVM), Naïve Bayes, Random Forests, and Decision Trees were widely used. These models were trained on labelled datasets, such as YelpCHI and TripAdvisor, to classify reviews as fake or genuine. Research showed that combining linguistic and behavioural features significantly improved the accuracy of these models. However, the performance of supervised learning models was highly dependent on the quality and volume of labelled data, which was often difficult to obtain. Crowdsourced datasets like those from Amazon Mechanical Turk (AMT) raised concerns about authenticity, while proprietary datasets were limited in access and scope.

To overcome the scarcity of labelled data, unsupervised learning approaches were introduced. These methods, including clustering and graph-based techniques, aimed to detect suspicious patterns or reviewer groups without explicit labels. For example, clique percolation and K-means clustering were employed to identify reviewer communities exhibiting abnormal behaviours. Graph-based models treated the review ecosystem as a network of users, products, and reviews, detecting anomalous interactions indicative of spam. These methods proved useful, especially for real-time applications where manual labelling was not feasible.

Semi-supervised learning offered another solution, blending small labelled datasets with larger unlabelled ones. Techniques like Positive-Unlabelled (PU) learning, Expectation-Maximization, and ramp loss SVMs allowed systems to learn more flexibly and adapt to dynamic spam tactics. Some studies even integrated user review graphs with text classification to enhance detection accuracy, combining the best of both worlds.

More recently, the field has witnessed a surge in the use of deep learning, particularly with the rise of transformer-based language models. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as LSTMs and Bi-LSTMs, have been applied to detect fake reviews with impressive results. However, it is transformer models like BERT, RoBERTa, and DistilBERT that have revolutionized this area. These models can understand the nuanced structure of language, capture context effectively, and are less reliant on manual feature engineering. RoBERTa, for instance, has shown state-of-the-art performance on multiple fake review benchmarks, significantly outperforming classical ML approaches.

Despite these advancements, certain limitations persist. One of the biggest challenges is the issue of domain transferability. A model trained on hotel reviews may not perform well on restaurant or product reviews due to variations in language, context, and user intent. Additionally, the dynamic nature of fake review tactics—often referred to as "concept drift"—requires models to continuously adapt to new spam strategies. Another critical concern is explainability. While deep learning and transformer models offer high accuracy, they often behave like black boxes, making it difficult for users or regulators to understand why a particular review was flagged as fake.

Both the literature surveys reviewed—"Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review" and "Fake Reviews Detection: A Survey"—highlight the importance of combining multiple types of features, adopting flexible model architectures, and ensuring interpretability. They also emphasize the role of benchmark datasets such as YelpCHI, Dianping, and Amazon, each with its own advantages and drawbacks. While behavioural and linguistic features continue to play a significant role, the integration of semantic understanding through language models and the use of graph-based features are shaping the future of this field.

The insights drawn from this literature survey serve as the foundation for the current mini project. Unlike traditional ML-based approaches, our system uses a zero-shot method enabled by a large language model (Mistral-Nemo), hosted locally through the Ollama API. By prompting the model with a structured instruction set, the system analyses not only the surface content of a review but also deeper elements such as emotional tone, specificity, and exaggeration. This strategy enables the application to offer a real-time probability score, a biased conclusion, and a human-readable explanation—without the need for large training.

CHAPTER 3

PROBLEM STATEMENT

3. Problem Statement

With the growing influence of online product reviews, consumers increasingly rely on them to make purchasing decisions. However, the rise of fake reviews—either promotional or defamatory—has become a significant issue, misleading consumers and damaging the reputation of businesses. Fake reviews can misrepresent a product's quality, affecting both consumer trust and the integrity of e- commerce platforms.

The problem is to **develop an automated system** that can accurately distinguish between fake and genuine product reviews. Detecting fake reviews is challenging due to factors like **imbalanced datasets**, where genuine reviews far outnumber fake ones, making it difficult for machine learning models to identify subtle patterns of deception. Additionally, fake reviews can vary in form, including overly positive or negative sentiment, suspicious repetition, or generic content, which makes them hard to identify without a sophisticated approach.

Another challenge arises from **multilingualism** since e-commerce platforms are global, and reviews are often written in different languages. Detecting fake reviews across multiple languages further complicates the issue. Furthermore, **fraudulent tactics** continue to evolve, as individuals and businesses develop new methods to bypass detection systems.

This project aims to develop a machine learning-based solution that uses **textual analysis** and **natural language processing** to classify reviews as fake or real. By accurately detecting fraudulent reviews, this system will help maintain consumer trust and ensure businesses can protect their reputation in the digital marketplace.

CHAPTER 4

EXPERIMENTAL SETUP

4. Experimental Setup

4.1 Hardware Setup

The hardware setup required for the Fake Product Review Detection project typically involves the following components:

1. **Computing Device(CPU/GPU):**

   The core of the hardware setup is a computer or server with sufficient processing power to run machine learning algorithms and handle large datasets. For basic model training and testing, a standard CPU-based system with at least 8 GB of RAM and a modern Intel i5 or i7 processor can suffice. However, for deep learning models and larger datasets, a GPU-enabled machine with a CUDA-compatible GPU (such as an NVIDIA GTX or RTX series) will significantly speed up model training and processing time.

2. **Storage:**

   Given the large volume of text data (reviews), a system with sufficient storage space (SSD preferred for faster data access) is important. A storage capacity of at least 100 GB is recommended for datasets, model storage, and intermediate processing files. For more complex projects or larger datasets, storage may need to be expanded or use of cloud storage (like AWS S3 or Google Cloud Storage) might be considered.

3. **RAM:**

   Memory is critical when working with large datasets or running deep learning models. 16 GB or more of RAM is recommended to ensure smooth data processing, model training, and execution without excessive swapping.

4. **Internet Connectivity:**

   A reliable internet connection is necessary for downloading datasets, installing packages, and leveraging cloud-based platforms like Google Colab or AWS for training large models if required.

5. **Optional Cloud Computing:**

   For larger-scale projects or for training deep learning models, cloud computing services like AWS EC2, Google Cloud Platform, or Microsoft Azure can be used. These platforms provide on-demand access to high-performance CPUs/GPUs, scalable storage, and processing.

4.2 Software Setup

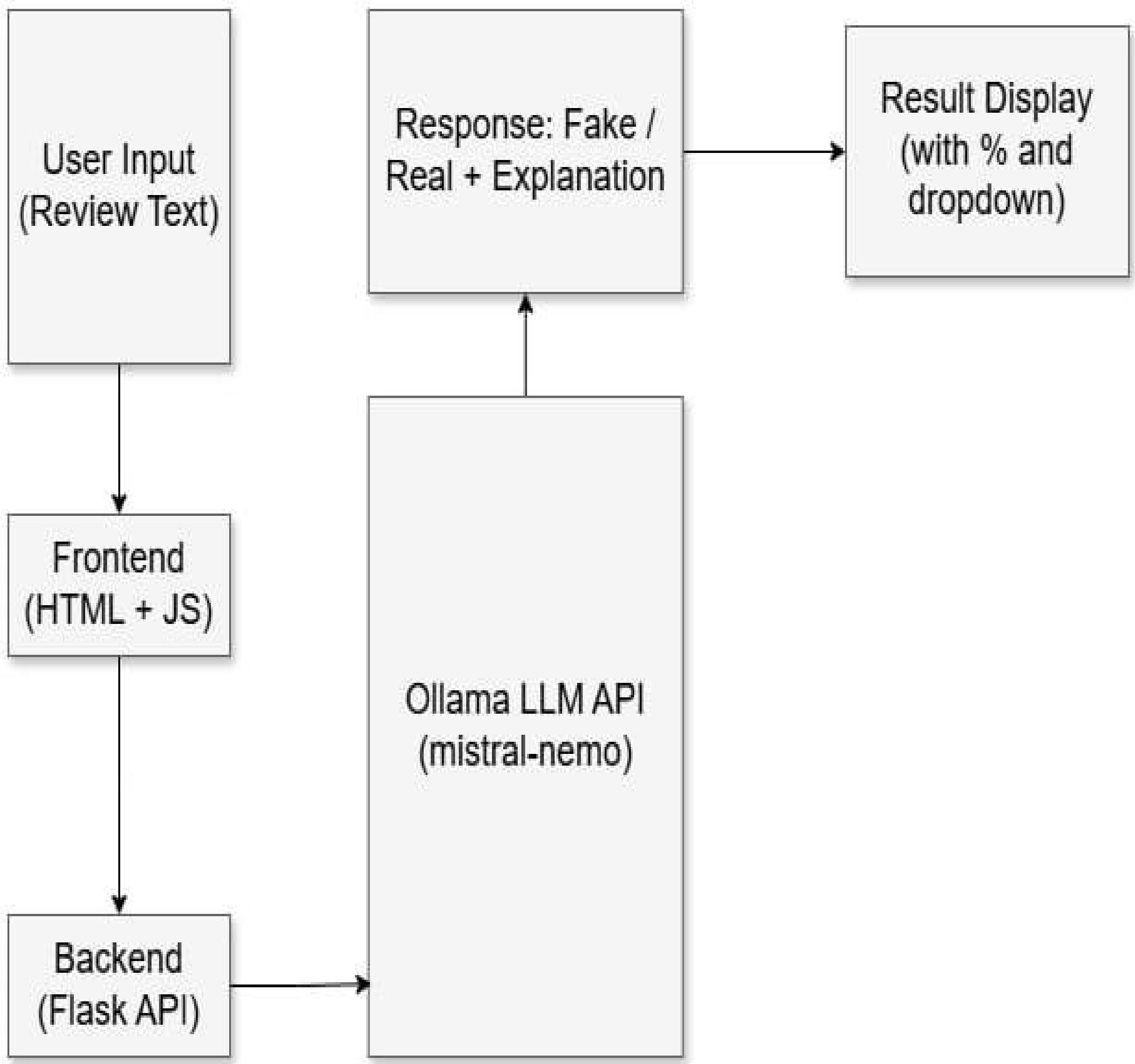
To develop a fake product review detection system, the following software setup is required:

1. Operating System: Windows, macOS, or Linux (Linux preferred for data handling).
2. Programming Language: Python 3.x, chosen for its extensive libraries for machine learning and natural language processing.
3. Package Manager: pip for managing Python libraries; Conda (Anaconda) for managing complex dependencies.
4. IDE/Text Editor: Use VS Code, PyCharm, or Jupyter Notebook for efficient coding and experimentation.
5. Libraries/Frameworks: Ollama Library used to load and run LLMs and AI Models locally on consumer grade devices.
6. Flask: Python library used for backend to handle API request and dispalying results dynamically on web page.
7. Version Control: Git for tracking code changes; GitHub or GitLab for hosting repositories.
8. Cloud Resources: Google Colab (free GPUs/TPUs) or AWS/Google Cloud for scalable computing; SQL/NoSQL databases for storing reviews.

CHAPTER 5
PROPOSED SYSTEM
& IMPLEMENTATION

5. Proposed system & Implementation

5.1 Block diagram of proposed system



5.2 Description of block diagram

1. Frontend (Client-Side)

   Technologies Used: HTML, CSS, JavaScript

   Components: Browser (User Interface)

   Role:

- The user accesses the web app through a browser.
- They click a button to generate a fake review.
- The request is sent to the backend (Flask).

2. Backend (Flask Server)

   Technology Used: Flask (Python)

   Role:

- Receives the request from the frontend.
- Processes the request and forwards it to the Ollama API (which handles the LLM processing).

3. Ollama (Local LLM Processing)

   Technology Used: Ollama (to run a local LLM like Mistral, Llama, etc.)

   Role:

- Receives a query from Flask (e.g., "Generate a fake product review").
- The LLM generates a realistic but fake review.
- Sends the generated review back to Flask.

4. Backend (Flask) - Response Handling

   Role:

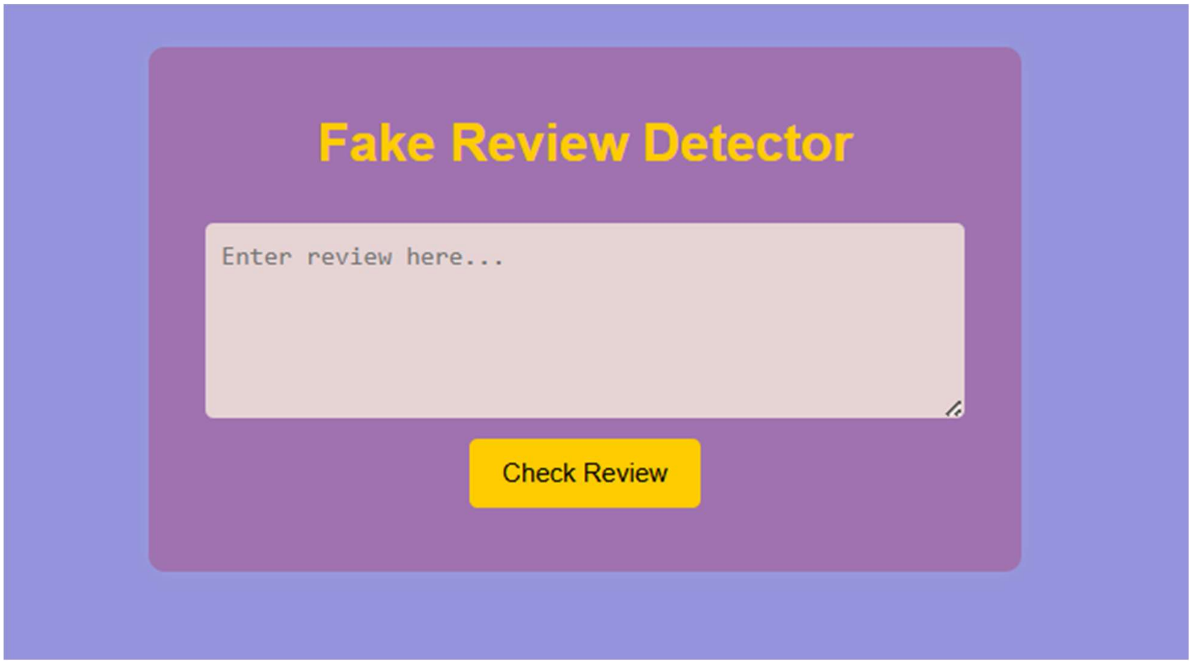
- Receives the fake review from Ollama.
- Sends the response (fake review) back to the frontend.

5. Frontend (Display Fake Review)

Role:

- Receives the fake review from the backend.
- Displays it on the webpage for the user.

5.3 Implementation



5.4 Advantages/ Application/ result table can be included in this subsection

Advantages of Fake Review Detection System:

1) Prevents Misinformation: Helps consumers make informed purchasing decisions.
2) Enhances Trust: Builds credibility for genuine products and sellers.
3) Improves Market Fairness: Reduces the influence of misleading marketing tactics.
4) Automated Detection: Saves time compared to manual review analysis.

Applications of Fake Review Detection:

1) E-commerce Platforms (Amazon, Flipkart, eBay): Detects fake product reviews.
2) Travel & Hospitality Industry (TripAdvisor, Booking.com): Identifies misleading hotel reviews.
3) Social Media & Online Forums: Flags manipulated comments or testimonials.
4) Corporate & Brand Reputation Management: Monitors and removes fake feedback.

Result Table:

Review Id	**Review Text**	**Fake/Real Prediction**	**Score(%)**
1.	"Amazing product! Highly recommended!"	Fake	85%
2.	"Worked fine for a week, then stopped. Disappointed."	Real	90%
3.	"Buy this now! Best deal ever! Wow!"	Fake	92%

CHAPTER 6

CONCLUSION

Conclusion

The Fake Product Review Detection mini project helps combat fraudulent online reviews, which mislead consumers and create unfair competition. As e-commerce and digital marketing grow, ensuring review authenticity is essential. This system uses Natural Language Processing (NLP) and Machine Learning (ML) to classify reviews as real or fake by analyzing sentiment, repetition patterns, and user behavior.

The project enhances transparency, protects consumers, and promotes fairness in online marketplaces. Automating the detection process saves time and resources compared to manual verification. It can be integrated into e-commerce, travel, hospitality, and social media platforms to improve content authenticity and manage brand reputation.

The system assigns a confidence score to each review, helping platforms flag high-risk fake reviews while minimizing false positives. Future enhancements could include deep learning, real-time detection, and cross-platform integration to improve accuracy and efficiency.

In conclusion, this system builds trust, ensures ethical business practices, and enhances digital marketplaces. As online shopping expands, such solutions will be crucial in maintaining authenticity and enabling informed consumer decisions.

References

Research paper

[1] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding Deceptive Opinion Spam by Any Stretch of the Imagination.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, 309–319.

[2] Rayana, S., & Akoglu, L. (2015). *Collective Opinion Spam Detection: Bridging Review Networks and Metadata.* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 985–994. DOI: 10.1145/2783258.2783370.

[3] Mukherjee, A., Liu, B., & Glance, N. (2012). *Spotting Fake Reviewer Groups in Consumer Reviews.* Proceedings of the 21st International Conference on World Wide Web, 191–200. DOI: 10.1145/2187836.2187863.

[4] Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). *Towards a General Rule for Identifying Deceptive Opinion Spam.* Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 1566–1576.

[5] Kumar, S., Spezzano, F., Subrahmanian, V. S., & Faloutsos, C. (2018). *Edge Weight Prediction in Weighted Signed Networks.* IEEE Transactions on Knowledge and Data Engineering, 29(12), 2573–2586. DOI: 10.1109/TKDE.2017.2731896.

[6] Jindal, N., & Liu, B. (2008). *Opinion Spam and Analysis.* Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM), 219–230.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention Is All You Need.* Advances in Neural Information Processing Systems, 30.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL-HLT, 4171–4186.

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners.* Advances in Neural Information Processing Systems, 33, 1877–1901.

[10] Zhang, Y., & Chen, X. (2020). *Explainable Fake Review Detection Using Deep Learning and Attention Mechanism.* IEEE Access, 8, 109556–109567. DOI: 10.1109/ACCESS.2020.3001364.

URLs

https://ieeexplore.ieee.org/abstract/document/9416474/
https://link.springer.com/article/10.1007/s10115-024-02118-2