FAKE REVIEW DETECTION USING DEEP LEARNING

Muthukumaran.S[1], Jenifer Nivetha.S[2], Priyanka.S[2], Ramya.TV[2]
Assistant Professor[1], Final Year[2]
Department of Information Technology
St.Joseph College of Engineering
Sriperumbudur, Chennai-602 117

Abstract - In the digital era, online reviews significantly influence consumer decisions, making them a target for manipulation through fake or deceptive reviews. This project focuses on detecting fake reviews from Google Maps using deep learning techniques, aiming to enhance trust and transparency for users relying on location-based services.

The system employs SerpAPI to collect real-time Google Maps reviews for various businesses and locations. These reviews are then preprocessed and analyzed using the DeepSeek R1 language model - a state-of-the-art, decoder-only transformer model with strong language understanding capabilities. DeepSeek R1 is utilized to classify each review as genuine or fake based on its linguistic patterns, sentiment coherence, and contextual indicators.

The proposed methodology explores prompt-based classification, making it efficient and suitable for real-world deployment even with limited computational resources. The system can be integrated into a full-stack application where users input a business or location, and the backend outputs annotated reviews with confidence scores.

By leveraging advanced deep learning models and scalable APIs, this project demonstrates a practical and intelligent solution to mitigate the impact of fraudulent online content, thereby contributing to the integrity of public review platforms.

I.INTRODUCTION

In today's digital world, user-generated content significantly influences consumer decisions, especially in the form of online reviews. While platforms like Google Maps provide a convenient way to assess businesses, they are increasingly vulnerable to manipulation through fake or deceptive reviews. These fraudulent reviews can mislead users, harm business reputations, and compromise the reliability of online information.

This project aims to address the growing concern of fake reviews by leveraging deep learning techniques. By analyzing patterns in review data and language structure, the system can differentiate between genuine and deceptive content. Our project is divided into two main components:

1. **Detection of fake Google Maps reviews** using a prompt-based approach with the DeepSeek R1 model and real-time data extraction via SerpAPI.
2. **Development of a full-stack web application** that generates automated reviews, which are then analyzed for authenticity.

Through this dual approach, we seek to explore both the detection and generation aspects of review systems, offering insights into the capabilities of AI in identifying manipulated content. The ultimate goal is to enhance online trustworthiness and promote fair user experiences.

II.EXISTING AND PROSPOSED SYSTEM

Online platforms such as Google Maps, Amazon, and Yelp rely heavily on user-generated reviews to inform customers and guide their purchasing or decision-making processes. To maintain the authenticity of these platforms, detecting and filtering out fake or misleading reviews is crucial. The current systems in place for fake review detection mostly rely on rule-based approaches or classical machine learning algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines (SVM). These models analyze basic features like review length, frequency of reviews, reviewer profile characteristics, and simple textual patterns. While these approaches offer some level of detection, they lack the sophistication to deeply understand the context and semantics of the reviews.

Moreover, most of these systems are not adaptive to evolving linguistic styles or capable of catching AI-generated or highly structured deceptive reviews. Manual moderation, another common method, is

inefficient, time-consuming, and non-scalable in real-world, large-volume scenarios. As a result, many fake reviews still bypass existing filters and continue to mislead consumers, ultimately damaging the trustworthiness and integrity of online platforms.

To address the limitations of the existing systems, we propose a comprehensive solution that integrates deep learning and real-time data extraction for effective fake review detection. Our system has two main components. The first component focuses on detecting fake Google Maps reviews using a prompt-based Deep Learning approach powered by the DeepSeek R1 model. This advanced language model is capable of understanding complex semantics, contextual relationships, and subtle patterns in textual data. Reviews are collected dynamically from Google Maps using the SerpAPI, which ensures real-world and up-to-date input. The collected reviews are then passed through DeepSeek R1, where prompt engineering techniques help analyze and classify them as genuine or fake based on linguistic cues, sentiment inconsistency, and contextual anomalies.

The second component of our system involves the development of a full-stack web application that allows users to generate automated reviews and test them using our detection system. This module serves two purposes: it provides an interface for users to interact with the system, and it also helps us study how the detection model responds to synthetically generated content. By integrating both detection and generation modules, our project presents a holistic view of how deep learning can be employed both defensively and analytically in the context of fake reviews.

Overall, the proposed system aims to provide a scalable, accurate, and intelligent solution to the problem of fake reviews. It utilizes state-of-the-art NLP techniques and modern web technologies to deliver a robust tool for ensuring content authenticity and promoting trust in digital ecosystems.

III.SYSTEM STUDY

A. Technical Feasibility

The technical feasibility assesses whether the existing technologies, tools, and resources are adequate for the successful implementation of the project. In our case, the system utilizes well-established and reliable technologies such as SerpAPI for data extraction, the DeepSeek R1 language model for fake review detection, and a full-stack web application built using React.js, Node.js, and MongoDB. All required development environments, frameworks, and libraries are accessible and compatible with our infrastructure. The deep learning model can be hosted on cloud platforms like Google Colab or integrated through APIs. The technical expertise required for this system, including API integration, prompt engineering, and full-stack development, is well within the capabilities of the project team, making the project technically feasible.

B. Economic Feasibility

Economic feasibility evaluates whether the proposed system is cost-effective and worth the investment. Since this project is developed in an academic environment, most of the tools used are open-source or free for academic use. Platforms like SerpAPI offer limited free access which is sufficient for our dataset needs during the development phase. The use of cloud-based resources such as Google Colab further reduces the need for expensive computing infrastructure. Additionally, the project does not require any hardware investments, and all software development is done using freely available tools and IDEs. Therefore, the overall cost of the project is minimal and economically feasible for a student-level implementation.

C. Operational Feasibility

Operational feasibility checks whether the system will function efficiently and be accepted by users. The proposed system is designed to be simple, intuitive, and user-friendly, targeting both technical and non-technical users who wish to analyze the authenticity of reviews. The interface of the web application is minimal and interactive, making it easy for users to input, generate, and analyze reviews. Moreover, the use of deep learning adds significant value by providing accurate and context-aware results, which boosts user confidence and usability. As a result, the system is expected to be operationally feasible and effective for its intended use.

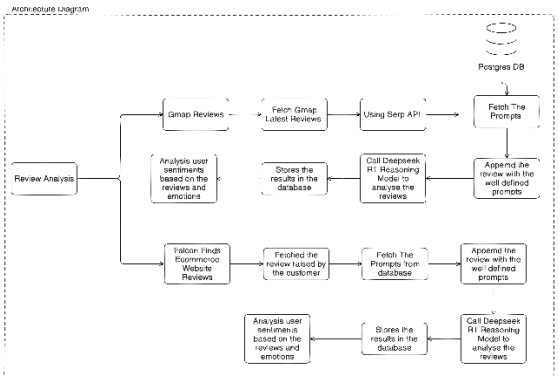
D. Legal and Ethical Feasibility

Legal feasibility involves ensuring that the project complies with data protection laws and ethical guidelines. The project collects only publicly available review data from Google Maps using SerpAPI, which is within the bounds of ethical scraping practices as long as it respects rate limits

and terms of service. No personal or sensitive user data is stored or used beyond what is already publicly visible. Additionally, the generated reviews used in the project are for educational and testing purposes only. Therefore, the project is legally and ethically sound for academic and research-based deployment.

IV. ARCHITECTURE DIAGRAM

The project "Fake Review Detection Using Deep Learning" has a full-stack architecture where users interact through a web application. The backend fetches real reviews from Google Maps using SerpAPI and also generates synthetic reviews. These reviews are processed by a deep learning model (DeepSeek R1) which analyzes and classifies them as real or fake. The results are stored in a database for analysis and review through an admin panel.



V. MODULES

1. Review Collection
2. Preprocessing
3. Prompt Generation
4. Model Prediction
5. Web App Review Generation
6. Output Display

A. Review Collection Module

The Review Collection Module is responsible for gathering user reviews from Google Maps using SerpAPI. This module replaces traditional web scraping techniques, offering a more reliable and scalable way to access live review data. By sending a search query (like a business or restaurant name), SerpAPI returns structured review information including the review content, ratings, reviewer identity, and timestamps. This real-world dataset forms the foundation for fake review detection, ensuring the system is exposed to diverse writing styles, languages, and tones found in genuine reviews.

B. Preprocessing Module

Once the reviews are collected, the Preprocessing Module prepares the raw text for input into the deep learning model. It cleans and normalizes the data by removing unnecessary characters, converting all text to lowercase, eliminating extremely short or irrelevant entries, and optionally filtering out non-English content. This step ensures consistency and reduces noise, making it easier for the language model to accurately interpret the content. The cleaner and more structured the input, the better the model performs in analyzing semantic and linguistic patterns.

C. Prompt Generation Module

This module transforms each cleaned review into a structured prompt that can be understood by a large language model (LLM) like DeepSeek R1. It embeds the review inside an instructional format, guiding the model to analyze and classify the text as either fake or genuine. For instance, it might present the task as: "Classify the following review based on authenticity." The prompt design plays a critical role in steering the model's understanding and ensuring consistent output without any additional training, relying purely on zero-shot inference.

D. Model Prediction Module (DeepSeek R1)

The Model Prediction Module forms the core intelligence layer of the system. It takes the generated prompt and sends it to the DeepSeek R1 model, which performs a deep semantic analysis of the review. The model evaluates factors such as tone, emotional exaggeration, repetition, and generic phrasing to determine whether a review is genuine or fake. The output includes a binary classification along with optional reasoning, demonstrating the model's ability to understand language context at a human-like level without requiring labelled training data.

E. Web App Review Generation Module

The final module supports the creation of synthetic reviews through a web interface. Users can either manually write reviews or automatically generate them using built-in templates or small language models. These reviews are then passed through the same preprocessing and prediction pipeline to test

how well the system can detect artificially created content. This module enables robust testing, data augmentation, and demo presentations, making the system more versatile and comprehensive.

F. Output Display Module

Once the model classifies a review, the Output Display Module presents the result in an interactive and user-friendly format. It displays the review text, the predicted label (FAKE or GENUINE), and any explanation provided by the model. Color-coded indicators help users quickly interpret the result, and additional features like saving the analysis or viewing statistics may be included. This module enhances the system's usability and serves as a visual feedback tool for evaluators and users.

VI. SCOPE OF FUTURE ENHANCEMENT

Platform-specific Fake Review Detection - Extend the model to be compatible with major review-based platforms like Amazon, Flipkart, Yelp, and Google Maps. Each platform has unique review structures and behaviors, so fine-tuning the model to suit each one can greatly improve detection accuracy.

Multilingual Review Analysis - Incorporate models like XLM-RoBERTa or mBERT to detect fake reviews written in different Indian and global languages, making the system more inclusive and applicable across regional and international markets.

User Behavior& Metadata Analysis - Go beyond text analysis by examining the reviewer's profile – such as account age, review frequency, IP address patterns, geolocation, and reviewing history – to identify suspicious users.

Multimodal Content Detection (Text + Media) - Extend the detection to analyze images or videos that accompany reviews. This could include checking for reused stock photos, inconsistent timestamps, or manipulated media.

Crowd sourced Data Labelling and Feedback - Allow users to report suspected fake reviews and contribute to data labeling, making the dataset more dynamic and accurate for future training iterations.

Time Series Analysis of Reviews - Track the timeline of reviews. A large number of similar reviews posted within a short time span may indicate spam or bot activity.

VII. CONCLUSION

The rise of fake reviews on online platforms poses a serious challenge to digital trust, user experience, and business credibility. This project presents a deep learning-based solution to effectively identify and filter out deceptive reviews using advanced text analysis techniques. By leveraging models such as transformer-based architectures, we demonstrated how artificial intelligence can be used to detect subtle linguistic patterns and user behaviours associated with fake content.

The system developed not only improves the reliability of online reviews but also provides a scalable and intelligent tool that can be adapted across different domains and platforms. With further enhancements such as real-time detection, multilingual support, and cross-platform integration, this project has the potential to become a comprehensive solution for ensuring authenticity in user-generated content. Thus, it serves as a significant step toward fostering transparency, fairness, and trust in the digital ecosystem.

VIII. REFERENCES

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

[2] 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.

[3] Zhang, Y., & Wang, Y. (2020). Detecting Fake Reviews via Deep Learning and Contextual Word Embeddings. Journal of Intelligent & Fuzzy Systems, 38(3), pp. 3515–3525.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

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

[6] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of

Review Spam Detection Using Machine Learning Techniques. Journal of Big Data, 2(1), pp. 1–24.

[7] Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to Identify Review Spam. IJCAI Proceedings - International Joint Conference on Artificial Intelligence, 2488–2493.

[8] Kumar, S., Spezzano, F., Subrahmanian, V. S., &Faloutsos, C. (2015). Edge Weight Prediction in Weighted Signed Networks. IEEE 31st International Conference on Data Engineering, 221–232.

[9] 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, pp. 985–994.

[10] Yin, H., Luo, X., Wang, Q., Lee, W. C., Wang, L., & Zhou, X. (2016). A Parameterized Co-Embedding Framework for Recommender Systems. ACM Transactions on Information Systems (TOIS), 35(4), pp. 1–31.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

[12] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

[13] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), pp. 1–135.