



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

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Khaled Eskaf

College of Informatics, Midocean University, Comoros

[2025]

Acknowledgements

We give thanks to God foremost, for all the wonderful things we accomplished, and no knowledge is gained without his guidance and will. we are also grateful and appreciative for our supervisor Dr. Khaled Eskaf whose guidance, expertise, and unwavering encouragement were pivotal in shaping the direction and quality of this thesis.

His insightful observations, patience, and commitment to academic excellence were invaluable at every stage of this thesis. I am deeply grateful for the time and effort he dedicated to supporting my work and helping me develop.

I also extend my sincere thanks to the faculty and staff of the department, whose knowledge, resources, and constructive discussions fostered a rich academic environment. Their contributions significantly enhanced the development of this study.

I am also deeply grateful to my colleagues and friends for their support, motivation, and collaborative spirit. Their willingness to share ideas and provide feedback created a positive and inspiring atmosphere throughout this journey.

Finally, I thank my family, whose constant encouragement, understanding, and moral support were the foundation of my perseverance. Their belief in me was a constant source of strength and inspiration.

Abstract

The rapid growth of e-commerce platforms has led to a significant increase in misleading and inappropriate content, such as fake and inaccurate reviews, deceptive advertisements, and misleading product descriptions. This content negatively impacts customer trust, influences purchasing decisions, and reduces the credibility of websites.

This project aims to develop an intelligent tool capable of verifying inaccurate and fake reviews in Arabic e-commerce using ML techniques.

The proposed solution includes a multi-stage processing sequence encompassing data preprocessing, feature extraction, and model training for tool evaluation. It purifies Arabic texts using NLP techniques such as Tokenization, Stopword Removal, and Stemming.

Two methods are used for feature extraction and model training: TF-IDF for capturing statistical word importance and Word2Vec for learning semantic representations. Several machine learning algorithms are trained and taught, including LR, Naive Bayes, and Linear SVM, and evaluated using accuracy, precision, recall, and F1-score. The best-performing tool models are then integrated into an interactive tool built using Streamlit. This tool allows users to analyze individual reviews or a portfolio of reviews and compare model performance.

The results indicate that machine learning significantly improves the detection of misleading reviews, thereby boosting customer trust in e-commerce. This project provides a foundation for integrating artificial intelligence and e-commerce.

List of Publications

Table of Contents

List of Abbreviations	vi
List of Figures	vii
Chapter 1: Introduction.....	1
1.1 Problem Definition.....	1
1.2 Thesis Objectives.....	2
1.3 Thesis Contribution.....	3
1.4 Thesis Organization.....	4
Chapter 2: Overview.....	4
2.1 Project overview.....	4
Chapter 3: Related Work.....	5
3.1 Introduction.....	5
3.2 History.....	5
Chapter 4: Proposed Approach.....	6
4.1 Data Collection.....	6
4.2 Data Preprocessing.....	7
4.3 Model/Method Selection.....	8
4.4 Testing Evaluation.....	8
4.5 Expected Results.....	10
Chapter 5: Project Steps and Timeline	10
5.1 Phase 1: Data Collection and Preparation.....	10
5.2 Phase 2: Model Building and Training.....	11
5.3 Phase 3: Testing and Evaluation.....	11
5.4 Phase 4: Final Adjustments and Reporting.....	11
5.5 Resources Needed.....	12
Chapter 6: Conclusion and Future Work	12
6.1 Conclusion	12

6.2 Future Work	13
References	15

List of Abbreviations

- **NLP:** Natural Language Processing
- **ML:** Machine Learning
- **SVM:** Support Vector Machine
- **DL:** Deep learning
- **LR:** Logistic Regression

List of Figures

Figure1: Thesis Organization and Development Process.....	4
Figure2: Proposed Framework.....	8
Figure3: Machine Model Performance.....	9
Figure4: Streamlit UI Screenshot.....	11

Chapter 1

Introduction

With the increasing reliance on e-commerce platforms, user reviews have become a crucial factor influencing purchasing decisions. These reviews aim to provide honest opinions and help customers assess product quality. However, the rise of fake and misleading reviews has posed a significant challenge for both customers and online merchants. Misleading content not only distorts the perceived value of products but also diminishes the credibility of e-commerce in general.

Detecting this content is extremely difficult, especially in Arabic, due to its linguistic complexity and the lack of specialized tools designed for analyzing Arabic texts. Traditional manual moderation is time-consuming, inefficient, and unable to keep pace with the growing volume of online reviews. Therefore, there is a need for an intelligent, automated solution capable of analyzing Arabic texts and identifying misleading or inaccurate content.

This project aims to address this challenge by developing a ML based tool capable of processing Arabic reviews, extracting their useful characteristics, and classifying them as either genuine or misleading. The tool combines modern NLP techniques and multiple classification models to explore the most effective approaches for detecting fake Arabic reviews. Through this work, the project aims to support e-commerce platforms in maintaining user trust.

1.1 Problem Definition

With the increasing volume of user-generated reviews on e-commerce platforms, identifying misleading or fake content has become a major challenge. Although reviews are meant to assist consumers in making well-informed purchasing decisions, the presence of false, misleading, or poorly written reviews seriously skews product assessments. The legitimacy of online marketplaces is damaged, consumer trust is immediately impacted, and purchasing behavior is influenced.

Thus, the main issue this project attempts to solve is:

How can we create an automatic, precise, and effective approach to identify phony or deceptive Arabic reviews on e-commerce sites?

There are various sub-challenges in this problem:

- Processing and cleaning unstructured Arabic material that contains casual writing, slang, or noise.
- Identifying significant linguistic characteristics that differentiate authentic reviews from fraudulent ones.
- Training machine learning algorithms to correctly identify the authenticity of reviews.
- Supplying a useful tool that enables real-time review analysis for both bulk datasets and individual reviews.

By creating and deploying a machine learning-based detection system that enhances transparency and facilitates content moderation on e-commerce platforms, this project seeks to address these issues.

1.2 Thesis Objectives

The main objective of this project is to develop a model that detects inappropriate and misleading content across e-commerce platforms. The project goals include the following objectives:

- **Detecting fake reviews:** This aspect involves identifying and classifying misleading or fake reviews, which helps prevent consumers from relying on inaccurate information when making purchasing decisions.
- **Identifying Misleading Product Descriptions:** This aspect aims to uncover discrepancies between the advertised product descriptions and their actual features, thus helping to protect consumers from making purchasing decisions based on false or exaggerated claims.
- **Improved Content Management:** This aspect aims to increase the efficiency and effectiveness of content management processes on e-commerce websites by automating the detection of harmful content, thereby improving accuracy.

1.3 Thesis Contribution

This work makes important contributions in the field of misleading content on e-commerce platforms and NLP in Arabic.

- **Creation of an Arabic Fake-Review Identification tool:** A Comprehensive ML-based framework that can identify fraudulent or deceptive Arabic reviews. This contains a comprehensive preprocessing, feature extraction, model training, and assessment pipeline, specifically designed to handle the intricacies of the Arabic language.
- **Comparative Evaluation of Various ML Models:** A complete comparison is made between different classification algorithms, such as LR, Naive Bayes, Linear SVM, and Word2Vec-based models, using several assessment measures. This sheds light on the best models for Arabic review authenticity analysis.
- **Creation of a Cleaned and Structured Arabic Review Dataset:** The project processes and prepares an Arabic dataset for fake-review detection, including detailed Cleaning, Normalization, and Labeling. This dataset contributes to the limited availability of structured Arabic resources in this domain.
- **Implementation of an Interactive Tool for Real-World Use:** A practical Streamlit-based application is developed that enables users to analyze individual reviews or batch datasets and visualize model performance. This tool demonstrates the real-world applicability of the research and supports potential integration into e-commerce moderation systems.
- **Contribution to Arabic NLP Research and E-Commerce Transparency:** The thesis highlights the challenges of Arabic text analysis in misleading-content detection and proposes effective techniques to address them. The results provide a foundation for future studies and help advance automated solutions that enhance trust and credibility in online marketplaces.

1.4 Thesis Organization

The remainder of the thesis is structured as follows:

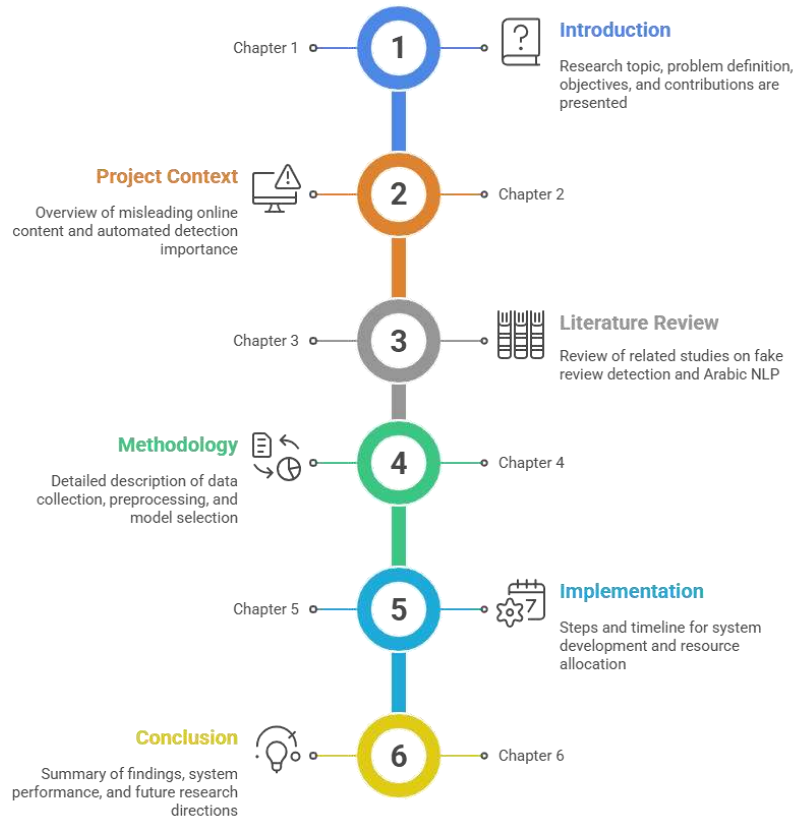


Figure 1 Thesis Organization and Development Process.

Chapter 2 Overview

2.1 Project Overview

E-commerce platforms are integral to modern-day shopping experiences, but their expanding user base has created an increase in deceptive content, which includes fake product reviews, deceptive ads, and misleading product information.

This misrepresentation not only damages the trust consumers place in these platforms but also results in financial loss and even legal challenges for businesses. The development of strong automated systems that detect and remove deceptive

content has become essential because of this situation. This project focuses on using machine learning techniques to detect misleading content to protect the reliability and safety of online shopping. The proposed solution builds consumer confidence and minimizes the distribution of deceptive or misleading content.

Chapter 3 Related Work

3.1 Introduction

With the rapid expansion of e-commerce platforms, reviews and ratings have become one of the most influential factors in consumers' purchasing decisions. However, this proliferation has been accompanied by a significant increase in misleading content, such as fake reviews, deceptive advertisements, and inaccurate product descriptions. This type of content not only misleads users but also erodes trust in the platform and negatively impacts the credibility of e-commerce.

Many early studies relied on traditional language processing and text analysis methods, but most focused on English, while Arabic content remained less researched despite its widespread use and diverse dialects and lexical styles. Furthermore, the scarcity of tagged data and the difficulty of processing Arabic texts have made the need for more advanced and effective solutions urgent.

This background underscores the importance of developing an intelligent tool capable of detecting misleading content in Arabic reviews using ML and NLP techniques, which is the aim of this research.

3.2 History

Early research on fake review detection primarily relied on classical machine learning models such as Logistic Regression, Naive Bayes, and SVM. These models typically used TF-IDF or Bag-of-Words to represent textual features, enabling basic classification but offering limited semantic understanding.

With advancements in NLP, studies introduced word embeddings such as Word2Vec, GloVe, and fastText, which improved the ability to capture

contextual meaning. However, many of these studies evaluated only a single model and did not provide comprehensive comparisons between algorithms. Additionally, most prior work focused on English datasets, leaving a gap in Arabic review analysis.

Research targeting Arabic misleading content remains limited. Existing studies suffer from challenges such as dialect variation, lack of annotated data, and complex morphological structures. Some attempted deep learning approaches, but results varied due to insufficient data and high computational demands.

Compared to these studies, the current project provides several contributions:

- It evaluates multiple machine learning models rather than relying on a single approach.
- It combines TF-IDF and Word2Vec for improved feature representation.
- It focuses specifically on Arabic misleading reviews, addressing a research gap.
- It delivers a functional, interactive Streamlit tool, offering real-world usability that many previous works lacked.

Chapter 4 Proposed Approach

To achieve the objectives of this research, a multi-stage methodology was adopted to build an effective system for detecting misleading Arabic reviews. This methodology follows a structured sequence. Each step was carefully designed to ensure the model's ability to understand Arabic text and accurately extract patterns indicative of genuine and misleading reviews.

4.1 Data Collection

This project requires collecting a wide range of data, including customer reviews and product descriptions, from various e-commerce platforms. This data will be sourced from publicly available websites such as Kaggle and Amazon, while taking into account each platform's legal aspects and usage policies. The data will consist of textual content, including user comments, product names, and their associated

descriptions, to provide a comprehensive database for subsequent analysis and modeling.

4.2 Data Preprocessing

To ensure the efficient performance of machine learning models, raw Arabic revisions undergo several preprocessing steps aimed at cleaning and standardizing the text before being fed into the training phase. This process includes text noise removal, character normalization, removal of unnecessary symbols, and other Arabic language processing techniques. These steps have been consistently applied to both training and test data to ensure the reliability of the results and the accuracy of the model.

- **Text Cleaning:** This includes removing URLs, non-Arabic characters and shortening the text to standardize word formatting
- **Tokenization:** The cleaned text is divided into symbols (words) based on the number of spaces.
- **Removing Stopword:** Arabic Stopwords (e.g., على، من) are removed using Natural Language Toolkit (NLTK) to retain only meaningful tokens.
- **Stemming:** The ISRIStemmer from Natural Language Toolkit (NLTK) is applied to reduce each word to its root. This also improves generalization and reduces vocabulary size.
- **Feature Extraction:** use two different vectorization techniques to extract features. Apply TF-IDF and word embeddings to turn text into numerical data for machine learning.

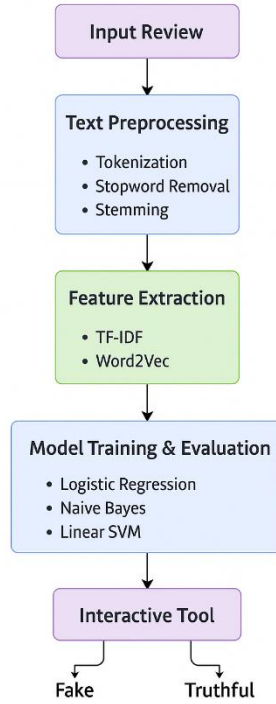


Figure 2 Proposed Framework.

4.3 Model/Method Selection

Three machine learning models and one embedding-based model were chosen to compare their performance on content classification tasks and test their ability to effectively handle textual data. Each model uses a different method, allowing for a focused evaluation of various text representation techniques.

- TF-IDF + LR model is a simple and efficient linear classifier often used for text classification. It works well with sparse text features and provides a solid baseline for spotting misleading reviews.
- TF-IDF + Linear SVM model is a strong margin-based classifier for high-dimensional text data and achieves high accuracy in binary classification tasks.
- TF-IDF + Naive Bayesian Algorithm is a fast probabilistic approach that works especially well for short and simple texts. This method is lightweight and efficient for classifying reviews.
- Word2Vec + LR uses dense word embeddings to capture the meaning and context of words in the text. This approach adds a more detailed linguistic perspective compared to TF-IDF models.

Comparing these models enables a comprehensive evaluation of their performance and helps determine the best method for detecting misleading content on E-commerce platforms.

4.4 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.

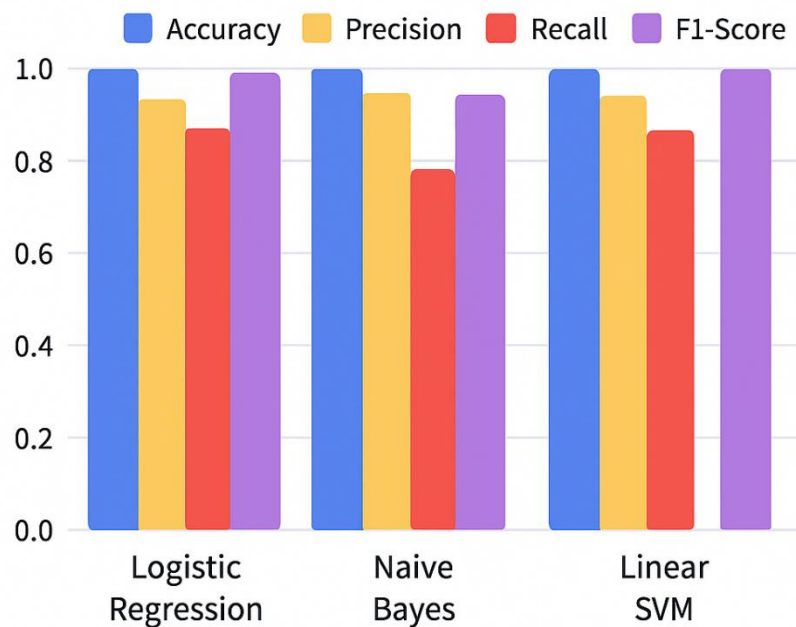


Figure 3 Model Performance Chart.

4.5 Expected Results

The expected outcomes of the project include:

- **Improved Accuracy in Detecting Misleading Content:** The project's machine learning models aim to identify misleading reviews, product descriptions, and other inappropriate content on e-commerce platforms, thereby improving user trust and reducing exposure to fraudulent information.
- **New Insights into Arabic NLP for Content Management:** Applying multiple preprocessing techniques and comparing traditional machine learning models with embedding-based methods, the project aims to provide insights into analyzing Arabic texts, benefiting future research in content management and fake review detection.
- **A Practical Tool for Application:** The project will deliver an efficient tool for analyzing single reviews or entire files, which can be integrated into online store content monitoring systems.

Chapter 5

Project Steps and Timeline

This chapter provides an overview of the project implementation phases, outlining the workflow followed from data collection to final report preparation. The work was divided into four main phases: data preparation, model building, performance evaluation, and final refinements.

5.1 Phase 1: Data Collection and Preparation.

- Identify and gather datasets of product descriptions and reviews.
- Clean and preprocess the data, using techniques such as URL removal, Arabic Normalization, Stopword removal, Tokenization, and Stemming.
- Convert text into numerical using TF-IDF and Word2Vec.
- Split the dataset into training and testing sets.

5.2 Phase 2: Model Building and Training.

- Training a machine learning model, including LR, Naive Bayes, and Linear SVM, and a Word2Vec-based model.
- The trained models were integrated into the Streamlit interface to facilitate user interaction.

5.3 Phase 3: Testing and Evaluation.

- Evaluate model performance using standard metrics such as accuracy, recall, positive accuracy, and F1 score.
- Analyze errors and identify strengths and weaknesses in each model.
- Test models on new data to ensure generalizability and minimize overlearning.
- Update the tool's interface to display evaluation results interactively and clearly.

5.4 Phase 4: Final Adjustments and Reporting.

- Improve the model's performance and the tool's interface based on the evaluation results.
- Document the steps, methods, experiments, results, and challenges in the final project report.
- Prepare presentations and visual materials to showcase the project.

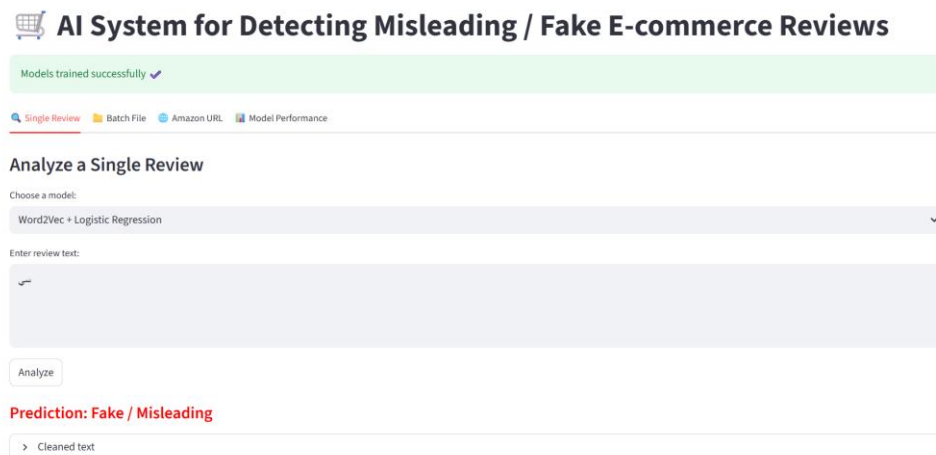


Figure 4 Streamlit UI Screenshot.

5.5 Resources Needed

- **Software and Tools:**
 - Language:
 - The primary programming language for the project is Python
 - ML Libraries:
 - Scikit-learn for traditional machine learning (ML) models and TensorFlow for potential DL extensions.
 - NLP Libraries:
 - NLTK for Arabic text preprocessing and Gensim for Word2Vec embeddings.
 - User Interface framework
 - Streamlit to building an interactive tool to display results.
- **Datasets:**
 - The dataset from Kaggle contains Arabic reviews categorized as Fake or Truthful.
 - Additional datasets from e-commerce platforms may be integrated into future projects.
- **Computing Resources**
 - A computer with moderate processing power for training models.
 - Google Colab to support faster model training if needed.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis presents an intelligent system for detecting misleading Arabic reviews on e-commerce platforms, relying on natural language processing and machine learning techniques. The methodology included integrated steps, starting with text cleaning and feature extraction using TF-IDF and Word2Vec, followed by training several models such as Logistic Regression, Naive Bayes, and Linear SVM, and

concluding with performance evaluation using rigorous scientific metrics. The results showed that the Naive Bayes model achieved the best overall performance, reflecting its effectiveness in analyzing short, direct Arabic texts.

The research also resulted in the development of an application using Streamlit that allows for the analysis of individual reviews or entire datasets, making the system practically applicable in real-world environments. This project contributes to filling a clear research gap in the field of detecting misleading Arabic reviews and provides a framework upon which to build more advanced tools in the future.

6.2 Future Work

Despite the effectiveness of the developed tool for detecting misleading Arabic reviews, several aspects can be improved in the future to enhance the accuracy and practicality of the tool. Key areas for future development include:

1. Integrating the system with e-commerce platforms via APIs

Due to technical limitations in collecting reviews directly from sites like Amazon and AliExpress, as a result of their anti-scraping and security systems, the current model does not include a live review fetching process.

- In future developments, the system can be integrated with e-commerce platforms through:
- Using official APIs or external APIs (such as RapidAPI) to fetch reviews directly.
- Analyzing the content of dynamic pages using advanced tools such as Selenium or Playwright.
- Integrating the tool with a real platform that automatically reads and categorizes reviews as soon as they are published, without human intervention.
- Adding this step will make the system more practical and prepare it to be part of a company's content moderation system.

2. Use DL models (Transformers) such as BERT or AraBERT to improve understanding of linguistic context, especially in complex or long texts.

These models may offer higher accuracy but require more computing resources and more data.

3. Expand the size and quality of the data The model's performance can be enhanced by expanding the dataset to include: reviews from multiple platforms, different Arabic dialects, and longer, more detailed texts.

4. Integrate additional non-textual features such as star count, user behavior, review posting frequency, and review date (the time the review was written). These features can help detect fake reviews more accurately.

5. Develop a more integrated interface. In the future, the system could be integrated with e-commerce platforms via API to enable automatic and real-time review monitoring.

6. Expand the system to be multilingual, supporting both Arabic and English reviews, especially for international stores.

References

- A Survey on content moderation in social media platforms. (n.d.). *IEEE Access.*
- Al-Adhaileh, M. H., & Alsaade, F. W. (n.d.). *Detecting and analysing fake opinions using artificial intelligence algorithms.*
- Chollet, F. (n.d.). *Deep learning with Python.*
- Detecting misleading information on the web using natural language processing. (n.d.). *Journal of Computer Science.*
- Elmogy, A. M., Tariq, U, Ibrahim, A., & Mohammed, A. (n.d.). *Fake reviews detection using supervised machine learning.*
- Fake Review Detection: A data mining perspective. (n.d.). *Journal of Data Science.*
- Géron, A. (n.d.). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.*
- Jain, A. M., & Jain, A. (n.d.). *AI-based content creation and product recommendation applications in e-commerce: An ethical overview.*
- Jayasinghe, J. M. T., & Dassanayaka, S. (n.d.). *Detecting deception: Employing deep neural networks for fraudulent review detection on Amazon.*
- Mutemi, A., & Bacao, F. (n.d.). *E-commerce fraud detection based on machine learning techniques: Systematic literature review.*
- Soleimani, F., Noorbehbahani, F., & Mahdavi, M. (n.d.). *Mitigating review and rating fraud in e-commerce platforms: A blockchain-based reputation system with AI-driven review validation.*