



Project Proposal

AI in Detecting inappropriate and misleading content on E-commerce Platforms

Prepared By

Name: Wiam Khalid Alharbi

ID No.:251000399

Name: Zakia Hani Alghamdi

ID No.:251000552

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Khaled Eskaf

Collage of Informatics, Midocean University, Comoros

Email:.....

1. Project Overview

E-commerce platforms are integral to modern-day shopping experiences, but the growth of these platforms has also led to a surge in inappropriate and misleading content, including fake product reviews, deceptive advertisements, and fraudulent product descriptions. This misrepresentation not only damages the trust consumers place in these platforms but also results in financial loss and even legal challenges for businesses. As such, it is crucial to develop robust systems that can automatically detect and filter misleading content. This project focuses on using machine learning techniques to identify such content, thereby making online shopping experiences safer and more reliable. The solution aims to improve trust, enhance user experience, and reduce the spread of misleading or fraudulent information.

2. Project Goal

The main objective of this project is to develop a model that detects inappropriate and misleading content across e-commerce platforms. The project goals can be outlined as follows:

- **Identify Misleading Product Descriptions:** Detect discrepancies between what is advertised and the actual product, helping consumers avoid purchasing items based on false claims.
- **Detect Fake Reviews:** Identify and filter reviews that are artificially inflated or fabricated to deceive customers into making a purchase.
- **Identify Deceptive Ads:** Develop algorithms to detect misleading advertisements that promise unrealistic outcomes or false benefits.
- **Improve Content Moderation:** Enhance the efficiency and effectiveness of content moderation processes on e-commerce websites by automating the identification of harmful or misleading content.

3. Approach

To achieve these goals, a multi-step approach will be used:

- **Data Collection:**

The project will require the collection of large datasets containing product descriptions, customer reviews, and advertisements from various e-commerce platforms.

- **Sources:** These datasets will be sourced from platforms like Amazon, eBay, Alibaba, or scraped from publicly accessible e-commerce sites (subject to legal considerations).
- **Data Types:** The data will primarily consist of textual content, including product names, descriptions, customer feedback, and promotional advertisements. Some platforms might also provide metadata (like ratings or sales data) that could further enhance the analysis.

- **Data Processing:**
 - Text Preprocessing: Raw data will be cleaned by removing special characters, irrelevant words, and normalizing the text (lowercasing, removing stopwords).
 - Tokenization and Lemmatization: Tokenization will break down the text into individual words or phrases, and lemmatization will be used to reduce words to their base forms (e.g., “running” becomes “run”).
 - Feature Extraction: Natural Language Processing (NLP) techniques will be employed to extract useful features from the text. Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings (e.g., Word2Vec, GloVe) will be used to convert textual data into numerical representations that can be processed by machine learning models.
- **Model/Method Selection:**
 - Traditional Models: Initially, traditional machine learning algorithms such as Logistic Regression, Naive Bayes, and SVM (Support Vector Machine) will be applied to identify patterns of misleading content.
 - Deep Learning Models: For better performance, more sophisticated models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), or Transformer-based models like BERT and GPT could be employed, especially for handling longer texts and extracting contextual relationships.
 - Supervised Learning: Since the task involves categorizing content as misleading or not, supervised learning techniques will be used where labeled data (e.g., pre-tagged reviews as fake or real) will help train the models.
- **Testing and Evaluation:**

The model’s performance will be evaluated using standard metrics like:

 - Accuracy: The overall rate at which the model correctly classifies content.
 - Precision: The percentage of relevant content detected as misleading from the content flagged by the model.
 - Recall: The percentage of all misleading content correctly identified by the model.
 - F1-Score: The harmonic mean of precision and recall, providing a balanced evaluation metric.
 - Cross-validation will be used to ensure the model performs well across different subsets of data and generalizes effectively.

4. Project Steps and Timeline

Phase 1: Data Collection and Preparation (1-2 months)

- Identify and gather datasets of product descriptions, reviews, and advertisements.

- Clean and preprocess the data, including tokenization, lemmatization, and feature extraction.

Phase 2: Model Building and Training (2-3 months)

- Develop and test traditional machine learning models (e.g., SVM, Logistic Regression) for initial detection of misleading content.
- Experiment with deep learning approaches (e.g., RNNs, LSTMs, BERT).
- Fine-tune the models, adjusting parameters to optimize accuracy.

Phase 3: Testing and Evaluation (1-2 months)

- Evaluate the model using accuracy, precision, recall, and F1-score on validation and test datasets.
- Analyze the results, identifying strengths and weaknesses of the model.
- Implement improvements based on evaluation metrics.

Phase 4: Final Adjustments and Reporting (1 month)

- Make necessary adjustments to improve the model's accuracy and generalization.
- Compile the results, methodologies, and future work into a final report.
- Prepare a presentation summarizing key findings and recommendations for e-commerce platform integration.

5. Expected Results

The expected outcomes of the project include:

- Improved Accuracy in Detecting Misleading Content: The model should be able to identify fake reviews, deceptive ads, and misleading product descriptions with high precision.
- Development of a Practical Tool: A model that can be directly applied to e-commerce platforms for automated content moderation. This could be integrated into existing content moderation systems or developed as a standalone application for businesses.
- Contributions to NLP and E-Commerce Research: The project will provide new insights into how machine learning and NLP can be leveraged for content moderation and could lead to further research into improving detection methods.

6. Resources Needed

Software and Tools:

- Python: The primary programming language for the project.
- Libraries: Scikit-learn, TensorFlow, PyTorch, Keras for machine learning and deep learning model implementation; NLTK or SpaCy for NLP tasks.
- Cloud Resources: Computing resources (e.g., AWS, Google Cloud) for model training and storage of large datasets.

Datasets:

- Publicly available e-commerce datasets or web scraping tools for collecting real-time data from platforms like Amazon, eBay, and others.
- Pre-labeled datasets of reviews and ads, if available, for training purposes.

Support:

- Access to academic resources and journals on machine learning and NLP to guide model selection and development.
- Guidance from faculty or industry professionals specializing in e-commerce, content moderation, and machine learning.

Project Summary:

This project focuses on detecting inappropriate and misleading content on e-commerce platforms like Amazon and eBay. As the volume of transactions and interactions on these platforms grows, deceptive practices such as false advertisements, manipulated product images, and misleading descriptions have become more common. These practices negatively impact consumer trust and lead to poor purchasing decisions.

To address this issue, advanced techniques are used to analyze both text and images. Reviews, comments, and product descriptions are analyzed using natural language processing (NLP) to detect suspicious patterns or misleading statements. Additionally, computer vision techniques are applied to analyze product images and ensure they accurately represent the products, preventing any alterations or misrepresentations.

As content manipulation methods continue to evolve, the project aims to continuously update and improve these models to ensure their effectiveness. The goal is to ensure accurate content moderation, protect user privacy, and avoid removing legitimate content. Ultimately, the project seeks to enhance the user experience on e-commerce platforms by improving the quality of displayed information and increasing consumer trust.

The project also aims to adapt to emerging trends in content manipulation, making it a dynamic and evolving solution to maintaining content integrity over time.

	Project Name	Authors	Summary	Methodology
1	AI-based Content Creation and Product Recommendation Applications in E-commerce: An Ethical Overview	Aditi M Jain, Ayush Jain	This study discusses the rapid integration of artificial intelligence (AI) in e-commerce for content creation and product recommendations. It highlights the benefits of AI in personalization and efficiency but also explores critical ethical challenges such as data privacy, algorithmic bias, and consumer autonomy. The paper proposes frameworks to ensure fairness, transparency, and the removal of biases in AI models used for content creation.	The paper analyzes AI-driven systems in e-commerce platforms like Amazon and Shopify. It discusses the biases embedded in AI models and proposes practices such as conducting regular algorithm audits, diversifying training data, and incorporating fairness metrics into the AI models.

2	E - C o m m e r c e F r a u d D e t e c t i o n B a s e d o n M a c h i n e L e a r n i n g T e c h n i q u e s : S y s t e m a t i c L i t e r a t u r e R e v i e w	Abed Mutemi, Fernando Bacao	This study addresses the challenges of e-commerce fraud detection, particularly focusing on the use of machine learning (ML) and data mining techniques. With the rapid growth of e-commerce, especially accelerated by the COVID-19 pandemic, the frequency of digital fraud incidents has increased. The study conducts a systematic literature review using the PRISMA methodology, exploring the effectiveness of machine learning techniques for fraud detection in digital marketplaces and e-commerce platforms. It identifies key trends, recent techniques, and research gaps, and highlights the increased use of artificial neural networks (ANNs) in fraud detection.	The researchers conducted a systematic literature review (SLR) using the PRISMA methodology. They analyzed 101 publications, focusing on fraud detection techniques, identifying key machine learning algorithms applied in the domain, and providing insights into emerging trends and gaps in research. This methodology involved defining search strategies, conducting data acquisition, cleaning, and content analysis, followed by testing and revision of publication results
---	---	--------------------------------	--	---

3	M i t i g a t i n g R e v i e w a n d R a t i n g F r a u d i n E - C o m m e r c e P l a t f o r m s : A B l o c k c h a i n - B a s e d R e p u t a t i o n S y s t e m w i t h A I - D r i v e n R e v i e w V a l i d a t i o n	F a t e m e h S o l e i m a n i , F a k h r o d d i n N o o r b e h b a h a n i , M o j t a b a M a h d a v i	This paper presents a novel approach to mitigating review and rating fraud in e-commerce platforms by integrating blockchain technology with artificial intelligence (AI). The goal is to establish a decentralized, immutable system for handling customer reviews and ratings. The proposed system uses blockchain to ensure transparency, security, and trustworthiness in reviews and ratings. It incorporates smart contracts to automate processes such as review validation, reward distribution, and product listing. Additionally, AI-driven review validation is used to distinguish between genuine and fake reviews, enhancing trust among users.	The research proposes a blockchain-based reputation system that employs AI to validate customer reviews and prevent fraudulent behavior in e-commerce platforms. The system uses smart contracts to automate the entire review process, ensuring that reviews can only be submitted by verified buyers. The AI model, stored in a decentralized system (IPFS), validates reviews based on predefined criteria, and rewards users for providing honest and genuine feedback. The methodology includes setting up a decentralized review system, validation via AI, and the distribution of rewards through smart contracts.
---	--	--	---	--

4		Mosleh Hmoud Al-Adhaileh, F a w a z W a s e l a l l a h Alsaade	This paper tackles the challenge of detecting fake opinions, particularly fraudulent product reviews, which are widespread on e-commerce and social media platforms. Fake reviews are often written by fraudsters who aim to mislead customers and damage the reputation of businesses. The study proposes a method to distinguish between fake and truthful product reviews using linguistic features and deep learning algorithms. Two models—BiLSTM (Bidirectional Long Short-Term Memory) and CNN (Convolutional Neural Network)—were tested and compared. The BiLSTM model showed higher accuracy in detecting fake reviews, providing an effective approach to prevent deceptive content in online marketplaces.	The methodology consists of seven phases: 1. Product Review Dataset: The dataset consists of 9456 fake reviews from various e-commerce platforms. 2. Analyzing Features: Linguistic features like word count, sentiment (positive/negative), and authenticity were extracted using tools like the Linguistic Inquiry and Word Count (LIWC). 3. Preprocessing: This phase involved text cleaning, stopword removal, tokenization, and word stemming to prepare the dataset for model training. 4. Word Embedding: Word2Vec was used to map words into vector representations based on their semantic meaning. 5. Classification: The reviews were classified using BiLSTM and CNN deep learning models. Both models were trained and evaluated using performance
---	--	--	---	--

5	Fake Reviews Detection using Supervised Machine Learning	A h m e d M . Elmogy, Usman T a r i q , A t e f Ibrahim, Ammar Mohammed	This paper addresses the problem of detecting fake reviews in e-commerce systems, which play a crucial role in decision-making for customers and contribute to the reputation of businesses. Fake reviews are often written to artificially boost product ratings. The study uses supervised machine learning techniques to detect fake reviews by applying multiple classifiers, including KNN, Naive Bayes, SVM, Logistic Regression, and Random Forest, on a real Yelp dataset. The results show that incorporating user behavior features, such as review timing and writing style, improves detection accuracy, with KNN (K=7) achieving the best f-score of 82.40%.	1. Dataset: Real Yelp dataset consisting of restaurant reviews. 2. Feature Extraction: Textual features (e.g., sentiment, Cosine similarity, TF-IDF) were extracted from the reviews, alongside behavioral features like the reviewer's writing style, timestamp, and frequency of reviews. 3. Classifiers: Five classifiers were evaluated—KNN, Naive Bayes, SVM, Logistic Regression, and Random Forest. 4. Language Models: Two language models, bi-gram and tri-gram, were used to improve feature extraction. 5. Performance Metrics: Accuracy, precision, recall, and f1-score were used to evaluate classifier performance. The results showed an improvement in performance when reviewer behavioral features were included.
---	--	---	--	---

6		J. M. Thilini Jayasinghe, Sachith Dassanayaka	This study addresses the issue of fake reviews on Amazon, one of the largest e-commerce platforms in the USA. Fake reviews undermine the integrity of consumer feedback systems. The paper proposes a new methodology to classify genuine and fraudulent reviews using a combination of convolutional neural networks (CNNs), word embedding, and emotion-mining techniques. The model achieves over 96% accuracy in detecting fake reviews, providing valuable insights into the linguistic and emotional markers that distinguish fake reviews from authentic ones. This research aims to enhance the authenticity of online reviews and rebuild consumer trust in e-commerce platforms.	- Dataset: Utilized the Amazon 2018 dataset, focusing on reviews from categories like Cell Phones, Digital Music, and Electronics. - Approach: A two-step process was used to detect fake reviews: - Cosine similarity between verified and non-verified purchase reviews was used to identify fake reviews. - Deep Neural Networks (CNNs) were trained using features like review length, sentiment scores, and product ratings to classify reviews. - Model Architecture: The CNN model was designed with three convolutional layers, using a kernel size of 5 to capture essential text patterns. Max Pooling and GlobalMaxPooling1D layers were applied to reduce data complexity. Dropout was used to prevent overfitting. - Loss Function: A hybrid adaptive cross-entropy (HACE) loss function was
---	--	---	---	--

Detecting Deception: Employing Deep Neural Networks for Fraudulent

7. References

Books:

- “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow” by Aurélien Géron
- “Deep Learning with Python” by François Chollet

Research Papers and Journals:

- “Fake Review Detection: A Data Mining Perspective” – Journal of Data Science.
- “A Survey on Content Moderation in Social Media Platforms” – IEEE Access.
- “Detecting Misleading Information on the Web Using Natural Language Processing” – Journal of Computer Science.