



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.

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.