

زكيه الغامدي

Final Project 1.3 wiam %2C Zakia.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:338010252

Submission Date

Dec 17, 2025, 7:45 PM GMT+2

Download Date

Dec 17, 2025, 9:45 PM GMT+2

File Name

Final Project 1.3 wiam , Zakia.docx

File Size

1.5 MB

23 Pages

3,940 Words

23,827 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





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

To begin with, no achievements can be undertaken without God's will and direction, so we thank Him first and foremost.

We owe a great debt of gratitude to Dr. Khaled Eskaf, our supervisor. He played a crucial role with his direction, professionalism, unrelenting support and motivation to impact the outcome of this thesis.

It was his comments during the thesis on the several occasions, his tireless efforts and his devotion toward scholarship that I thank the most.

A special thank you goes to the department faculty and support staff, whose resourcefulness, academic materials, and positive feedback participated paving the way for scholarship to thrive. Their efforts notably contributed to the completion of this study.

Lastly this research project was a positive and motivating experience for which I thank my friends and colleagues for the collaboration and team spirit they displayed. Their idea and comment exchanges were the motivation that kept the spirit of the project alive.

My deepest gratitude goes to my family for their unchanging support, encouragement, and faith in me, which motivated me to keep going. I drew strength and inspiration from their support.

Abstract

The gradual expansion of e-commerce stores has caused the emergence and proliferation of false and inappropriate e-commerce-related information, including inauthentic and inaccurate customer reviews, false ad notifications, and inappropriate product offerings. This kind of information harms customer trust, distorts purchasing behavior, and decreases site authenticity.

The purpose of the study is to design and develop a sophisticated system that identifies and classifies false reviews in the Arabic language for e-commerce applications, utilizing machine learning approaches.

Data preprocessing, feature engineering, and model training sequentially evaluate tools in the proposed approach. With the aid of the aforementioned NLP techniques, the Arabic text refinement stage of the process is accomplished by Tokenization, Stopword Removal, and Stemming.

Feature extraction and model training are accomplished by TF-IDF, which records a word's frequency, and Word2Vec, used for concept mappings. LR, Naive Bayes, and Linear SVM are the obtained supervised ML algorithms, and are assessed using a set of metrics; accuracy, precision, recall, and F1-score. The trained model tools are embedded in a streaming application using Streamlit. Users of the application can upload and compare a set of documents, and the individual review analysis will aid model performance analysis.

More customer trust on online shopping due to review analysis automation was the main impact of the presented ML model. This also acted as a stepping stone for ML and AI in 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

The increasing reliance on e-commerce platforms has made user reviews an important factor used to influence other user potential customer's buying decisions. The aim of these reviews are to help potential customers make an informed decision on a product. Unfortunately, the emergence of fake reviews has created a new problem for potential customers and online merchants. Fraudulent reviews create an imbalanced alter reality on the perceived value of a product and e-commerce via online reviews.

The Saudi Vision 2030, a blueprint of the Kingdom's e-commerce, relies on a large number of e-commerce merchants and online customers. The Saudi Data and Artificial Intelligence Authority (SDAIA) has spearheaded several businesses to use artificial intelligence to bridge the gap between online merchants and customers, most especially within the e-commerce platforms.

The aim of this project is to develop a Machine Learning tool to process and analyze the attributes of reviews. The project has its aim to plan and support e-commerce platforms in the user trust retention in order for them to build a highly sophisticated response to the challenges.

1.1 Problem Definition

As there is more and more user-generated review content on e-commerce site, it is becoming more and more difficult to determine user-generated reviews that are fake or misleading. While reviews should help consumers make better purchases, fake, misleading and poorly written reviews seriously distort a review of a product. This damages the online marketplace, trust of the consumer, and impacts consumer behavior.

Thus, the main concern the current project is trying to address is:

How can an automated system be made that accurately and reliably identifies fake or misleading reviews in the Arabic language on e-commerce platforms?

As is to be expected, there are some challenges that arise from this.

- news and shiny. Most Arabic review text is unstructured, and informal language, slang, and other forms of noisy content can be expected.
- There are many fine and subtle linguistic features that can be learned by a model to identify genuine and deceptive reviews.
- There is a need to accurately determine the authenticity of the review text via the application of machine learning.
- The ability to access a review text that is either in a large collection of review documents or in an isolated state, and in either case to provide the user with an environment that is responsive.

There is machine learning content, and it will make a clear contribution to e-commerce platforms content moderation.

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

In the most recent works about Arabic Language Processing and the issues of deceptive content on e-commerce platforms this research does the following:

- **Building an Arabic Fake-Review Identification Tool:** A fully automated ML-based fraud detection system for Arabic deceptive reviews. The system incorporates a complete preprocessing, feature extraction, model training, and modeling assessment pipeline structured around the complex and rich system of the Arabic language. .
- **Alternative ML Models Assessment:** Inclusion of different classification algorithms: LR, Naive Bayes, Linear SVM, and Word2Vec models. A spate of models are utilized in an assessment of the models and review their performance in a classification task. This assessment adds to the Arabic review authenticity classification.
- **Cleaning and Structuring Arabic Reviews Dataset:** The developed model includes the +2,000-strong Arabic Dataset dedicated to fake reviews. The Dataset undergoes a full cycle of Processing in the domain of Dirty Data, in a structured format, available, and Accessible. The dataset available increases the limited structured resources around the Arabic language.
- **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 discusses Arabic text analysis challenges in misleading-content detection and proposes techniques to address ever-present challenges. The results contribute to avoiding trust erosion in automated solutions in e-commerce and provide building blocks for future studies to automate solutions that nurture trust on e-commerce platforms.

1.4 Thesis Organization

This thesis is organized as follows:

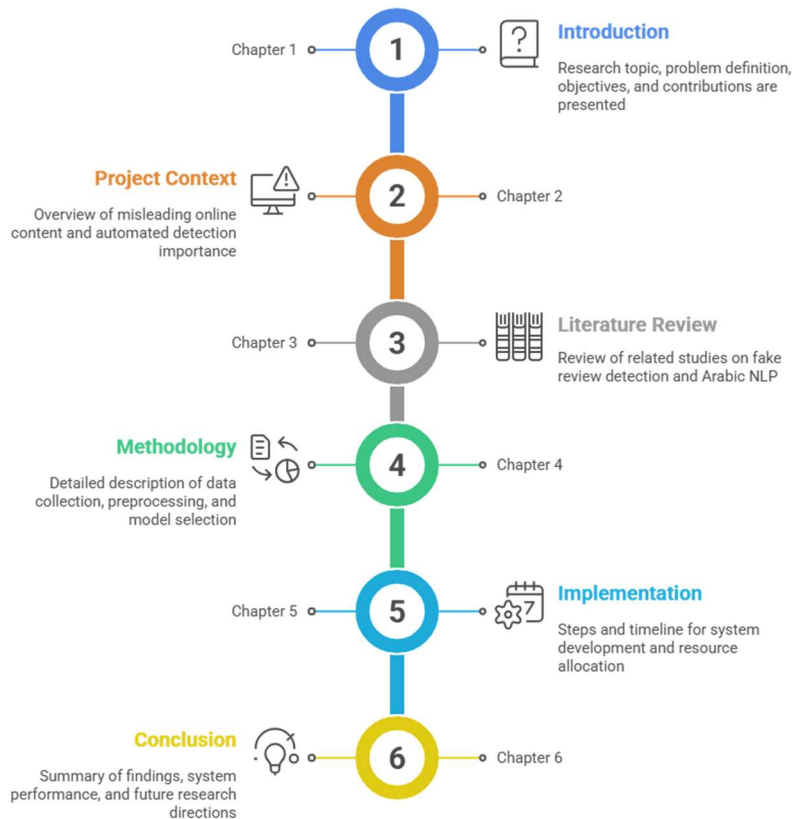


Figure 1 Thesis Organization and Development Process.

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

Because of the rapid growth of reviews and ratings on e-commerce sites, reviews and ratings have become critical to purchase decisions. However, the expansion of reviews and ratings has also led to the expansion of misleading reviews, fake advertisements, and inaccurate product descriptions. Such content misleads users, erodes trust in the platform, and damages the credibility of e-commerce. Earlier studies in this area focused on traditional language processing and text analysis. Such studies, however, concentrated on the English language. Meanwhile, the Arabic language, and its different dialects and lexical styles, has also been widely used, although it remains under discussed in analysis studies. Additionally, the limited availability of annotated datasets and the challenges of processing Arabic text have only intensified the urgency of this issue. This sets the stage for the objectives of this research focused on finding methods to recognize misleading reviews in Arabic language reviews. The goal is to eliminate the content accurately and efficiently using machine learning (ML) and natural language processing (NLP).

3.2 History

Classical machine learning approaches, such as Logistic Regression and Naive Bayes, feature time series analysis as used in SVM reviews. The use of TF-IDF and Bag of Words feature sets, allows for classification of the text, albeit for the more elementary categories of classification.

Research has shown that systems that utilize Natural Language Processing (NLP) techniques such as Word2Vec, GloVe, and fastText are able to capture contextual meaning and significance more effectively. However, most of these studies examined a single model and did not do a thorough comparison of different algorithms. Moreover, most of these studies focused on English

datasets and largely ignored Arabic datasets. This is a particular gap in the review analysis of Arabic writings. Thus, driven studies targeting the Arabic Misleading Content are also limited. The existing studies face major obstacles such as relative dialect diversity, the absence of annotated corpora, and the complexity of the morphology. Others used deep learning techniques, but the results are often unsatisfactory due to the high demand of resources and lack of data. Compared to such studies, the current research project has numerous merits, such as the following: a) It studies and analyzes multiple machine learning models as opposed to a single model. b) It introduces the combination of TF-IDF and Word2Vec to achieve a superior representation of the features. c) It concentrates on the Arabic Misleading Reviews, thus filling in the gap of lack of studies. d) It provides an interactive and tangible Streamlit program, thus providing something usable in the real world, something that many of the previous works did not do.

Chapter 4 Proposed Approach

In addressing the goals of this research, a multi-step approach was taken to develop an effective system to detect fraudulent Arabic reviews. This approach takes a methodical path, as each stage is designed to guarantee the model's ability to comprehend Arabic texts while ensuring the effective extraction of patterns to discriminate between actual reviews and fraudulent ones.

4.1 Data Collection

This project entails the collection of a large volume of data such as consumer reviews and product descriptions from disparate e-commerce platforms. This data will be collected from the public domains of websites such as Kaggle and Amazon, bearing in mind the legal implications and acceptable use policies of each platform. This data will take the form of text, in the comments of the users, and the titles of the products along with their descriptions to form a detailed database for the later stages of analysis and modeling.

4.2 Data Preprocessing

In a bid to allow the machine learning models to function optimally, the Arabic reviews must go through a number of pre-processing stages that work to clean and format the text before looking to train the models. Such tasks as removal of text noise, character canonicalization, removal of extraneous punctuation, and other Arabic language processing techniques are used to standard the text before training. Such measures are taken to the training and test sets to increase result reliability and model calibration.

- Removing specific patterns: Previously existing URLs and non-Arabic characters are expunged and also editing text blocks in accordance to predefined word formats.
- Tokenization: The cleaned text is divided into symbols (words) based on the number of whitespaces or white characters in general.
- Stopwords Removal: Using NLTK, Arabic stopwords (e.g. على، من) or tokens are removed so that resulting tokens have, or preserve, some semantic content.
- Stemming: The NLTK's ISRISemmer is used to stem or truncate words to their roots, etc. which also increases generalization and lessens the size of the vocabulary.
- Feature Representation or Extraction: for TF-IDF and for word embeddings(turning text to numeric for machine learning) two widely used feature extraction or representation methods would be carried out.

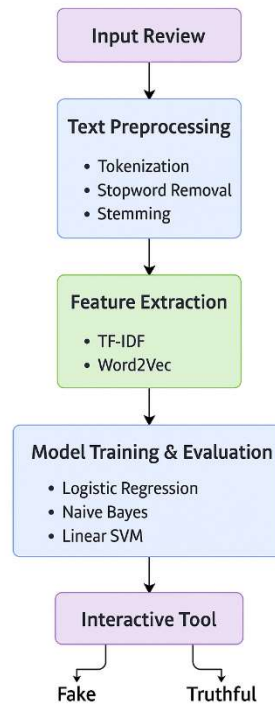


Figure 2 Proposed Framework.

4.3 Model/Method Selection

Out of the three machine learning models and the embedding-based model, the machine learning models were chosen based on their ability to handle text, so as to compare the performance of each of the models on the content classification business task, while each of the models provided a different functionality to focus on distinct evaluation of text representation.

- The TF-IDF + LR model is a benchmarked and recognized linear classifier that is highly efficient and effective due to its simplicity for text classification, and also for sparse text feature classification and evaluation in order to identify and filter out fraudulent or counterfeit reviews.
- TF-IDF + Linear SVM is a boundary-based classifier which performs very well for high-dimensional text data and achieves high accuracy rates on binary classification problems.
- TF-IDF + Naive Bayesian is one of the fastest methods and is a very effective classifier for short texts. Therefore, it is a very low-cost, effective and fast algorithm for this task.

- Word2Vec + LR adds a dense word embedding layer to capture the meaning and context of what is being said. Other than TF-IDF, this adds a richer linguistic detail.

Evaluating these models helps us to understand their effect and ascertain a suitable one for detecting misleading data on E-commerce sites.

4.4 Testing and Evaluation:

Some of the evaluation metrics to be used include:

- Accuracy: the rate at which the model classifies the content correctly. This is the overall classifier performance.
- Precision: This is the percentage of misleading content that is relevant to the flagged content.
- Recall: This is the percentage of misleading content that the model was able to uncover.
- F1-Score: This is the evaluation of average between Precision and Recall and used to give a balancing metric.
- For the model to be accepted, it should demonstrate the ability to perform on different strata of the data and be able to generalize. This will be captured through cross-validation.

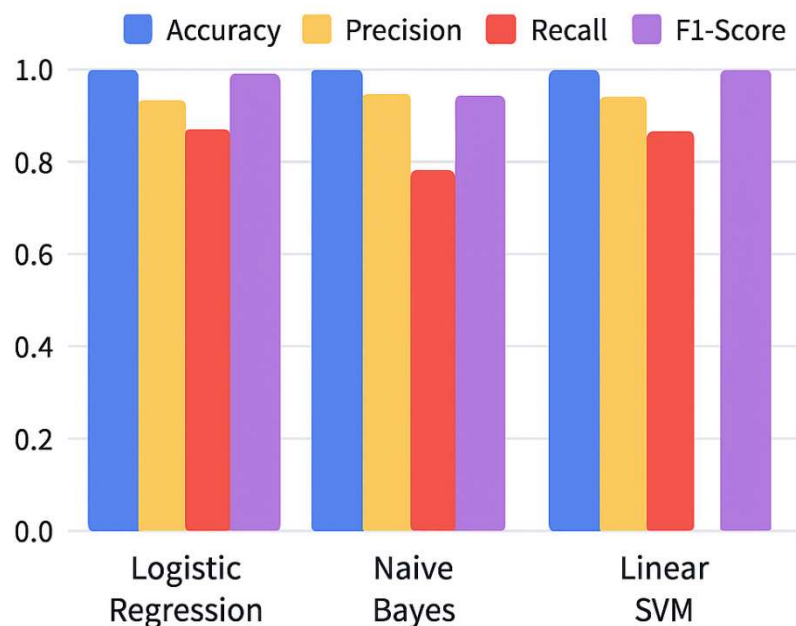


Figure 3 Model Performance Chart.

4.5 Expected Results

Here are the expected outcomes of the project:

- **Gained Enhanced Detection of Misleading Content:** With the project's developed machine learning models, e-commerce platforms are expected to have greater capability of misleading reviews, inappropriate product descriptions, and more, thereby increasing exposure to fraudulent data and reducing user trust.
- **Gained Additional Knowledge for Arabic NLP:** Integrating several preprocessing techniques, the project aims to compare the classical machine learning models and embedding techniques to provide additional knowledge for Arabic NLP.

Chapter 5

Project Steps and Timeline

This chapter explains what the proposed system will look like, and how it will function, detailing the workflow from initial data collection through to the last step, evaluation and reporting. The project will happen in data preparation, model development, performance evaluation, and final fitting, all of which happen in four main steps.

5.1 Step one: Data collection and preparation

This stage is the extraction of relevant data sets with product descriptions and customer reviews. After relevant data are collected, they are subjected to data cleaning and improvement of quality and consistency factors such as removal of urls, Arabic text normalizing, stop word removal, tokenization and stemming. After data cleaning, the text data will be processed and converted to numerical data with TF-IDF and Word2Vec techniques. Then,

the dataset will be split to training and testing to enable evaluation of the model with some level of confidence.

5.2 Step 2: Developing and training the model

This is the phase that training happens. The focus should be on model training, which comprises of some machine learning models such as Logistic Regression, Naive bayesian, Linear Support Vector Machine, and Word2Vec model, in addition to a model training custom-built interface on Streamlit designed for ease of user testing and interaction to be flexibly added to the system.

5.3 Step 3: Testing and Evaluation

- Standard metrics like accuracy, precision, recall, and F1 score, will be used on the trained models and assessed, focusing on the performance of the models and determining their overall effectiveness and their strengths and weaknesses. In addition, this evaluation phase seeks to determine how well the models can be used on completely new and unseen data.
- 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 Step 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.

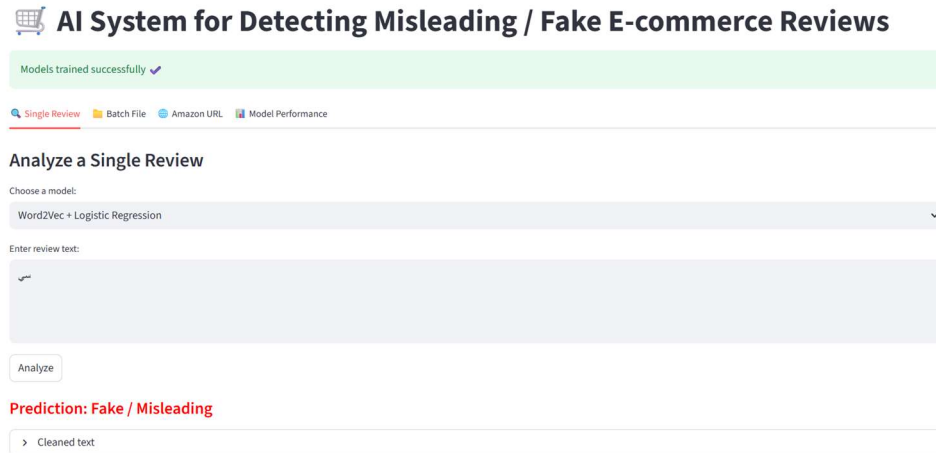


Figure 4 Streamlit UI Screenshot.

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

This thesis describes an intelligent system for detecting misleading Arabic reviews from e-commerce sites using Natural Language Processing and Machine Learning. The approach taken was a series of interconnected steps including, first, text pre-processing and feature extraction with both TF-IDF and Word2Vec. Then, a number of models were trained, including Logistic Regression, Naive Bayes, and Linear SVM, and performance of the models was evaluated using a range of well-established scientific performance indicators. The results showed Naive Bayes the best overall performance, demonstrating an ability to analyze short, direct Arabic texts.

In addition, this research has produced a functioning application on Streamlit, to analyze single reviews or entire datasets, thus making the system operational. This work goes some way to filling a significant research gap in the detection of misleading Arabic reviews and will support efforts to advance the development of other systems.

6.2 Future Work

Working with cutting-edge technological advancements, the developed tool demonstrates effectiveness in identifying deceptive Arabic reviews, however, certain aspects remain to be developed in order to facilitate overall effectiveness and versatility. Work in the future could involve the inclusion of more sophisticated deep learning models, the diversification of the dataset to encompass more types of reviews, and enhancements in the application interface to improve usability and flexibility for end-users. Some of the spearheaded potential modifications may involve Looking through the prism of potential future enhancements, the hub of activities for the future lies in the following for the initial actions/ steps, it covers the following:

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.*