

FAKE REVIEW DETECTION IN ARABIC TEXT

BY

Reem Abdulrahman Abdullatif AlJunaid

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

May 2024

FAKE REVIEW DETECTION IN ARABIC TEXT

Reem Abdulrahman Abdullatif AlJunaid

COMPUTER SCIENCE

May 2024

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by Reem AlJunaid under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**.



Dr. Motaz AlFarraj
Department Chairman



Dr. Suliman Al-Homidan
Dean of Graduate Studies

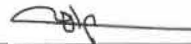
Date



Dr. Husni Al-Muhtaseb
(Advisor)



Dr. Moataz Ahmed
(Member)



Dr. Rabeah AlZaidy
(Member)



© Reem Abdulrahman Abdullatif AlJunaid

2024

*I dedicate this to my parents, my husband, my sisters, my brothers, and my friends for
their constant love and support.*

ACKNOWLEDGMENTS

At the very beginning, I would like to express my gratitude to the almighty Allah for giving me the energy, ability, and capabilities to prepare this dissertation and for giving me the ability to work hard successfully during my whole master's degree. All praise to Allah that the master's degree went smoothly with a great experience.

During my training, many people have helped me generously, whom I would like to mention here with deep gratitude. Words actually will never be enough to express my gratefulness for them. First and foremost, I would like to thank my advisor, Dr. Husni Al-Muhtaseb, for the guidance, encouragement, and advice he has provided throughout my dissertation. I have been extremely lucky to have an advisor who cared so much about my work, and who responded to my questions and queries so promptly. I equally express my most profound appreciation to the dissertation committee members, Dr. Moataz Ahmed and Dr. Rabeah AlZaidy. They always granted some of their valuable time to guide me and provide me with enlightening comments. Special thanks to my colleagues, Aisha AlAnsari and Asma Yamani, for all that they have taught me during my master's degree. They were always patient and willing to help when I had questions. I would also like to express my gratitude to my family for their patience, continued support, and encouragement. Furthermore, I extend my gratitude to Imam Abdulrahman Bin Faisal University for generously sponsoring my master's studies. I would also like to express my great thanks to all students, staff, and faculty members at King Fahd University of Petroleum & Minerals (KFUPM) who contributed to making my master's degree a valuable, fabulous, and unforgettable experience. |

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS.....	VI
LIST OF TABLES	IX
LIST OF FIGURES	XII
LIST OF ABBREVIATIONS	XIII
ABSTRACT.....	XVIII
ملخص الرسالة	XX
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement.....	2
1.2 Thesis Objectives.....	3
1.3 Thesis Questions.....	4
1.4 Thesis Motivation	4
1.5 Thesis Contribution	5
1.6 Thesis Methodology	6
1.7 Thesis Structure	7
CHAPTER 2 BACKGROUND	9
CHAPTER 3 LITERATURE REVIEW.....	11
3.1 Sentiment Analysis	11
3.1.1 Sentiment Analysis Levels.....	11
3.1.2 Sentiment Analysis Pre-Processing.....	14

3.1.3	Sentiment Analysis Techniques.....	19
3.1.4	Literature Review on Arabic Sentiment Analysis for Online Products	23
3.2	Fake Reviews Detection	30
3.2.1	Targeted Fake Reviews and Scope.....	30
3.2.2	Datasets	35
3.2.3	Feature Extraction	38
3.2.4	Fake Review Detection Methods	41
3.2.5	Literature Review on Fake Review Detection Using Sentiment Analysis	42
3.2.6	Literature Review on Fake Review Detection in Arabic.....	47
3.3	Gap Analysis	53
3.3.1	Targeted Scopes:	53
3.3.2	Datasets:	53
3.3.3	Features for Fake Review Detection:	53
3.3.4	Methods:.....	54
CHAPTER 4 APPROACH		55
4.1	Data Acquisition and Collection:	55
4.1.1	Saudi Amazon Dataset	56
4.1.2	AliExpress Dataset	57
4.1.3	Amazon dataset	59
4.1.4	Electronic Retailers Yelp Dataset	61
4.2	Sentiment Analysis	65
4.2.1	Data Preparation	66
4.2.2	Training Model.....	69
4.2.3	Sentiment Analysis Model Evaluation.....	73
4.2.4	Annotating the corpora.....	76

4.2.5	Building the Lexicons	78
4.3	Fake Reviews Detection	79
4.3.1	Feature Engineering	80
4.3.2	Fake Reviews Detection Approaches	85
CHAPTER 5 CONCLUSION		120
5.1	Remarkable Findings	122
5.2	Future Work	124
REFERENCES		126
VITAE.....		143

LIST OF TABLES

Table 1: Previous studies on Arabic Sentiment Analysis for Online Products.	28
Table 2: The utilized datasets for fake reviews from previous studies.	36
Table 3: The publicly available and annotated datasets for fake reviews detection.	37
Table 4: Previous studies on Fake Review Detection using Sentiment Analysis.	46
Table 5: Targeted Arabic fake reviews.	52
Table 6: The distribution of the Saudi Amazon dataset.	57
Table 7: The distribution of the AliExpress dataset.	59
Table 8: The distribution of the authentic reviews in the Amazon dataset.	60
Table 9: The distribution of Electronic Retailers Yelp dataset.	63
Table 10: The publicly available and annotated datasets for fake reviews detection.	64
Table 11: The training and testing sets of Authentic Amazon Reviews for SA using 2 labels.	68
Table 13: The training and testing sets of Authentic Electronic Retailers Yelp Reviews for SA using 2 labels.	68
Table 14: The training and testing sets of Authentic Electronic Retailers Yelp Reviews for SA using 3 labels.	68
Table 15: The results of Arabic sentiment analysis on authentic Amazon reviews.	76
Table 16: The results of Arabic sentiment analysis on authentic Electronic Retailers Yelp reviews.	76
Table 17: The evaluation of MARABERT sentiment analysis model on Saudi Amazon and AliExpress corpora.	78
Table 18: The evaluation of CAMELBERT-Mix-SA sentiment analysis model on Electronic Retailers Yelp corpus.	78
Table 19: The collected lexicons.	79
Table 20: The utilized feature sets for fake review detection.	81

Table 21: The utilized equations to compute the fake review detection features.	82
Table 22: The distribution of the training and testing set for supervised fake review detection. 87	
Table 23: The tested values for parameter tuning for decision tree model.	90
Table 24: The tested values for parameter tuning for random forest model	90
Table 25: The tested values for parameter tuning for gradient boosting model.....	90
Table 26: The tested values for parameter tuning for light gradient boosting model.	90
Table 27: The tested values for parameter tuning for extreme gradient boosting model.....	91
Table 28: The supervised fake review detection results.....	92
Table 29: Supervised fake review results with and without sentiment-based features.	93
Table 30: Paired sample t-test results.	94
Table 31: Comparing the results of supervised models from the study with a previous study on the same dataset.	96
Table 32: The distribution of the training and testing set for supervised fake review detection using YelpZip.....	97
Table 33: Comparing the results of supervised models from the study with a previous study on the same dataset.	98
Table 34: Similar reviews by the same reviewer on the same product.	101
Table 35: Similar reviews by the same reviewer on different products.	101
Table 36: Similar reviews by different reviewers on the same product.	102
Table 37: Reviews with question format and ratings deviating from 3.	102
Table 38: Reviews lacking product experience and deviating from a rating of 3.	102
Table 39: Reviews with contradictory elements.	102
Table 40: Identical reviews labeled as authentic reviews in electronic retailers' yelp dataset. 103	

Table 41: Identical reviews labeled as authentic and fake reviews in electronic retailers' yelp dataset.	105
Table 42: A review contains an email labeled as authentic review in electronic retailers yelp dataset.	106
Table 43: Identical reviews labeled as authentic reviews in YelpZip dataset.	107
Table 44: Identical reviews labeled as authentic and fake reviews in YelpZip dataset.	108
Table 45: Results of SSML training using the labeled Yelp dataset, unlabeled Saudi Amazon and AliExpress datasets.	116
Table 46: Results of SSML training using the labeled YelpZip dataset, unlabeled Saudi Amazon and AliExpress datasets.	116
Table 47: Results of Semi supervised training using the labeled Yelp dataset only.	117
Table 48: Results of Semi supervised training using the labeled YelpZip dataset only.	117
Table 49: Best-performing models using Semi-supervised approach.	118
Table 50: Best performing PU-learning models using Semi-supervised approach.	119

LIST OF FIGURES

Figure 1: Example of fake reviews from [8].	10
Figure 2: Fake Review Detection Approach.	55
Figure 3: The construction process of the Saudi Amazon dataset.....	57
Figure 4: The construction process of the AliExpress dataset.	58
Figure 5: The construction process of the authentic Amazon reviews dataset.....	60
Figure 6: The construction process of Electronic Retailers Yelp dataset.....	62
Figure 7: The framework of conducting Arabic sentiment analysis.	65
Figure 8: Data preparation for Arabic sentiment analysis.....	66
Figure 9: The concept of Transfer Learning.....	70
Figure 10: The framework of conducting fake review detection.	80
Figure 11: Supervised learning fake review detection approach.....	86
Figure 12: The Performance of the models with and without including the sentiment-based features.....	93
Figure 13: The XGB feature importance.....	96
Figure 14: Fake review detection using self-training approach.	111
Figure 15: Fake review detection using PU-learning approach.	113

|

LIST OF ABBREVIATIONS

AdaBoost	:	Adaptive Boosting
ADJ	:	Adjective
ADV	:	Adverb
AFFINN	:	Affective Norms for English Words
AI	:	Artificial Intelligence
API	:	Application Programming Interface
AraBERT	:	Arabic Bidirectional Encoder Representations from Transformers
AraSenTi	:	Arabic Sentiment Tweets
AraXLNet	:	Arabic Extra-Long Neural Network
ARBL	:	Arabic ResNet-based Language Model
ArSAS	:	Arabic Sentiment Analysis System
ASTD	:	Arabic Sentiment Tweets Dataset
AUC	:	Area Under the Curve
BERT	:	Bidirectional Encoder Representations from Transformers
BOW	:	Bag of Words
BRAD	:	Books Reviews in Arabic Dataset
BTO	:	Binary-Term Occurrence
CA	:	Classical Arabic
CAMeLBERT	:	Computational Approaches to Modeling Language BERT
CART	:	Classification and Regression Trees
CC	:	Character Count

CN	:	Convolutional Network
CNN		Convolutional Neural Networks
CSDRDP	:	Content Similarity of Different reviewers on Different products
CSDRSP	:	Content Similarity of Different Reviewers on the Same Product
CSF	:	Content Similarity Footprints
CSSRDP	:	Content Similarity of Same Reviewer on Different Products
CSSRSP	:	Content Similarity of Same Reviewer on the Same Product
CSV	:	Comma Separated Values
DA	:	Dialectal Arabic
DFNN	:	Deep Feedforward Neural Network
DL	:	Deep Learning
DLMNN	:	Dynamic Local Memory Neural Network
DT	:	Decision Tree
EC	:	Email Count
FIM	:	Frequent Itemset Mining
FRD	:	Fake Review Detection
GAN	:	Generative Adversarial Network.
GB	:	Gradient Boosting
GBM	:	Gradient Boosting Machine
GBRF	:	Gradient Boosting Random Forest
GPT	:	Generative Pre-trained Transformer
HARD	:	Hotel Arabic Reviews Dataset
HTTP	:	Hypertext Transfer Protocol

IANFIS	:	Interval Type-2 Adaptive Neuro-Fuzzy Inference System
KNN	:	K-Nearest Neighbors
LABR	:	Lexicon-based Arabic Review dataset
LC	:	Link Count
LDA	:	Latent Dirichlet Allocation
LDA	:	Latent Dirichlet Allocation
LGBM	:	Light Gradient Boosting Machine
LR	:	Logistic Regression
LSA	:	Latent Semantic Analysis
LSTM	:	Long Short-Term Memory
MARABERT	:	Multimodal Arabic BERT
ML	:	Machine Learning
MLPNN	:	Multilayer Perceptron Neural Network
MRC	:	Maximum Review Count Per Day
MSA	:	Modern Standard Arabic
MSE	:	Mean Squared Error
NB	:	Naive Bayes
NLP	:	Natural Language Processing
NN	:	Neural Network
NWC	:	Negative Words Count
PCF	:	Product-Centric Footprints
POS	:	Part of Speech
PROD	:	Product Review Dataset

PSI	:	Product Sentiment Inconsistency
PSIC	:	Reviews with Product Sentiment Inconsistency Count
PU	:	Positive Unlabeled
PWC	:	Positive Words Count
R	:	Review Rating
RBF	:	Rating Behavior Footprints
RC	:	Reviews Count
RCR	:	Reviews with Rating Count
RD	:	Rating Deviation
RF	:	Random Forest
RPF	:	Reviewer Profile Footprints
S	:	Sentiment
S140	:	Stanford Sentiment140
S3VM	:	Semi-Supervised Support Vector Machine
SA	:	Sentiment Analysis
SC	:	Sentence Count
SemEval	:	Semantic Evaluation
SentiWordNet	:	Sentiment WordNet
SSML	:	Semi-Supervised Machine Learning
SML	:	Supervised Machine Learning
SMOTE	:	Synthetic Minority Over-Sampling Technique
SRI	:	Sentiment To Rating Inconsistency
SRIC	:	Reviews with Sentiment to Rating Inconsistency Count

STLI	:	Sentiment to Text Length Inconsistency
SVM	:	Support Vector Machine
TC	:	Tips Count
TF	:	Temporal Footprints
TF-IDF	:	Term Frequency-Inverse Document Frequency
TLC	:	Text Length Category
UCI	:	University of California, Irvine
USI	:	User Sentiments Inconsistency
WC	:	Word Count
Word2Vec	:	Word to Vector
WSF	:	Writing Style Footprints
XGB	:	Extreme Gradient Boosting

ABSTRACT

Full Name : Reem Abdulrahman Abdullatif AlJunaid

Thesis Title : Fake Review Detection in Arabic Text

Major Field : Computer Science

Date of Degree : May 2024

Online shopping has become a trendy way for people to purchase their products. In online shopping, customers heavily rely on the product reviews provided by other customers to adjust their purchasing choices. However, these reviews may be fake and created by spammers to mislead customers and to let them make wrong purchase decisions. This thesis introduces an approach for detecting Arabic fake reviews in online product evaluations, leveraging machine and deep learning techniques coupled with sentiment analysis. The methodology begins with collecting reviews from popular e-commerce platforms, resulting in a dataset of 12,865 reviews from Saudi Amazon and AliExpress. Sentiment analysis is conducted using traditional machine learning models (Logistic Regression, Decision Tree, Support Vector Machine) and pre-trained deep learning models (MARABERT, CAMeLBERT-Mix-SA) on a set of 10,500 authentic reviews from Amazon. MARABERT, excelling in ternary classifications with an 84.12% F-score, annotated collected reviews with sentiments, validated through semi-manual verification based on ratings. A lexicon set is constructed with 2,146 terms for Arabic online product reviews using a trained Logistic Regression model. The study introduces proposed sentiment-based features alongside content and behavior-based features to identify fake reviews. Two annotated datasets, Electronic Retailers Yelp and YelpZip, are utilized for training machine learning models employing both supervised and semi-supervised learning. Supervised

learning involves five models (Decision Tree, Random Forest, Gradient Boosting, Light Gradient-boosting Machine, and Extreme Gradient Boosting) to evaluate the proposed feature set's effectiveness. Results indicate substantial improvements compared to previous studies. The semi-supervised approach incorporates a rule-based method to identify explicit fake reviews, followed by annotation of remaining unlabeled reviews using Decision Tree and Random Forest models with Self-Training and PU-Learning variants. The Random Forest Self-Training model achieves a remarkable F1 score of 96.9%, outperforming existing models in the literature when integrated with the proposed feature set. This comprehensive approach demonstrates its efficacy in enhancing fake review detection in Arabic online product evaluations.

ملخص الرسالة

الاسم الكامل: ريم عبد الرحمن عبد اللطيف الجنيد

عنوان الرسالة:

التخصص: علوم الحاسب

تاريخ الدرجة العلمية: مايو 2024

أصبح التسوق عبر الإنترنت وسيلة عصرية للناس لشراء منتجاتهم. في التسوق عبر الإنترنت، يعتمد العملاء بشكل كبير على مراجعات المنتجات المقدمة من العملاء الآخرين لضبط خيارات الشراء الخاصة بهم. ومع ذلك، قد تكون هذه المراجعات مزيفة وقد أنشأها مرسلو البريد العشوائي لتضليل العملاء والسماح لهم باتخاذ قرارات شراء خاطئة. تقدم هذه الأطروحة طريقة للكشف عن المراجعات المزيفة باللغة العربية في تقييمات المنتجات عبر الإنترنت، والاستفادة من تقنيات التعلم الآلي والعميق إلى جانب تحليل المشاعر. تبدأ المنهجية بجمع المراجعات من منصات التجارة الإلكترونية الشهيرة، مما يؤدي إلى مجموعة بيانات مكونة من 12,865 مراجعة من أمازون السعودية وعلى إكسبريس. يتم إجراء تحليل المشاعر باستخدام نماذج التعلم الآلي التقليدية (Logistic Regression, Decision Tree, Support Vector Machine) ونماذج التعلم العميق المدربة مسبقاً (CAMELBERT-, MARABERT) Mix-SA) على مجموعة مكونة من 10500 مراجعة أصلية من أمازون. MARABERT، المتفوق في التصنيفات الثلاثية بدرجة 84.12% F، قام بتصنيف المراجعات المجمعة بالمشاعر، وتم التحقق من صحتها من خلال التحقق شبه اليدوي بناءً على التقييمات. تم إنشاء مجموعة مفردات تحتوي على 2,146 مصطلحاً لمراجعات المنتجات العربية عبر الإنترنت باستخدام نموذج Logistic Regression المدرب. تقدم الدراسة ميزات مقترحة قائمة على المشاعر إلى جانب الميزات القائمة على المحتوى والسلوك لتحديد المراجعات المزيفة. يتم استخدام مجموعتي بيانات مشروحتين، وهما Yelp و Electronic Retailers، و YelpZip، لتدريب نماذج التعلم الآلي التي تستخدم التعلم الخاضع للإشراف وشبه الإشراف. يتضمن التعلم الخاضع للإشراف خمسة نماذج (Decision Tree, Random Forest, Gradient

Boosting, Light Gradient-boosting Machine, and Extreme Gradient Boosting) لتقييم فعالية مجموعة الميزات المقترحة. تشير النتائج إلى تحسينات كبيرة مقارنة بالدراسات السابقة. يشتمل النهج شبه الخاضع للإشراف على طريقة قائمة على القواعد لتحديد المراجعات المزيفة الصريحة، متبوعة بتعليق توضيحي للمراجعات غير المسماة المتبقية باستخدام نماذج Decision Tree و Random Forest مع متغيرات التدريب الذاتي والتعلم PU. يحقق نموذج Random Forest درجة F1 رائعة تبلغ 96.9%، متفوقاً على النماذج الموجودة في الأدبيات عند دمجها مع مجموعة الميزات المقترحة. يوضح هذا النهج الشامل فعاليته في تعزيز اكتشاف المراجعات المزيفة في تقييمات المنتجات العربية عبر الإنترنت.

CHAPTER 1

INTRODUCTION

Sentiment analysis, also known as sentiment mining or opinion mining, stands as a vital classification task within the realm of Natural Language Processing (NLP). It involves the intricate process of gathering and scrutinizing people's opinions expressed through text or speech to unveil attitudes, sentiments, and emotions toward a particular entity—be it a product, service, or even an idea [1]. Sentiment analysis, also known as sentiment mining or opinion mining, stands as a vital classification task within the realm of Natural Language Processing (NLP). It involves the process of gathering and scrutinizing people's opinions expressed through text or speech to unveil attitudes, sentiments, and emotions toward a particular entity—be it a product, service, or even an idea [1]. The use of sentiment analysis has extended into various fields, demonstrating its significance in commercial [2], political [3], and educational [4] domains. It empowers businesses to gauge brand popularity, assess reputation, and scrutinize public perception over specific periods. Furthermore, it aids in evaluating new products, features, or services by analyzing user feedback, enabling enterprises to refine future business strategies. In political contexts, sentiment analysis becomes a potent tool for tracking public opinion about politicians or events, gleaned from social media posts. In education, it helps assess the efficacy of learning technologies through the analysis of student feedback.

The surge in e-commerce applications for product acquisition has been particularly noteworthy in recent times. A 2020 report indicates that around 80% of active internet users in Saudi Arabia engage with e-commerce applications [5]. Notably, one of the primary reasons people prefer online shopping is the ability to peruse reviews from fellow customers on e-commerce platforms [6]. However, the reliability of these reviews is often compromised, with some being deliberately misleading. Fake reviews, characterized by deceptive content intended to mislead or capture attention, pose a significant challenge with far-reaching negative consequences for both customers and the e-commerce industry.

Despite significant attention given to detecting fake reviews in English, there exists a notable gap in research focusing on Arabic reviews. The Arabic language's complex syntax and rich morphology present unique challenges, making the detection of fake reviews in Arabic even more demanding. Consequently, there is an urgent need to devise effective methods for distinguishing fake reviews from authentic ones in Arabic.

1.1 Problem Statement

During the last few years, there has been a huge expansion of e-commerce applications especially after the COVID-19 crisis. These e-commerce applications allow the customers to express their opinions and expertise on the product through the reviews sections. Online customers heavily rely on these written reviews to make decisions about the products they want to purchase. As businesses' reputations and profits are directly affected by the credibility of these reviews, some businesses utilize spammers just to generate fake positive reviews and consequently to maintain their businesses. Therefore, customers' purchasing decisions will be exploited by these fake reviews. Many studies have proposed different methods for detecting fake English reviews. However, fewer attempts were found to detect

fake Arabic reviews. This research aims to bridge this gap and contribute to the development of effective strategies for detecting fake reviews in Arabic e-commerce reviews.

1.2 Thesis Objectives

This research seeks to attain the following specific objectives, organized sequentially:

- **Methodology Development:**

Formulate a comprehensive and transparent methodology for the systematic development of an automated classification system for Arabic reviews. This methodology will serve as the foundation for subsequent stages.

- **Building the Necessary Resources:**

Establish and curate the requisite resources for training and testing the classification model. This involves creating and collecting a dataset of labeled Arabic reviews, encompassing both fake and authentic instances, and ensuring its adequacy for the subsequent stages of model development.

- **Model Design and Development:**

Design and implement a classification model. This model will leverage AI and sentiment analysis techniques, incorporating insights from the developed methodology and utilizing the established resources.

- **Model Evaluation:**

Conduct a comprehensive evaluation of the developed classification model. This assessment will encompass rigorous testing to gauge the model's accuracy, precision,

recall, and other pertinent metrics. The goal is to ensure the model's effectiveness and reliability in discriminating between fake and authentic reviews in the Arabic context.

By systematically addressing these objectives, this research aspires to contribute to the advancement of fake review detection in Arabic, fostering a more trustworthy and secure online shopping environment within the e-commerce domain.

1.3 Thesis Questions

There are some questions that we are trying to answer while conducting this research work.

The significant questions that will be answered are:

- **Q1:** *How can we utilize different scopes in detecting fake reviews?*
- **Q2:** *How can we construct a dataset of Arabic reviews for online products?*
- **Q3:** *What are the behavioral and content features that can be used to detect fake reviews?*
- **Q4:** *What types of classifiers can be utilized to detect fake reviews?*

1.4 Thesis Motivation

Our research is driven by a series of compelling motivations within the domains of Arabic sentiment analysis, natural language processing, and the pervasive issue of "fake reviews." Firstly, the escalating prevalence of fake reviews in e-commerce applications underscores the critical need for robust methodologies tailored to the nuances of the Arabic language. Furthermore, the dearth of Arabic product review resources and the limited scope of existing studies in this domain accentuate the urgency of our investigation. To address these gaps, we recognize the necessity for a domain-specific Arabic annotated corpus dedicated to product reviews, as well as the development of sentiment lexicons tailored for Arabic

sentiment analysis within the e-commerce context. These foundational resources are crucial for training and enhancing the performance of sentiment analysis models.

1.5 Thesis Contribution

The key contributions in this thesis are outlined below:

- **Development of Specialized Datasets:**

Compilation of comprehensive datasets featuring online product reviews in Arabic, labeled with sentiments.

Construction of a dedicated dataset of online product reviews in Arabic, labeled to distinguish between fake and authentic reviews.

- **Construction of Sentiment Lexicons:**

Creation of sentiment lexicons tailored specifically for online product reviews in Arabic.

- **Innovative Approaches to Dataset Labeling:**

Introduction of an innovative approach for labeling online product reviews, categorizing them as either fake or authentic, while concurrently capturing the associated sentiments.

- **Feature Proposal for Fake Review Detection:**

Proposition of a set of sentiment-based features designed specifically for the detection of fake reviews within the realm of online product evaluations.

- **Empirical Investigations:**

Rigorous empirical investigations examining the impact of sentiment-based features on the task of fake review detection in online product reviews.

1.6 Thesis Methodology

The research methodology is designed as follows:

1) Conducting a Literature Review:

- Review existing studies that focus on Fake Review Detection in both Arabic and English languages.
- Comprehensive analysis of studies dedicated to Sentiment Analysis in the context of online product reviews, considering the Arabic language.
- Exploration of available datasets for Fake Review Detection in Arabic and English.
- In-depth investigation of available datasets for Sentiment Analysis of online product reviews in Arabic.

2) Designing a Framework for Fake Review Detection:

- Formulation of a robust framework for the detection of fake reviews, integrating Artificial Intelligence (AI) and advanced Sentiment Analysis techniques. The design leverages insights gained from the literature review.

3) Building Resources:

- **Data Collection and Processing:**
 - Detailed data collection and processing for available datasets related to fake reviews of online products, ensuring quality and relevance.
 - Data extraction specifically focused on online product reviews in Arabic, employing systematic methods.
- **Labeling and Lexicon Construction:**

- Labeling of collected and extracted reviews with sentiments, enabling the creation of sentiment lexicons.
- Fake Review Labeling:
 - Manual labeling of a subset of extracted Arabic reviews, categorizing them as fake or authentic.
- 4) Feature Calculation:
 - Calculation of content-based and behavioral-based features for fake review detection across all datasets.
- 5) Developing the Classification Model:
 - Utilization of the designed framework to develop an advanced classification model for the detection of fake reviews. The model incorporates features derived from sentiment analysis and other content-based and behavioral features.
- 6) Evaluating the Classification Model:
 - Detailed evaluation of the developed classification model, considering metrics such as accuracy, precision, recall, and F1 score.

1.7 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2:** This chapter provides a foundational background on fake reviews, exploring their definition and various types.
- **Chapter 3:** In this chapter, we conduct a comprehensive survey of the literature on existing approaches and datasets about Fake Review Detection in both Arabic and

English languages. We summarize our findings, identify gaps, and establish the thesis objectives. The literature review also provides the necessary background on sentiment analysis techniques for online product reviews.

- **Chapter 4:** This chapter is dedicated to the design of a Framework for Fake Review Detection, which seamlessly integrates Artificial Intelligence (AI) and Sentiment Analysis techniques.
- **Chapter 5:** In this section, we showcase the significant discoveries made during the study, along with insights into future research directions.

CHAPTER 2

BACKGROUND

In the realm of online reviews, the proliferation of fake reviews has become a pervasive issue, often characterized as spam or deceptive content generated by individuals known as spammers. A nuanced classification system categorizes fake reviews into three main types [7]:

- **Untruthful Reviews:**

These reviews are authored with the explicit intent of either promoting or tarnishing the reputation of a business or a product. Their deceptive nature makes them challenging to identify through casual reading, and they are often referred to as false or deceptive reviews.

- **Brand-Only Reviews:**

This category encompasses reviews specifically crafted for a particular brand, seller, or product manufacturer. Such reviews may serve promotional or adversarial purposes, focusing solely on a particular entity within the market.

- **Non-Reviews:**

Non-reviews take the form of advertisements or content that is entirely irrelevant to the product or service being discussed. Though they pose little harm and are easily identifiable, they fall under the umbrella of disruptive spam reviews.

While brand-only reviews and non-reviews are relatively straightforward to identify, untruthful reviews pose a substantial challenge due to their close similarity and deceptive

nature. In a study by [8], examples of detected fake reviews are illustrated in Figure 1, specifically untruthful reviews, were highlighted. These reviews, identified as fake, exhibited similar content and were authored within a short period by different users, exemplifying the complexity of discerning false reviews.

In the context of this research, our focus centers on the detection of all three types of fake reviews associated with online products. By addressing these distinct categories, we aim to contribute to the development of robust methodologies for identifying and mitigating the impact of fake reviews in the online domain.

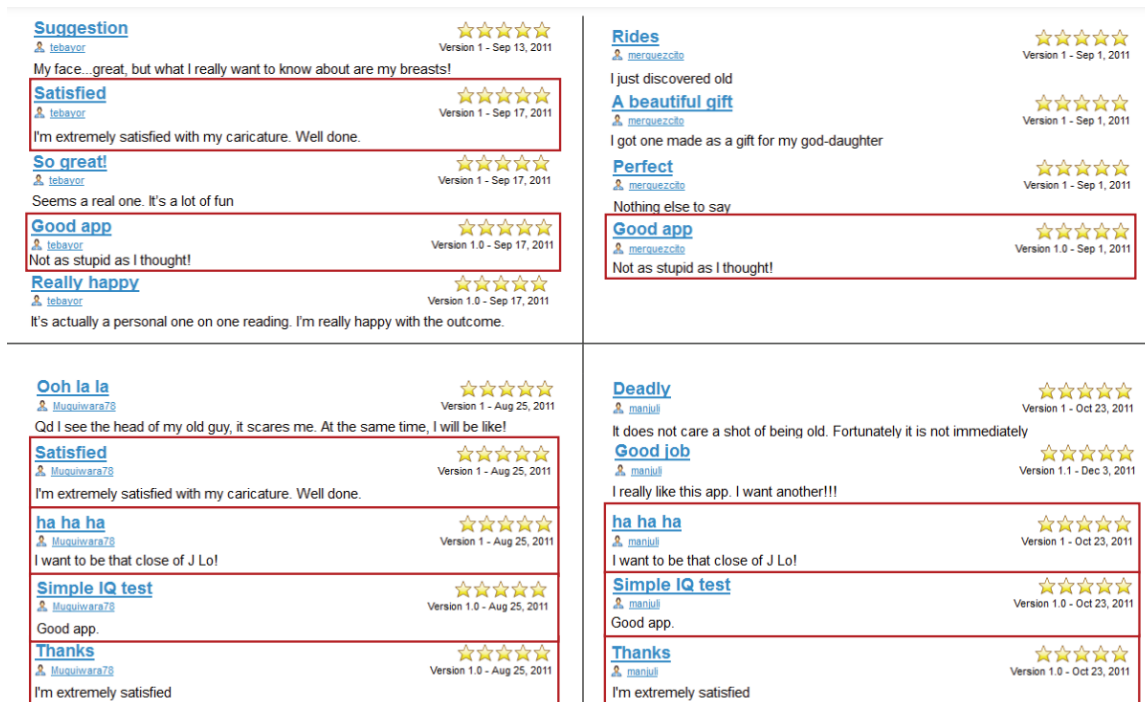


Figure 1: Example of fake reviews from [8].

CHAPTER 3

LITERATURE REVIEW

This chapter provides a comprehensive review of the existing literature, focusing on two pivotal aspects: Sentiment Analysis and Fake Reviews Detection. The primary objective of this chapter is to explore the current state of knowledge in these domains and to lay a solid foundation for the subsequent chapters. Additionally, this chapter identifies gaps in the literature to ensure that our research contributes meaningfully to the advancement of knowledge in Sentiment Analysis and Fake Reviews Detection.

3.1 Sentiment Analysis

The literature review on sentiment analysis encompasses a comprehensive exploration of various aspects, systematically organized into four sections, sentiment analysis levels, sentiment analysis pre-processing, sentiment analysis techniques, and lastly, a dedicated section focuses specifically on a literature review of Arabic sentiment analysis for online products, offering a detailed synthesis of relevant studies within this niche with their notable findings.

3.1.1 Sentiment Analysis Levels

The task of sentiment analysis has been investigated at many different levels, Document, Sentence, and aspect level. Based on whether the study is aimed to perform the sentiment analysis over a whole document (document level), a set of connected sentences (sentence level), or specific aspects of entities in the text (aspect level). The difference between the

tasks of sentiment analysis at each of these levels will be explained in detail in the following sections.

1) Document Level Sentiment Analysis

The sentiment analysis task at the document level is to determine a single polarity for a whole document, whether a whole document expresses a positive, negative, or natural sentiment [9]. It works with the assumption that the document is classified based on the overall sentiment of a single opinion holder like a product review [10]. As an example, the classifier analyzes whether a product review expresses an overall positive, negative, or natural sentiment. Most of the utilized approaches at this level are machine learning-based approaches [6]. Various traditional machine learning models have been used such as Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB) [11], [12]. However, deep learning-based (DL) approaches have captured the attention of researchers in recent years due to their superior performance at this level [11], [13]. Despite this, this approach of sentiment analysis is currently not used a lot [9], [14] since the classification at this level does not always produce accurate results compared with the sentence-level approach [10], [13], [15].

2) Sentence Level Sentiment Analysis

The sentiment analysis task at the sentence level is to analyze each sentence to determine its polarity, whether expresses a positive, negative, or natural sentiment [16]. According to the polarity of the sentence, a subsequent analysis can be performed to determine if it is a subjective or objective sentence [9]-[11]. This type of Sentiment analysis is useful when analyzing a document with a mix of diverse sentiments or with conditional or ambiguous statements [9]. At this level, the polarity of each sentence will be determined independently

using the same approach of the document-level sentiment analysis but with more preprocessing steps and more training data. These polarities can be used later to calculate the overall polarity of a document. Despite the importance and usefulness of sentence and document-level sentiment analysis, they are unable to describe the entity in detail [17], and unable to find out exactly what people like and dislike [10]. Further, most of the studies conducted on Arabic sentiment analysis applied the task at this level [14].

3) Aspect Level Sentiment Analysis

This type of sentiment analysis determines the sentiment based on specific aspects of entities. The task of sentiment analysis at the aspect level doesn't concentrate only on whether the product or the review is positive or negative, but it also concentrates on which aspect of the product people like or dislike. The process works first by calculating the polarity of each aspect of the entity and then an aggregate sentiment is calculated for the sentence [18]. Many real-life applications require this level of detailed analysis. Companies can identify which aspects of their products are interesting to customers to enhance their products concerning these aspects [19], [20]. However, aspect-based sentiment analysis requires a variety of subtasks that include the extraction of implicit or explicit aspect terms, the calculation of weight terms, the detection of aspect groups, and the calculation of the total polarity of aspect groups [21]. Assigning a polarity to each mined aspect can be achieved using either machine learning algorithms or a lexicon-based approach [9]. Performing the sentiment analysis at this level in Arabic requires further investigation, although, there have been a few attempts to do so [14].

3.1.2 Sentiment Analysis Pre-Processing

1) Data Extraction and Collection

Obtaining a suitable dataset is a critical step in sentiment analysis. Several methods can be employed to gather and prepare data for sentiment analysis tasks, including: [10]:

1. **Public Datasets:** Numerous sentiment-labeled datasets are freely available on the internet, created by companies or researchers. These datasets span various domains and can be directly downloaded from websites like Kaggle, the UCI Machine Learning Repository, and Sentiment140. These resources provide a solid foundation for initial experiments and model training.
2. **Crowdsourcing Platforms:** Platforms such as Amazon Mechanical Turk [22] enable the creation or annotation of data through crowdsourcing. This approach is advantageous for rapidly generating large volumes of labeled data. Workers on these platforms can be tasked with labeling data according to predefined sentiment categories, significantly speeding up the data collection process.
3. **Manual Annotation:** For smaller datasets, manual annotation is a viable option. This process involves human annotators labeling each data point according to its sentiment. Although time-consuming, manual annotation ensures high-quality, accurately labeled data, which is crucial for training reliable sentiment analysis models.
4. **Application Programming Interface (API):** APIs provide a streamlined method for extracting data from websites with associated sentiment labels. For instance, the Twitter API allows developers to collect tweets along with metadata that can be used for sentiment analysis. APIs offer an efficient and rapid means of data

extraction, though they may not be available for all websites and can have usage limitations.

5. **Sentiment Lexicons:** Sentiment lexicons are dictionaries that link words to their sentiment polarity (positive, negative, or neutral). These lexicons can be applied to text data to infer sentiment based on word presence. This method is useful for initial sentiment labeling and can complement other data collection techniques.
6. **Supervised Learning from Scratch:** Starting with a small, manually labeled seed dataset, a sentiment classifier can be trained. As the classifier's accuracy improves, it can be used to automatically label additional data. This iterative process gradually expands the dataset, enhancing the model's performance over time.
7. **Transfer Learning:** Pre-trained sentiment analysis models can be fine-tuned for specific domains or datasets, requiring a smaller amount of labeled data. Transfer learning leverages existing knowledge from general sentiment analysis models, making it a powerful approach when labeled data is scarce.

Each of these methods offers unique advantages and can be selected based on the specific requirements and constraints of the sentiment analysis task at hand.

2) Data and Arabic Text Pre-Processing

High dimensionality and large size of textual data are the major obstacles to data mining [23]. NLP tools and techniques should be applied to preprocessing the data and text in order to reduce these two factors. Cleaning the dataset involves converting some values to a standard range format like the rating values (e.g. converting 3.8 to 4 and Excellent to 5), and removing some instances such as duplicated instances, instances with empty or null values, or instances with long text. After cleaning the dataset, the text pre-processing steps

should take place for each instance of the dataset. The pre-processing of Arabic text is the most important step in all Arabic NLP applications as the Arabic language is characterized by its rich morphology and complex syntax [24]. In previous studies on sentiment analysis of Arabic text, a variety of preprocessing steps have been employed [14], [25], [26], including:

- Non-Arabic Text Removal: Eliminating all non-Arabic text.
- Text Filtering: Removing unnecessary words that do not affect the sentiment of the text (e.g., على، إلى، و، في)، as well as digits, special characters, and links.
- Diacritics Removal: Stripping diacritical marks (e.g., َ, ُ, ّ, ّ) from Arabic text (e.g., جيد to جَيِّد).
- Emoji Conversion: Converting emojis into text representations of their sentiments (e.g., 😊 to سعيد).
- Dialect Conversion: Translating regional variations or slang words into Modern Standard Arabic (MSA) (e.g., كثير to وايد).
- Text Normalization: Standardizing certain Arabic letters (e.g., آ، إ، أ to ا) and removing repeated letters (e.g., ممتاز to ممتاز) or elongated text (e.g., رائع to رائع).
- Spelling Correction: Correcting spelling mistakes in the text.
- Text Tokenization: Breaking the text into smaller units called tokens, which can be words, sub-words, or characters.
- Text Stemming: Reducing words to their stem forms (e.g., أعجبتني to أعجب).

3) Feature Extraction

Feature extraction plays a vital role in sentiment analysis tasks. The extracted features contribute heavily to the overall performance of sentiment analysis models [27]. Moreover, the choice of these features depends on the specific sentiment analysis task, language, and available resources. The features utilized for sentiment analysis can be categorized as follows [9]:

1. Lexical Features:

These features involve analyzing individual words or terms within the text. It includes sentiment lexicons, which assign sentiment scores or labels to words, as well as word frequencies, word presence/absence in the text.

2. Syntactic Features:

These features consider the grammatical structure of the text, including the use of part-of-speech (POS) tags, parsing information, and syntactic patterns. They help capture the relationships between words and their impact on sentiment.

POS tagging involves assigning a specific tag to each word in the sentence (e.g. noun, verb, adjective) based on its context or grammatical category. By employing stemming and part-of-speech (POS) tagging techniques, the authors of [28] achieved enhanced results in MSA news subjectivity classification compared to the existing state-of-the-art methods. However, the majority of POS taggers are designed for Modern MSA, with limited initial efforts focused on the Egyptian dialect [29]. As a result, many research studies opt not to include POS tags as features in their classification models, citing that existing POS taggers are unsuitable for social media text that includes dialectal elements [8].

Other Syntactic features can be obtained using a combination of different features. In [30], the authors utilized syntactic features which include POS tags along with n-grams, lemmas, a bag of words (BOW), and a bag of lemmas.

3. Semantic Features:

These features focus on the meaning and semantics of the text. They can involve techniques such as word embeddings, semantic role labeling, word sense disambiguation, and sentiment propagation through semantic networks.

4. Stylistic Features:

These features analyze the stylistic elements of the text, including capitalization, punctuation, emoticons, sentence length, and other textual characteristics associated with sentiment expression.

5. Contextual Features:

These features consider the contextual information surrounding the text, such as the sentiment of neighboring sentences or the overall sentiment of the document. They help capture the influence of context on sentiment interpretation. N-grams can be considered as both contextual and semantic features.

N-grams are a consecutive sequence of N elements, these elements can be a word, a character, or any other unit of analysis. N-gram features are utilized to measure the occurrence or frequency of these sequences within a given text. Unigrams, bigrams, and trigrams are the most used n-grams. Unigrams consist of single words, bigrams consist of two consecutive words, and trigrams consist of three consecutive words. The N-gram features have been utilized in various studies of Arabic sentiment analysis. Studies [31] and [23] found that unigrams surpass bigrams and trigrams based on the classification

performance. While [32] and [33] showed that utilizing trigrams resulted in optimal performance. The experiments carried out by [34] demonstrated that employing a combination of trigram features enhances classification accuracy.

3.1.3 Sentiment Analysis Techniques

Sentiment analysis techniques, also known as data mining or opinion mining, aim to automatically determine the sentiment or subjective information expressed in texts. These techniques employ various approaches and algorithms to analyze and classify sentiment in textual data. The choice of approach depends on different factors such as available data, task requirements, and the specific characteristics of the text or language being analyzed. Different techniques have their strengths and limitations, and selecting the most suitable technique is crucial for accurate sentiment analysis results. The categorization of sentiment analysis approaches can be based on different perspectives, such as the way the text is viewed or the level of detail in text analysis. In general, researchers categorize sentiment analysis approaches into three categories: (1) Machine Learning-Based, (2) Lexicon-Based, and (3) Hybrid approaches [10]. In this section, an overview of these approaches is presented.

1) Machine Learning Approaches

This is the most employed approach for sentiment analysis. This approach utilizes machine learning models along with linguistic features to automatically classify the sentiment expressed in text. Machine Learning (ML) approach for sentiment analysis can be applied in supervised, unsupervised, or semi-supervised settings, depending on the availability of labeled data and the specific requirements of the task [10].

- **Supervised Learning:**

In supervised learning, machine learning models are trained using labeled data, in which each text sample is paired with a sentiment label (e.g., positive, negative, neutral). The models learn to recognize patterns and relationships between the text features and the corresponding sentiment labels during the training phase. Once trained, the models can classify new, unseen text samples based on the learned patterns. Supervised learning is effective when a sufficient amount of labeled training data is available, and it allows for accurate sentiment classification [35]. However, supervised models may struggle with generalizing to unseen or out-of-domain data if the training data is not representative of all possible variations or contexts encountered during inference [36], [37].

- **Unsupervised Learning:**

Unsupervised learning approaches do not rely on labeled data for sentiment analysis. They aim to discover patterns or clusters in the data without prior knowledge of sentiment labels. Unsupervised learning techniques, such as clustering algorithms or topic modeling, group similar texts based on their content or context. By examining the generated clusters or topics, sentiment can be inferred [38]. Unsupervised learning is useful when labeled data is unavailable or scarce, but it may yield less accurate results compared to supervised approaches. Moreover, unsupervised approaches heavily rely on the assumptions and algorithms used for clustering or topic modeling, which can introduce subjectivity and bias into the sentiment analysis results. Further, without labeled data, it is challenging to evaluate the accuracy of unsupervised sentiment analysis, making it difficult to quantify the performance or compare it against other methods [39].

- **Semi-Supervised Learning:**

Semi-supervised learning integrates aspects of both supervised and unsupervised learning methodologies. In this setting, a small portion of the data is labeled, and the majority remains unlabeled. ML models are initially trained on the labeled data using supervised learning techniques. To further improve their performance, these models leverage the unlabeled data. Unsupervised pre-training is one approach where models learn useful representations or features from the unlabeled data, which are then fine-tuned using the labeled data [40]. Semi-supervised learning proves advantageous when labeled data is limited, but unlabeled data is abundant, allowing for enhanced sentiment analysis models [41]. However, it introduces additional complexity, including steps like unsupervised pre-training or co-training, compared to traditional supervised learning [42], [43]. The quality of the unlabeled data also plays a crucial role; noisy or irrelevant unlabeled data can introduce errors or biases into the sentiment analysis models [44].

2) Lexicon-Based Approach

Lexicon-based techniques rely on sentiment lexicons or dictionaries that assign sentiment scores or labels to words. Each word is linked to a sentiment polarity (positive, negative, or neutral) or a sentiment intensity score. The sentiment of a text is determined by aggregating the sentiment scores of the individual words present in the text. The feasibility of this approach depends on the coverage and quality of the sentiment lexicon used. The lexicon should include a wide range of words and phrases, including domain-specific terms, to accurately capture the sentiment in different contexts. Building and maintaining comprehensive and up-to-date sentiment lexicons can be a challenge, especially for niche domains or specialized languages [45]. Moreover, this approach may face challenges in

handling words or phrases with multiple sentiment orientations or ambiguous sentiments. Some words may have different sentiment associations depending on the context [46]. For example, the word "فظيع" in Modern Standard Arabic (MSA) can imply a negative sentiment, meaning extremely awful or terrible. However, in Dialect Arabic (DA), the same word "فظيع" can imply a positive sentiment, meaning extremely good or fantastic. Moreover, the word "صدم" in different contexts can carry opposite sentiments. In the context of "هذا الفنان صدم المتابعين بروعة أدائه," the word "صدم" has a positive sentiment, meaning the artist impressed the audience with the excellence of his performance. While in the context of "السائق صدم بسيارته الحائط," the word "صدم" has a negative sentiment, meaning the driver crashed his car into the wall. Addressing sentiment ambiguity requires additional contextual analysis techniques [47] or heuristics [48]. While Lexicon-based approaches have their limitations, their advantages make them valuable tools in sentiment analysis tasks. Lexicon-based approaches provide transparent and interpretable results. The sentiment lexicons used in this approach assign sentiment scores or labels to individual words or phrases, making it easy to understand how sentiment is determined [49]. This interpretability is particularly useful in applications where explainability is important, such as in legal or regulatory contexts [50]. Further, they are relatively straightforward to implement. Once a sentiment lexicon is available, sentiment analysis can be performed by matching words or phrases in the text with the lexicon entries. This simplicity makes Lexicon-based approaches quick to deploy and well-suited for applications where real-time or near real-time sentiment analysis is required [51], [52]. In addition, they do not heavily rely on large amounts of labeled training data. Instead, they utilize sentiment lexicons that contain pre-defined sentiment information. This reduces the need for extensive manual

annotation or resource-intensive training processes, making Lexicon-based approaches applicable in scenarios where labeled training data is scarce or expensive to obtain [53].

3) Hybrid Approach

Hybrid approaches combine multiple techniques to leverage their strengths. For example, combining lexicon-based approaches with machine learning algorithms can allow for more flexible and adaptive sentiment analysis. Hybrid models may utilize lexicon-based techniques for specific linguistic patterns or domain-specific rules while employing machine learning algorithms for overall sentiment classification. By leveraging the complementary strengths of multiple approaches, hybrid models can achieve higher accuracy, better handle sentiment ambiguity, and adapt to specific domains [37], [54], [55]. On the other hand, the hybrid approach may also introduce additional complexity, require more computational resources, and need careful integration and optimization of different components to achieve optimal performance [56].

3.1.4 Literature Review on Arabic Sentiment Analysis for Online Products

The literature on Arabic sentiment analysis encompasses numerous attempts that employ a diverse range of approaches in different domains, including customer feedback analysis, social media analysis, and political and social contexts. This section highlights significant studies that have contributed to the field of Arabic sentiment analysis for online products. Table 1 provides a comprehensive review highlighting these diverse studies. It summarizes the studies based on their utilized dataset, lexicons, approach, sentiment analysis level, pre-processing steps, features, and classification method.

Generally, the studies utilized datasets with different sizes ranging from 250 to 300K reviews, showcasing the scalability of these analyses. From the table, we can see that three public datasets for online products were utilized in these studies including LABR [57], BRAD [58], and PROD [59], and these datasets worked as benchmarks for Arabic sentiment analysis. Both LABR and BRAD are datasets for only book reviews. LABR stands for Large Scale Arabic Book Reviews, it is a huge Arabic sentiment analysis resource that consists of more than 63K reviews written in both MSA and Classical Arabic (CA). The reviews in this dataset were collected from the GoodReads.com website and each review in this dataset is associated with its rating score. BRAD, Books Reviews in Arabic Dataset, is an extension of LABR that contains more than 500K reviews. While, PROD, Product Reviews, is a dataset for different product types. PROD contains more than 15K reviews, written in both MSA and CA, extracted from Souq.com, and all the reviews are linked with their rating scores. However, these datasets are annotated based on their ratings. Likewise, most of the studies relied also on the rating in annotating their datasets with the sentiments. Relying on the rating to label the reviews is not always accurate as some reviews in LABR found by [60] had conflicting ratings and sentiments. Moreover, the study [61] demonstrated that consistent expansion of the lexical dictionary and increasing data size positively influenced the classification performance.

For the sentiment analysis approaches, the ML approach was the most utilized method in Arabic sentiment analysis for online products [60], [62], [63], [64], [65], [66], [67]. Utilizing DL models outperformed ML models when tested on the same datasets even if it was balanced or imbalanced [66], however, ML models exhibited sensitivity to imbalanced datasets [60], [62]. When the distribution of classes was uneven, the ML model struggled

to accurately predict the minority class, leading to biased results. As a comparison between ML models, the authors of [63] revealed that employing two classification algorithms, namely SVM and NB, yielded superior performance compared to K-Nearest Neighbors (KNN) on the identical dataset. Furthermore, the researchers of [62] utilized different hierarchical classification structures, flat, 2-level, and 4-level, to perform multi-way sentiment analysis using four core classifiers, DT, NB, SVM, and KNN. The flat structure uses a single classifier to distinguish the five classes. The 2-level structure first divides the reviews into negative, neutral, and positive categories, then further splits the negative and positive categories into weak and strong ones. The 4-level structure uses a top-down approach to separate each class from the others in a binary fashion. They show that hierarchical classifiers give significant improvements over flat classifiers in both accuracy and Mean Square Error (MSE). They also showed that KNN benefits the most from increasing the hierarchical structure levels. On the contrary, the authors of [65] introduced a Dynamic Local Memory Neural Network (DLMNN) model for sentiment analysis (SA) and an Interval Type-2 Adaptive Neuro-Fuzzy Inference System (IANFIS) model for predicting future online product trends. They classified the dataset into 3 main categories, content-based (CB), grade-based (GB), and collaboration-based (CLB). DLMNN analyzed reviews in each category, classifying them based on their sentiments. Meanwhile, IANFIS utilized weighting factors and classification techniques for predicting future product trends. The experimental outcomes demonstrated that the proposed models surpassed existing methods, delivering enhanced accuracy and efficiency in both sentiment analysis and future product prediction tasks. Study [68] proposed an approach that works at feature-level SA. They combined their approach with ontology information to help extract explicit

product features from the review and determine the important features, then generate a feature-based summary. Their proposed approach outperformed ML models and dictionary and ontology baselines. Few studies utilized a lexicon-based approach for Arabic sentiment analysis for online products [69], [70]. These studies showed promising performance of lexicon-based approach in online products. The hybrid approach, integrating different SA techniques, was commonly employed, and it enhanced the performance of the classification [71].

Studies addressed different levels of analysis, from document-based, sentence-based, and aspects-based sentiment classification. Robust preprocessing, including tokenization, normalization, stemming, and stop-word removal, was consistently applied. Many papers demonstrated that text Pre-processing is a key factor in sentiment analysis classification [63], [66], [67], [72]. The findings reported in [66] demonstrated that the suggested pre-trained model, Arabic Extra-Long Neural Network (AraXLNet), coupled with the FARASA segmenter, yielded accuracy rates of 94.78%, 93.01%, and 85.77% in the sentiment analysis task for Arabic across various benchmark datasets, including AraSenTi [73], Arabic Sentiment Tweets Dataset (ASTD) [74], and SemEval [75]. Notably, AraXLNet outperformed Arabic Bidirectional Encoder Representations from Transformers (AraBERT), which achieved accuracy scores of 84.65%, 92.13%, and 85.05% on the corresponding datasets.

Research conducted by [64] indicated that optimal outcomes were attained by combining BOW with word clustering. This approach yielded F1 scores within the range of 80% to 92% for polarity classification and 63% to 96% for subjectivity classification, contingent upon the specific domain under consideration. Also, they noticed that utilizing word

clustering alone yielded suboptimal results, except in certain domains and cluster sizes. However, the incorporation of word clustering proved beneficial in capturing both the similarity and sentiment orientation of words. The researchers in [63] demonstrated that the performance of NB improves when employing the N-Gram feature. Conversely, the use of the same feature led to a decline in performance when applied to SVM. The study presented in [67] demonstrated that fastText word embedding exhibits superior performance compared to Word to Vector (Word2Vec) word embedding. In [76], various machine learning techniques were employed, utilizing multiple features to emphasize the importance of merging and pruning terms. The study revealed an accuracy improvement by merging terms after pruning. Additionally, it highlighted the selection of features dependent on lexicon-based sentiment analysis without incorporating any other features. These findings collectively underscore the diverse and effective strategies employed in Arabic sentiment analysis, paving the way for advancements in understanding consumer sentiments in online product reviews.

Table 1: Previous studies on Arabic Sentiment Analysis for Online Products.

Ref	Year	Domain	Dataset Name	Data Size	Language	Annotation	Lexicons	Approach	Level	Pre-processing	Features	Classification
[76]	2014	Books	LABR	4000 reviews	MSA, CA	Using the ratings	LABR lexicons	ML-based	Document	Removing stop words, negations, and singletons; Eliminating five proclitic	Lexicon-based Weights, Unigrams, Empirical Features	NB, SVM, KNN
[63]	2016	TV, Phone, Tablet, PC, Camera, Notebook	-	250 documents	CA	Manually	-	ML-based	Document	Normalization; Segmentation; Removing stop words, non-Arabic words, and special characters; Converting emoticons; Correcting the elongated words; Stemming	N-grams	NB, SVM, and KNN.
[62]	2016	Books	LABR	63K reviews	MSA, CA	Using the ratings	-	ML-based	Sentence	Removing newline characters, HTML tags, hyperlinks, repeated dots, non-Arabic characters, and some special Unicode characters	BoW	NB, SVM, DT, and KNN with three classification structures, flat, 2-level, and 4-level.
[69]	2016	Products	-	200 reviews	CA	Manually and Using the Ratings	Corpus-based lexicons	Lexicon-based	Sentence	Normalization; Stemming; Removing stop words; POS tagging	Nouns, Noun Phrases	-
[60]	2016	Books	LABR	63K reviews	MSA, CA	Using the ratings	-	ML-based	Document	Tokenization; Removing stop words; Stemming	BoW, N-grams	LR
[72]	2016	Muasher products	-	1,000 tweets	MSA, CA	Manually	-	Hybrid	Sentence	Tokenization; Normalization; Removing stop words; Stemming	TF-IDF, BTO, N-Grams	NB, SVM
[70]	2016	Products	-	500 reviews for movies, hotels, restaurants, and products	MSA	-	ARBL lexicons	Lexicon-based	Aspect	-	Roots, Patterns, Opinion Words, Intensifiers, Negations, Syntactic rules	-
[64]	2017	Products	-	6268 documents on online products and other domains	MSA, CA	Manually	-	ML-based	Document and Sentence	Normalization; Removing stop words; Stemming	BOW, Word Clustering	SVM

[68]	2017	Books	LABR	2000 book reviews	MSA, CA	Using the ratings	Arabic ontology for hotels and books domains	Hybrid	Sentence	Tokenization; Removing stop words; Stemming; POS tagging	Ontology-based Features	NB, SVM, KNN, DT, Ontology baseline, Ontology with important feature
[71]	2018	Product	PROD	3962 reviews	CA	Using the ratings	-	Hybrid	Document	Tokenization; Removing stop words; Stemming; N-grams extraction	Word Frequency, Polarity Scores	NB, MLP
[65]	2020	Food	-	5000 reviews	-	Using the ratings	SentiWordNet	ML-based	Sentence	Tokenization; Stemming; Removing stop words, hashtags, and usernames	Count of Positive and Negative Emoticons, Exclamation Marks, Question Marks, Positive and Negative Gazetteer Words Occurrences, N-Grams, POS	DLMNN, and IANFIS on three scenarios (GB, CB, and CLB)
[66]	2021	Products	-	11K tweets	CA	Manually	-	ML-based	Sentence	Tokenization; Stemming	BoW	DL, DT, NB, KNN, SVM
[77]	2022	Books	BRAD	Pretraining: 60 M sentences. Fine-tuning: 17 K tweets.	MSA, CA	Using the ratings	-	ML-based	Sentence	Segmentation; Tokenization	Word Embeddings	AraXLNet
[61]	2022	Sustainable Products	-	300 K reviews	CA	Using the ratings	Emotional terminology library	Hybrid	Sentence	Segmentation; Removing stop words and non-Arabic characters; Word Vector	Lexical Score, Word Vector	Fully connected layer with Softmax classifier
[67]	2023	Books	LABR	63K reviews	MSA, CA	Using the ratings	-	ML-based	Sentence	Removing English and Arabic punctuations, diacritic, repeating characters, stop words, non-Arabic characters, and digits; Normalization; Tokenization; Stemming	fastText, Word2Vec	LSTM, CNN, and CNN-LSTM (binary and multi-classification)

3.2 Fake Reviews Detection

During the last decade, countless studies have been conducted to detect fake reviews in different domains using different techniques. Most of these studies focused on the detection of fake reviews in English text with little work being done for fake reviews in Arabic text. This section presents a summary of previous attempts at detecting fake reviews in both English and Arabic languages.

3.2.1 Targeted Fake Reviews and Scope

The targeted fake reviews vary from study to study based on the different utilized scopes and perspectives used. The current studies that focused on fake reviews detection in English utilized the following scopes [78]:

- **Content Similarity Footprints (CSF):**

Fake reviews can be detected by identifying non-reviews, duplicate or near duplicate reviews using different similarity measures. Studies [7], [79], and [80], suggested that non-reviews exhibit high levels of pairwise similarity. Jindal and Liu [7] proposed an approach to detect duplicate and near duplicate reviews by calculating the Jaccard similarity, which is the ratio of the length of the intersection of the 2-grams of the two reviews to the length of the union of their 2-grams. The similarity score of at least 90% between two reviews represents duplicate or near duplicate reviews. The authors of [81] claimed that spammers utilized conceptually similar words in their reviews which can help in detecting duplicate and near duplicate reviews by using conceptual-level similarity measures.

- **Writing Style Footprints (WSF):**

Exploring the writing footprints of the reviewers can help in detecting spammers and therefore fake reviews. Spammers and real customers have relatively unique writing styles and use specific words at different frequencies [78]. It is assumed by many studies that spammers often write fake reviews without even experiencing the product or without knowing the detailed information about the product. Therefore, fake reviews are assumed to be less diverse [82] and contain false information about the product [83]. In contrast, other studies support the opposite hypothesis, which states that fake reviews provide more details than non-fake reviews [78]. To figure out the writing footprints of the reviewers, researchers analyzed the specificity of the information about the product provided in the review [84], [85], [86]. Others analyzed the lexical diversity in the review to track the proportions of different POS tags, and the frequency of temporal and spatial words in the review [78]. On another hand, a study [83] utilized profile compatibility features to represent the level of compatibility between the product information and the personal experience provided in the review [83]. Mukherjee et al. [87] showed that spammers never provide tips in the reviews so that the number of tips provided in the reviews could be used to detect spammers. Furthermore, various studies have posited the hypothesis that individuals engaging in spamming activities may inadvertently leave linguistic clues when crafting non-reviews. These discernible cues can be identified and leveraged through diverse feature categories, including syntactical features (such as the percentage of verbs, nouns, and pronouns in the review), lexical features (encompassing sentiment and text orientation, content, and lexical diversity), and stylistic features (including capitalization, repetition, and emotiveness diversity) [88], [89].

- **Rating Behavior Footprints (RBF):**

Examining patterns in how reviewers assign ratings, such as consistently giving extreme ratings (very low or very high), can indicate potential bias or deception. Exploring the set of reviews written by a user can reveal his/ her intent. Moreover, due to the presence of several genuine negative reviews, spammers may be compelled to compose excessively positive reviews to mitigate the impact of the previous negative ratings. Also, spammers try to make the reviews look natural by posing extremely positive singleton reviews from different accounts. Therefore, the spread of singleton reviews for a product and the rating spikes can be clear indications of spammer activities [78]. A list of eight potential spam indicators based on ratings was suggested by Wu et al. [90]. Among these indicators, truncated rating (the huge difference in average rating after removing the rating of very positive consecutive reviews) and reactive positive singletons (the duration between positive singletons and genuine negative reviews) were the strongest indicators. Further, Rajshri et al. [91] considered providing only a star rating without a review text as deceptive behavior by spammers.

- **Temporal Footprints (TF):**

Analyzing the timing and frequency of review submissions can help detect unusual bursts of activity, potentially indicating coordinated efforts to manipulate the review system. Mukherjee et al. [92] identified spammers by observing behaviors such as posting many reviews in a single day and submitting multiple ratings/reviews for a single product. Similarly, Similarly, Savage et al. [93] considered the posting of numerous reviews in a single day .as a key indicator of spammers. Wu et al. [94] confirmed that fake reviews can deceive and alter the popularity ranking of hotels. They quantified this distortion by comparing rankings before and after the deletion of suspected fake reviews, using two

indicators: the proportion of positive singletons (the ratio of the positive singleton reviews to the total number of reviews for a hotel) and the concentration of positive singletons (the degree of time-based clustering of positive singletons reviews for a hotel).

- **Product-Centric Footprints (PCF):**

Product-centric footprints in fake review detection involve examining features and patterns related to the products being reviewed. These footprints can provide valuable insights into the authenticity of reviews and the potential presence of fake or deceptive reviews. A group of spammers can be hired to promote low-quality or low-demand products to artificial boost sales. Therefore, utilizing product-centric features is effective for identifying target products or likely target products by spammers and, consequently, identifying fake reviews. In a study conducted by Li et al. [95], it was discovered that the characteristics of product descriptions were correlated with the occurrence of fake reviews for a particular product. Additionally, knowing that spammers who target a product in an organized manner are probably hired as a group. So, another way to detect fake reviews is by finding the home base where the spammers are hired. The hiring of spammers is commonly accomplished through crowdsourcing platforms. Fayazi et al. [96] proposed a sampling method to identify targeted products and an initial set of fake reviewers recruited through these crowdsourcing platforms. Kumar et al. [97] proposed three quality scores: goodness of a product, fairness of a user, and reliability of a rating, and they formulated six axioms to establish the inter-dependency between these scores. The axiom "higher quality products receive more reliable positive ratings" and the axiom "reliable ratings are given by fairer users" are among these axioms. Wang et al. [98] proposed eleven relations that were automatically generated from two discovered patterns between products and reviewers. In

their approach, they represented the reviews by concatenating the review text, the embeddings of the reviewed product, and the embeddings of the reviewer. Based on their proposed representations, the classifier identified the fake reviews more precisely. Lu et al. [99] designed a factor graph model that covers three entities: the reviews, reviewers, and products. This model simultaneously identifies fake reviews and review spammers.

- **Reviewer Profile Footprints (RPF):**

Exploring rating footprints by utilizing meta-data features can help identify spammer group and, consequently, fake reviews. Examining the completeness and consistency of reviewer profiles, including factors like profile age, number of reviews, and other metadata, can provide insights into reviewer authenticity. Many researchers considered meta-data features to identify fake reviews [100], [101], [102], [103]. Meta-data features are computationally inexpensive to analyze and easy to obtain. Moreover, the meta-data-based models are domain-independent which can be generalized and implemented for different platforms [78].

- **Collusion Footprints (CF):**

Identifying coordinated efforts among multiple individuals or accounts to submit fake or deceptive reviews may involve shared objectives, similar review patterns, or interactions among colluding accounts. Exploring patterns in the review text, reviewer behavior, and products can help identify spammers and consequently, fake reviews. Mukherjee et al. [104] introduced a set of features or indicators to identify spammer groups. These indicators are: group deviation, time window, member content similarity, group content similarity, early time frame, group size, ratio of group size, and support count. Allahbakhsh et al. [105] proposed a model based on frequent itemset mining techniques to identify

collusion groups. They also introduced several features to estimate the extent of damage that may be caused by these collusion groups. Wang et al. [94] proposed a framework to detect spammer collusion groups through graphs that capture the relationships among users to reveal the co-review collusiveness.

3.2.2 Datasets

The construction of review datasets for fake review detection involves gathering and preparing the data that will be used in the fake review detection system. There are several methods utilized in constructing such datasets:

- **Public Datasets:** There are several datasets available on the web for fake review detection of different domains which are collected or created by companies or researchers and can be downloaded directly[7], [79], [106].
- **Crowdsourcing Platforms:** Utilizing platforms like Amazon Mechanical Turk [22] or CrowdFlower to have human workers generate reviews, both genuine and fake, under controlled conditions. Several studies utilized this approach in constructing their datasets for fake reviews [107], [108], [109].
- **Manual Annotation:** Manually curating datasets by experts who identify and label reviews as genuine or fake based on specific criteria [95], [110].
- **Filtering Algorithm:** Using user-item interaction data from online platforms to infer the likelihood of a review being fake based on patterns of user behavior. Yelp [79] is the most popular online review platform which implements automated systems to detect and filter out fake or suspicious reviews and has been utilized by several studies [87], [111], [112].

- **Supervised Learning from Scratch:** This method starts with a small seed dataset and uses it to train a fake review detection classifier. As the classifier becomes more accurate, it can be used to label additional data and iteratively improve the model.
- **Transfer Learning:** Leveraging pre-trained fake review detection models and fine-tuning them on a specific domain or dataset. This approach requires a smaller amount of labeled data for fine-tuning.

Table 2 illustrates some of the utilized datasets for fake reviews from previous studies.

Table 3 represents the publicly available and annotated datasets for fake reviews detection.

Table 2: The utilized datasets for fake reviews from previous studies.

Ref	Year	Source	Domain	Volume
[95]	2011	Epinions	Online Products	6,000
[113]	2011	ResellerRatings	Stores	408 K
[84]	2011	Tripadvisor	Hotels	800
[95]	2011	Opinions	Hotels	6,000
[114]	2012	Tripadvisor	Hotels	27 K
[115]	2012	Amazon	Online Products	109 K
[79]	2013	Yelp	Hotels and Restaurants	-
[108]	2013	Tripadvisor	Hotels	1,600
[109]	2013	Tripadvisor	Hotels	2,848
[116]	2013	ResellerRatings	Stores	628 K
[117]	2013	Datatang	Online Products	10,020
[118]	2013	Datatang	Hotels and Restaurants	493 K
[119]	2013	Amazon	Online Products	65 K
[99]	2013	Amazon	Online Products	195 K
[79]	2013	Yelp	Hotels and Restaurants	67 K
[107]	2014	Tripadvisor	Hotels	3,032
[120]	2014	Amazon	Online Products	3 M

[121]	2014	Amazon	Online Products	6,819
[122]	2014	Tripadvisor	Hotels	3000
[112]	2014	Yelp	Hotels and Restaurants	359 K
[112]	2014	Yelp	Hotels and Restaurants	608 K
[7]	2015	Amazon	Books, Music, DVD	5.8 M
[123]	2015	Amazon	Online Products	542 K
[100]	2015	Datatang	Hotels and Restaurants	9,765
[124]	2018	Amazon	Online Products	21 K
[125]	2019	Yelp	-	18 K
[106]	2020	Amazon	Online Products	26 M

Table 3: The publicly available and annotated datasets for fake reviews detection.

Ref	Year	Dataset	Domain	Labeling Approach	Volume	Link
[79]	2013	Yelp	Hotels and Restaurants	Yelp Filter	67 K	Here
[108]	2013	DOSC	Hotels	Crowdsourcing	1,600	Here
[121]	2014	Amazon	Online Products (Books)	Rule-Based	6,819	Here
[112]	2014	Datatang	Hotels and Restaurants	Yelp Filter	9,765	Here
[100]	2015	YelpZip	Hotels and Restaurants	Yelp Filter	608 K	Here
[100]	2015	YelpCHI	Hotels and Restaurants	Yelp Filter	67 K	Here
[100]	2015	YelpNYC	Hotels and Restaurants	Yelp Filter	395 K	Here
[124]	2018	Amazon	Online Products	Amazon Filter	21 K	Here
[125]	2019	Yelp	Electronics Retailers	Yelp Filter	30 K	Here
[106]	2020	Amazon	Online Products	Manual	26 M	Here

3.2.3 Feature Extraction

The features used in fake review detection can be categorized into two main categories: behavioral features and content-based features.

1) Behavioral features:

Behavioral features refer to patterns of behavior exhibited by reviewers that may indicate the presence of fake or suspicious reviews. These features are derived from the actions and interactions of reviewers on online review platforms. Here are some common behavioral features used in fake review detection:

- **Review Velocity:**
The rate at which a reviewer posts reviews [126]. Unusual high review velocity, especially for a new account, may be indicative of fake reviews.
- **Consistency in Ratings:**
Whether a reviewer consistently gives similar ratings across different reviews, or if there are significant deviations [91], [126].
- **Rating Deviation:**
It refers to the difference between the rating given in a specific review and the average rating for the same product [127].
- **Burstiness:**
The phenomenon where a sudden and abnormal surge in the number of reviews occurs within a short period of time for a particular product [106].
- **Analytical thinking of the reviewer:**
It involves evaluating reviewer's ability to provide detailed and reasoned assessments [126].
- **Count of posted tips:**
The count of tips or advices shared by a reviewer on a platform [125]. It provides insights into the reviewer's level of engagement and activity on a platform.
- **Early Review Ratio:**

The proportion of reviews that are submitted shortly after a product or service becomes available for review [127].

- **Reviewing Range:**

The diversity of product categories, brands, or stores that a reviewer covers [128].

A very narrow or very broad range may indicate suspicious behavior.

2) Content-based features

Content-based features refer to characteristics or attributes of the text content within a review that can be analyzed to determine the likelihood of the review being fake or deceptive. These features focus on the actual language, style, and content of the review itself. Here are some common content-based features used in fake review detection:

- **Review Length:**

The number of words or characters in the review. Fake reviews may be excessively short or unusually long [91], [126].

- **Grammar and Spelling:**

The quality of language usage, including grammar, spelling, and punctuation [129].

Fake reviews may exhibit poor language skills.

- **Content similarity [106]:**

The percentage of negation words in the review text [126] and the ratio of comparing words [126].

- **Use of Emoticons and Symbols:**

The presence of excessive emoticons, emojis, or special symbols, may indicate fake or overly enthusiastic reviews [129].

- **Specific Product Details:**

The inclusion of specific details about the product or service being reviewed [91].

Fake reviews may lack specific information or use generic descriptions.

- **Sentiment Consistency:**
The alignment of sentiment expressed in the review with the overall rating given [126]. Inconsistent sentiment may raise suspicion.
- **Use of First-Person Perspective:**
The frequency of first-person pronouns (e.g., "I," "me," "my") in the review, may provide insight into the reviewer's personal experience [126].
- **POS Frequency:**
The distribution and frequency of different parts of speech (such as nouns, verbs, adjectives, etc.) within the text [126].
- **Sentiment Intensity:**
The degree or strength of the sentiment expressed in the text [126]. It assesses how strongly positive, negative, or neutral the sentiments conveyed in the review are.
- **Sentiment Vocabulary Ratio:**
The ratio of positive and negative sentiment words to the total number of words in the review [126], [127], [126].
- **Review Duration:**
The difference between the first and last posted reviews of the reviewer [130].
- **Word embeddings:**
Word embeddings capture semantic relationships between words [131].
- **N-GRAM:**
N-Grams, especially bigrams (2-grams) and trigrams (3-grams), capture sequences of words. Analyzing these sequences can help identify specific language patterns that indicate fake reviews [131]. Unusual or rare phrases not commonly used in genuine reviews can be flagged using N-grams.

Studies have showed that considering both the behavioral and the content features can increase the accuracy of the detection [80], [132], [79], [106]. Moreover, Mukherjee et al. [115] found that behavioral features are more efficient than linguistic features. Based on their findings, both behavioral and content features are crucial for improving the performance of fake review detection models.

3.2.4 Fake Review Detection Methods

The classification methods of fake reviews can be divided into two types: traditional machine learning, and deep learning methods. Machine learning models play a vital role in fake review detection and can be categorized into three classes: supervised, unsupervised, and semi-supervised learning. Most of the studies on fake review detection in English have been based on supervised machine learning [125], [131], [133]. However, detecting fake reviews using supervised machine learning requires labeled data, which can be challenging to acquire, especially if fake reviews are domain-specific.

Previous studies have shown that utilizing ensemble models in fake review detection provides higher accuracy compared to the individual models [7], [127], [134], [135], [136]. Additionally, stacking ensemble classifiers have outperformed majority voting ensemble classifiers in terms of accuracy [136], [137], [138]. The performance of machine learning models also varies based on the selected features when tested on the same dataset [139].

To overcome the difficulty of obtaining labeled datasets in supervised learning, many studies have utilized unsupervised learning [1], [140], [141], or semi-supervised learning [126], [142], [143], as these methods do not require labeled datasets.

On the other hand, deep learning models have proven their power in many classification tasks within the natural language processing field. They can automatically and quickly perform feature extraction compared to traditional machine learning models and can capture the semantic meanings of the text using word embeddings. Recent studies have focused on utilizing deep learning models for fake review detection [127], [144]. The most utilized deep learning models for fake review detection in English were Convolutional Neural Networks (CNN) [144], [145], [146] and LSTM [147], [148], [149].

Despite the effectiveness of deep learning models, they encountered several issues. First, they require large datasets to perform well compared to traditional machine learning models. Second, they are often unexplainable, meaning they cannot provide declarative knowledge to explain the results. Third, they exhausted extensive computational resources.

3.2.5 Literature Review on Fake Review Detection Using Sentiment Analysis

This section provides an overview of studies focusing on the detection of fake reviews through sentiment analysis across various domains. Table 4 summarizes these studies, detailing the domain, dataset, lexicon, sentiment analysis approach, sentiment-based content and behavioral features, and the classification methods used in each study.

In 2018, a study in the hotel domain [156] for, utilized 1,600 labeled reviews with sentiment and fake labels. They employed sentiment scores along with LIWC, POS, and N-gram features with classification algorithms such as DT, NB, SVM, K-NN, RF, and LR to detect fake reviews. LR demonstrates superior performance compared to other classifiers, particularly when combining unigrams with LIWC sentiment scores.

Another study in the hotel domain in 2020 [131] collected 10,000 reviews from iranhotelonline.com and egardesh.com. The reviews underwent manual checks for spam detection after being labeled as spam based on matching certain features. The study employed a lexicon-based approach, creating a sentimental vocabulary using ADJ and ADV tags, supplemented by online dictionaries for synonyms and antonyms, and manually adding words to end up with 4,570 lexicons. The study computed several lexical, syntactic, behavioral, and semantic features. Sentiment-based features included sentiment score, count of sentiment words, the ratio of sentiment words to all review words, and the ratio of sentiment to overall product rating. Multiple classifiers such as DT, RF, NB, Rule Induction, SVM, NN, KNN, AdaBoost, and Bagging were employed, with DT and AdaBoost achieving the highest accuracy at 98.67% and 98.00%, respectively. The study highlighted the robustness of their extended feature set compared to previous studies using the same classifiers.

In another study in 2020 [150], spanning hotel, restaurant, doctor, and online products domains, the authors utilized four benchmark datasets, including hotel, restaurant, doctor, and Amazon datasets with 1,600, 400, 556, and 21,000 reviews, respectively. They incorporated lexicon-based sentiment analysis using various lexicons such as OpinionFinder, Bing Liu's, AFINN, S140, and SentiWordNet. The study applied classifiers such as SVM, AdaBoost, NB, KNN, RF, and Bagging as baseline models and CNN and DFFNN as their proposed models, incorporating feature sets that included the count of sentiment words in the review along with N-grams and Skip-Gram word embeddings. The reported experimental results demonstrated that their proposed models outperformed existing baseline approaches in terms of accuracy, AUC, and F-score.

The authors of [159] selected 1,971 Tech product reviews from the Amazon dataset, labeling reviews as fake if they exhibited a high degree of similarity and shared more than 3 significant words. They employed the AFINN lexicon for sentiment analysis on the reviews and applied clustering based on the inconsistency between the sentiment and the rating of the review to detect fake reviews. Their proposed approach achieved an accuracy of 85.53% in fake review detection.

A study in 2021 [158] built their lexicons by extracting all adjectives and adverbs to form the labeled dataset that contains 630 reviews of Nokia products. After that, the nouns close to these adjectives and adverbs were considered product aspects. Subsequently, using TF-IDF, only aspects with occurrences greater than 1 were selected. For the task of detecting fake reviews, Multinomial NB and Bernoulli NB were employed, considering factors such as the sentiment label, the presence of hyperlinks in the review, and whether the review exhibited advertising, promotional, or bonus characteristics. The outcomes revealed that Bernoulli NB outperformed Multinomial NB, achieving an accuracy of 96.7% compared to 94.5%.

The study referenced [1] investigated various categories of online products, including clothing, accessories, electronics, and beauty items. It utilized sentiment labels corresponding to predefined topics to analyze the dataset. After enriching the dataset with sentiment labels for each topic, the researchers applied the K-means algorithm for classification. Comparative experiments, including FIM, LSA, and LDA, demonstrated the superiority of their Grouped Spam Detection approach based on Nominated Topics (GSDNT) in identifying fake reviews. The investigator in [127] curated a dataset comprising 24 K labeled reviews from the Yelp dataset, distinguishing between fake and

authentic reviews. Subsequently, lexicon-based sentiment analysis, utilizing SentiWordNet 3.0, was employed to assign sentiment labels to the reviews. A huge number of lexical, stylistic, syntactic, behavioral, and semantic features were then extracted and computed. The proposed semantic features were the inconsistency between the rating and sentiment of the review, the sentiment inconsistency between the review and other reviews with the same star rating on the same product, and the inconsistency between the review and other reviews with the same star rating from the same reviewer. The findings demonstrated that the inclusion of review inconsistency features significantly enhances fake review detection performance, irrespective of the specific classification algorithms employed.

Table 4: Previous studies on Fake Review Detection using Sentiment Analysis.

Ref	Year	Domain	Dataset	Lexicons	Sentiment Analysis Approach	Sentiment-Based Content Features	Sentiment-Based Behavioral Features	Classification
[151]	2018	Hotel	1,600 reviews labeled with the sentiments	-	-	Sentiment Score	-	DT, Naive Bayes, SVM, K-NN, RF, and LR
[152]	2020	Hotel	10 K reviews	Corpus-based	Lexicon-based	Sentiment Score; Count of sentiment words; Ratio of sentiment words to all review words, Sentiment to overall product rating consistency.	-	DT, RF, NB, Rule Induction, SVM, NN, KNN, AdaBoost, and Bagging.
[153]	2020	Hotel, Restaurant, Doctor, Online products.	Hotel:1,600 reviews, Restaurant: 400 reviews, Doctor: 556 reviews, Amazon: 21 K reviews	OpinionFinder lexicon, Bing Liu's, AFINN, S140, SentiWordNet, NRC Hashtag, Emoticons.	Lexicon-based	Count of sentiment words	-	SVM, AdaBoost, Naive Bayes, KNN, RF, Bagging, CNN, DFFNN.
[154]	2020	Online Products: Tech products	1,971 reviews	AFINN	Lexicon-based, ML-based	Sentiment to review rating inconsistency	-	Clustering
[155]	2021	Nokia products	630 reviews	Corpus-based	Lexicon-based	Sentiment label	-	Multinomial NB and Bernoulli NB.
[1]	2021	Online Products: Clothes, Purses, Shoes, Cell Phones, Laptops, Solid-state Disks, Canteens, Makeup.	44 K reviews labeled with the sentiments	-	-	Sentiment label for nominated topic list	-	K-means
[156]	2021	Restaurants	24 K reviews	SentiWordNet 3.0	Lexicon-based	Sentiment Score; Ratio of sentiment words in a review; Sentiment to review rating inconsistency; Content inconsistency	Language inconsistency	SVM, NB, CART, RF, MLPNN.

3.2.6 Literature Review on Fake Review Detection in Arabic

While research on detecting fake Arabic reviews is limited, we've categorized existing studies based on their approach (scope, datasets, features, and classification methods) as shown in Table 5. Notably, previous studies focused on two main criteria for identifying fake Arabic reviews: content similarity and writing footprints.

The construction method utilized in collecting the Arabic review datasets was manual by humans [109], [157], [158], [159]. Ahmed and Alaa [109] used their dataset that was extracted from different websites, including TripAdvisor, Booking, and Agoda which resulted in 2,848 Arabic reviews. The reviews were labeled manually using their defined scoring rules. Also, Saeed et al. [160] utilized the Hotel Arabic Reviews Dataset (HARD) [157], which was collected from the Booking website and consists of 94,052 unlabeled Arabic reviews. They used the translated and labeled DOSC dataset [161] as a training set to label the reviews in the HARD dataset using the majority voting ensemble classifier. Wahsheh et al. [158] collected the reviews from Yahoo-Maktoob news with a total of 3,090 Arabic reviews. Najadat et al. [159] collected 3,000 Arabic reviews from Facebook and labeled them manually by experts based on their majority vote into spam and non-spam reviews. Alwayle et al. [162] also utilized the DOSC dataset [161] in their detection system. Additionally, other studies have utilized English fake reviews datasets after translating them into Arabic [160], [163].

For the feature extraction, most of the studies concentrated on one aspect for extracting features: the content of the reviews [158], [159], [160], [162], [163]. This means they

mainly looked at what the reviews said, without considering other factors like the reviewer's behavior or style.

Wahsheh et al. [158] classified Arabic reviews into spam or non-spam based on the presence of a Uniform Resource Locator (URL) in the review. As the inclusion of URLs in reviews can serve as a crucial indicator in fake review detection processes. Since fake reviews often originate from automated bots or paid reviewers, they may include URLs leading to unrelated websites or promotional content. Saeed et al [160] carried out several investigations on different methods of spam review detection. The extraction module extracted the important features that play a key role in classification, it first extracts different combinations of N-gram features and determines their polarity using sentiment lexicons. Then for each extracted N-gram feature, it applies negation handling to handle their polarities. After that, it applies content-based feature extraction to extract some other features like unique word percentage, word count, and review rating deviation.

Ziani et al. [163] developed new semantic features to enhance fake review detection in Arabic. These semantic features were derived from Arabic rhetoric relations and discourse parsing. Their approach starts by first applying automatic pre-processing to the reviews and then extracting both lexical and semantic features. The lexical features include the most selected statistical features from the previous studies. Their proposed semantic features include the explanation, cause, condition, likening, opposition, and difference relations polarity in Arabic.

Najadat et al. [159] proposed an approach to detect Arabic spam reviews based on keywords by calculating the TF-IDF to measure the importance of each word in the review

text. Then they applied different combinations of feature selection methods to select the most important features based on the deviation, correlation, information gain, chi-squared, and uncertainty.

Alwayle et al. [162] concentrated on feature extraction techniques, specifically targeting unigram and bigram models. These methods involve breaking down the text into individual tokens (unigrams) and adjacent token pairs (bigrams). By analyzing these features, they aimed to enhance spam detection.

However, to the best of our knowledge, only one study [109] utilized both behavioral and content features. It proposed an approach to classify Arabic opinion reviews using content-based features with some features about the reviewer and the hotel the review is written for.

Many studies found in the detection of fake reviews in Arabic have utilized traditional machine learning classification [109], [158], [160], [159].

Ahmed and Alaa [109] conducted three classification experiments, using data mining, text mining, and data-text mining on three machine learning models NB, SVM, and KNN. The results showed that data-text mining classifications with over-sample approaches achieved the highest accuracy. For ML classifiers, the NB classifier achieves the highest accuracy with 99.20%.

Wahsheh et al. [158] evaluated their proposed method by applying SVM to the results of the term frequency (TF). The results showed favorable results with an accuracy of 97.50% and an error rate of 2.49%.

Saeed et al [160] carried out several investigations on different methods of spam review detection. They examined four different types of classifiers in the spam detection module, which are rule-based classifiers, machine learning classifiers, majority voting ensemble classifiers, and stacking ensemble classifiers. According to their reported experiment results, the rule-based classifier surpassed machine learning-based classifiers. Moreover, the stacking ensemble classifier which combines rule-based and K-means classifiers surpassed all the classifiers with 95.25% accuracy on the DOSC dataset and 99.98% accuracy on the HARD dataset.

In [159], four classifiers, KNN, DT, NB, and SVM, were trained using the extracted keywords. For the evaluation, they utilized the k-fold cross-validation method. The highest accuracy was achieved using the DT classifier compared with the three other classifiers with a 92.63% accuracy value. For all classifier algorithms, the feature selection method that is based on correlation and deviation provided the best results. Moreover, they observed some properties of the reviews, the longest spam texts had a size of around 50 words, while the size of the longest non-spam texts was around 40 words. Also, most of the non-spam reviews were short.

Ziani et al [163] have utilized semi-supervised SVM (S3VM) was to classify the reviews based on their proposed semantic and lexical features. Considering their proposed semantic features enhanced the accuracy of the system from 89.33% (using the lexical features only in their previous study [164]) to 93% (using both the lexical and proposed semantic features).

Alharthi et al. [165], utilized both machine learning and deep learning models in detecting Arabic fake reviews. The machine learning algorithms they used were: LR, DT, GBM, RF, KNN, and SVM while the deep learning algorithm they used was AraBERT. The results showed that AraBERT exceeded the machine learning models in terms of accuracy with 93%. On the other hand, SVM has performed better than traditional machine learning algorithms with an accuracy of 87.61%. Alwayle et al. [162] proposed a fake reviews detection model called SRD-OSGRU based on GRU architecture. The proposed model was evaluated on the DOSC dataset [161] with a comparison with recent models, stacking ensemble, RB-Boosting, RB-RF model, RB-NN model, RB-Bagging, RB-KNN model, RB-K-Means. SRD-OSGRU model outperformed the recent models with an accuracy of 99.3%.

Table 5: Targeted Arabic fake reviews.

Ref	Year	Domain	Labeling Approach	Scope	Dataset Size	Features	Classification
[109]	2013	Hotels	Manually	CSF WSF	2,848 reviews	Content-based feature Behavioral features Hotel information features	SVM, NB, KNN
[158]	2013	News	Manually	CSF	3,090 reviews	Review with spam URL or not from BLD or ACLWSDS. The review contains ADP. Number of sentiment words and emoticons in the review.	Applying SVM to the results of the term frequency (TF).
[160]	2019	Hotels	Transfer Learning	WSF	HARD: 94,052 reviews DOSC: 1,600 reviews	Different combinations of N-gram features. Applying negation handling to each N-gram feature. Content-based features.	Rule-based, Machine learning, Stacking ensemble, Majority voting ensemble classifiers
[163]	2021	Hotels	Public dataset	WSF	DOSC: 1,600 reviews	Lexical and Semantic features	Semi-supervised SVM (S3VM)
[159]	2021	News	Manually	CSF	3,000 reviews	TF-IDF	DT, KNN, NB SVM.
[165]	2022	Products	Public dataset	WSF	40 K reviews	Automated feature extraction	LR, DT, GBRF, KNN, SVM, AraBERT.
[162]	2023	Hotels	Transfer Learning	WSF	DOSC: 1,600 reviews	Unigram, Bigram	SGRU

3.3 Gap Analysis

This section underscores the existing gaps identified in the research areas of Sentiment Analysis and Fake Reviews Detection. The gaps are categorized into four main points related to targeted scopes, datasets, features for fake review detection, and methods. The following subsections address these gaps.

3.3.1 Targeted Scopes:

- Few scopes, such as CSF and WSF, have been utilized in detecting fake reviews in Arabic. Other scopes require further investigation.

3.3.2 Datasets:

- Lack of publicly labeled datasets with sentiments for online product reviews in Arabic that are accurately labeled without relying on ratings to label the dataset.
- Lack of publicly labeled datasets for reviews in Arabic for fake review detection, accurately labeled without relying on crowdsourcing platforms or manual annotation.
- Lack of publicly labeled datasets for online product reviews in English for fake review detection with a complete set of features (e.g., review time, reviewer ID, product ID, etc.) and accurately labeled without relying on crowdsourcing platforms or manual annotation.

3.3.3 Features for Fake Review Detection:

- Many behavioral features remain unexplored, such as the ratio of posting a review with tips, links, advertisements, or emails by a reviewer.

- Several sentiment-based content features have not been investigated, including the correlation between the sentiment of the review and its length, the correlation between the sentiment of the review and its level of detail, and the correlation between sentiment intensity score with the severity score, etc.
- Many sentiment-based behavioral features have not been explored, such as the ratio of posting a review with sentiment to rating inconsistency and the ratio of posting a review with sentiment to average product sentiment inconsistency, etc.
- Most Arabic studies only utilized content features, with only one study incorporating both behavioral and content features [109].

3.3.4 Methods:

- Most studies on detecting fake reviews in Arabic employ traditional machine learning classification. One study uses deep learning models, using only word embeddings as a feature in a translated dataset [165].
- No study has utilized a labeled dataset in English to annotate an original Arabic unlabeled dataset using a semi-supervised approach, despite the proven effectiveness of this method in other applications, such as Arabic text classification [166].

CHAPTER 4

APPROACH

This chapter forms the operational core of our research, outlining the systematic methodology employed to construct the Fake Review Detection system, and addressing the research questions and objectives outlined in the preceding chapters. It consists of three primary sections: 'Data Acquisition and Collection,' which details the processes of obtaining essential resources; 'Sentiment Analysis,' where methods are explained for annotating the corpus with sentiments; and 'Fake Review Detection,' which focuses on identifying and classifying fake reviews using a proposed feature set and various classification approaches.

The outlined approach for detecting fake reviews in Arabic text is detailed in this chapter, as illustrated in Figure 2. Subsequent sections will provide a detailed exposition of the tasks in each stage of the process.



Figure 2: Fake Review Detection Approach.

4.1 Data Acquisition and Collection:

We have gathered four datasets for training and testing in Sentiment Analysis and Fake Review Detection tasks. Two datasets were sourced from prominent e-commerce websites, featuring Arabic reviews. The third dataset served as the training data for our Sentiment

Analysis model, while the fourth dataset was employed for training the Fake Review Detection model. The subsequent sections provide details about each dataset and outline the approach used for their collection.

4.1.1 Saudi Amazon Dataset

The Saudi Amazon dataset is unlabeled. We extracted this dataset from the Saudi Amazon website using a web scrapping tool called BeautifulSoup [167]. The dataset is extracted in September 2023, it includes reviews from 2011 to 2023. The construction of the dataset was undertaken in several steps as illustrated in Figure 3. The final dataset contains 8,593 reviews, 3,599 reviews were originally written in Arabic and 4,994 reviews were written in English and then translated into Arabic using Google API [168]. Furthermore, every review in the dataset has been provided with various details, such as whether it is a verified purchase, the rating, review time, review country, product ID, product category, product title, and review title. The reviews of the dataset garnered an average rating of 4.1, where the highest attainable score is five. 94.24% of the reviews were confirmed to be from verified purchases. Furthermore, most of the reviews were posted in Saudi Arabia with a percentage of 60.32%. The rest of the reviews were written in the United Kingdom of Emirates, Egypt, and other non-Arabic countries. The dataset includes reviews from 5 main categories, home products, electronics, beauty, fashion, and books, with different numbers of reviews for each category. Table 6 represents the number of reviews in each category concerning the ratings and concerning the original review language.

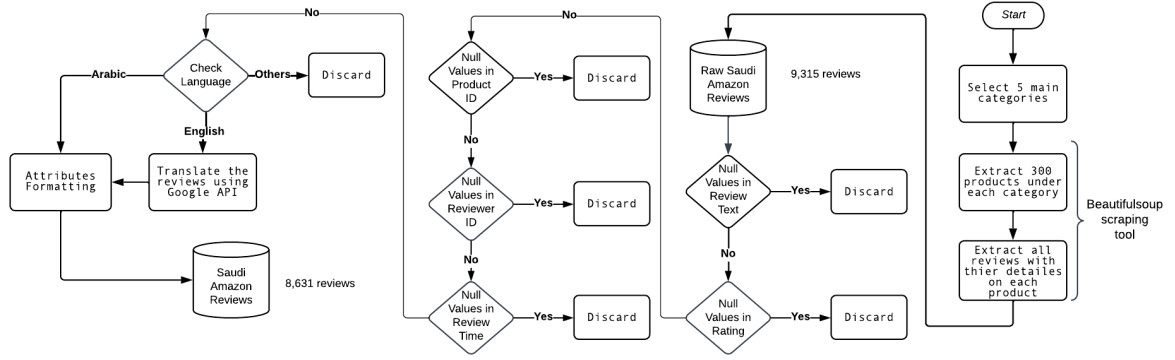


Figure 3: The construction process of the Saudi Amazon dataset.

Table 6: The distribution of the Saudi Amazon dataset.

Num	Product Category	Positive	Neutral	Negative	Original	Translated	Review Count
1	Home Products	1701	205	306	950	1262	2212
2	Beauty	1555	238	257	1010	1040	2050
3	Electronics	1596	225	376	1133	1064	2197
4	Books	940	82	157	583	596	1179
5	Fashion	808	102	83	127	866	993
Total		6600	852	1179	3803	4828	8631

4.1.2 AliExpress Dataset

The second dataset is the AliExpress dataset. We obtained this dataset through web scraping called Selenium [169] in March 2023, which encompasses reviews from both 2022 and 2023. We chose to use Selenium instead of BeautifulSoup for extracting reviews from AliExpress because the pages on AliExpress rely on JavaScript for loading. BeautifulSoup alone cannot handle JavaScript rendering, so we opted for Selenium, which is capable of interacting with dynamic content and rendering JavaScript.

The construction process of this dataset is identical to the construction process of the Saudi Amazon dataset, however, using a different scraping tool as shown in Figure 4. This dataset

comprises a total of 4,252 reviews, among which 123 were originally composed in Arabic, while 4,192 were initially in English and later translated into Arabic using Google API [168]. Additionally, each review in the dataset is accompanied by various details including the rating, review time, review country, product ID, product category, and reviewer ID. The reviews received an average rating of 4.53, with a maximum score of five. They originate from multiple Arabic-speaking countries such as Saudi Arabia, the United Arab Emirates, Oman, Qatar, Bahrain, Kuwait, Egypt, Palestine, Morocco, and Algeria. The dataset covers reviews from all specified categories in AliExpress, ranging from cellphones & telecommunications, consumer electronics, tools, automobiles & motorcycles, beauty & health, jewelry & accessories, hair, toys, watches, home improvements, home needs, security, sports, office, computers, women's clothing, men's clothing, kids, to pets, each with varying numbers of reviews. Then we categorized them into 4 main categories, home products, electronics, beauty, and fashion. Since AliExpress does not allow posting a review on a product without purchasing it, all the reviews were from verified purchases. Table 7 represents the number of reviews in each category concerning the ratings and concerning the original review language.

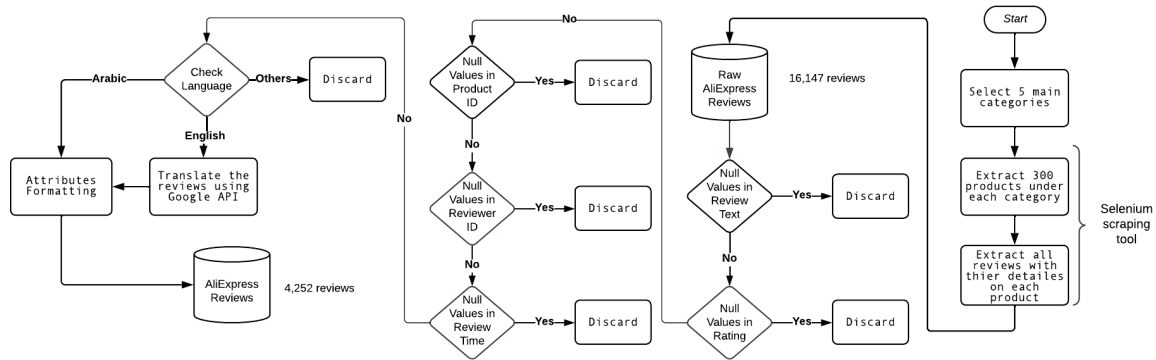


Figure 4: The construction process of the AliExpress dataset.

Table 7: The distribution of the AliExpress dataset.

Num	Product Category	Positive	Neutral	Negative	Original	Translated	Review Count
1	Home Products	947	45	83	25	1050	1075
2	Beauty	1102	90	107	58	1241	1299
3	Electronics	566	15	48	18	611	629
4	Fashion	1115	74	60	22	1227	1249
Total		3730	224	298	123	4129	4252

4.1.3 Amazon dataset

The Amazon dataset is publicly available [124]. It comprises 21,000 reviews provided and annotated by Amazon. Among them, 10,500 have been annotated as fake, while the other 10,500 are annotated as non-fake. Although there are many datasets for online products [7], [99], [106], [115], [117], [119], [120], [121], [123], [124], this dataset was chosen after careful consideration of several key factors. Firstly, it is labeled with both fake and authentic labels, the authentic reviews only will be used to train supervised machine learning models to identify the sentiment of the reviews. Since the authentic reviews in this dataset have a rating that is consistent with the sentiment of the reviews. Secondly, the labeling process relies on a filtering algorithm provided by the website, which offers a higher level of accuracy compared to manual annotations by experts or crowdsourcing platforms. Moreover, each review in the dataset is accompanied by a range of attributes, including verified purchase state, rating, product ID, product category, product title, and review title. For the verified purchase state in the authentic reviews, 84% of the data was confirmed as verified purchases. On average, the authentic reviews received a rating of 4.14, with five being the highest possible score. Additionally, the reviews are evenly spread across 30 distinct product categories such as PC, office products, kitchen, home improvement, electronics, health & personal care, and baby, each containing 700 reviews.

The reviews of the dataset are written in English, we translated these reviews into Arabic using Google API. Figure 5 shows the process followed to construct the dataset, while Table 8 illustrates the distribution of the dataset based on the product categories and the ratings.

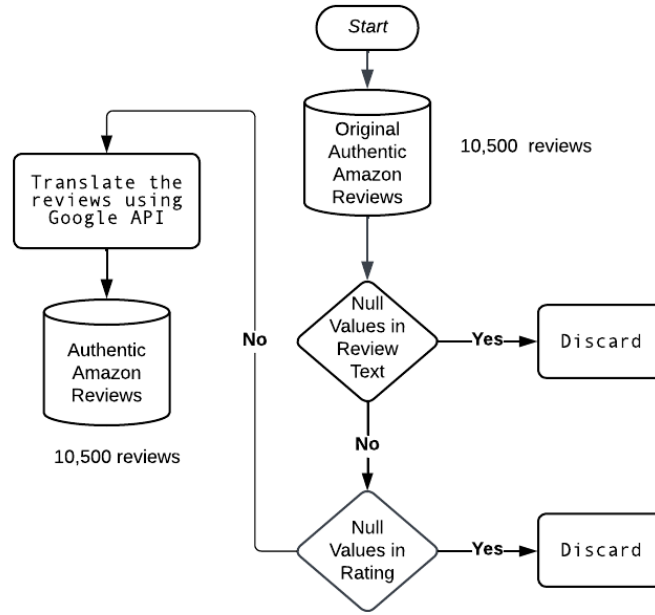


Figure 5: The construction process of the authentic Amazon reviews dataset.

Table 8: The distribution of the authentic reviews in the Amazon dataset.

Num	Product Category	Positive	Neutral	Negative	Review Count
1	Apparel	261	34	55	350
2	Automotive	274	29	47	350
3	Baby	267	36	47	350
4	Beauty	270	37	43	350
5	Books	285	25	40	350
6	Camera	276	36	38	350
7	Electronics	275	38	37	350
8	Furniture	265	30	55	350
9	Grocery	275	36	39	350

10	Health & Personal Care	271	39	40	350
11	Home	276	29	45	350
12	Home Entertainment	248	39	63	350
13	Home Improvement	273	29	48	350
14	Jewelry	273	31	46	350
15	Kitchen	285	23	42	350
16	Luggage	283	23	44	350
17	Musical Instruments	285	23	42	350
18	Office Products	252	30	68	350
19	Outdoors	287	22	41	350
20	PC	250	41	59	350
21	Pet Products	273	28	49	350
22	Shoes	258	38	54	350
23	Sports	285	40	25	350
24	Tools	289	31	30	350
25	Toys	268	31	51	350
26	Video DVD	283	28	39	350
27	Video Games	259	28	63	350
28	Watches	264	35	51	350
29	Wireless	245	28	77	350
30	Lawn and Garden	270	25	55	350
Total		8125	942	1433	10500

4.1.4 Electronic Retailers Yelp Dataset

Due to the lack of a public dataset for fake reviews in Arabic, we turned to searching for public databases in English. Table 10 showcases various public datasets available for fake reviews across different domains in English. There are three datasets found for online products [106], [121], [124]. However, the first two datasets [121], [124] were excluded due to the absence of the Reviewer-Id attribute, which is needed for calculating behavioral features—an essential objective of the thesis. The third dataset [106] emerged as the most recent and suitable option, boasting a substantial volume of reviews. Despite considerable effort spent on extracting reviews from this dataset using a supercomputer, we discovered that the dataset lacked authentic reviews with 4 and 5 ratings, as well as fake reviews with

1 and 2 ratings. Consequently, models achieved 100% accuracy in detecting fake reviews based solely on ratings. To address this, we pursued a pivot to a dataset closely aligned with the online products domain. The Yelp dataset for electronic retailers [125] emerged as the first choice. This dataset encompassed all necessary attributes for the system. The dataset contains two CSV files. The first file contains all the reviews with their details and labels, and the second file contains full information about each reviewer with different behavioral features. While the second file, containing detailed reviewer information and behavioral features, was available, it posed challenges as the calculated features did not align with the data in the first file. Discrepancies, such as mismatches in the number of reviews for each reviewer, led to the decision to base all behavioral features on the information provided in the first file for consistency, as we need to calculate extra behavioral features that were not included in the second file. The process we went through to construct this dataset is illustrated in Figure 6. The updated dataset contains 24,760 reviews in four cities, 32% of the dataset for fake reviews, and 68% of the dataset for authentic reviews. The average rating of the dataset is 3.5. The full distribution of the dataset is depicted in Table 9.

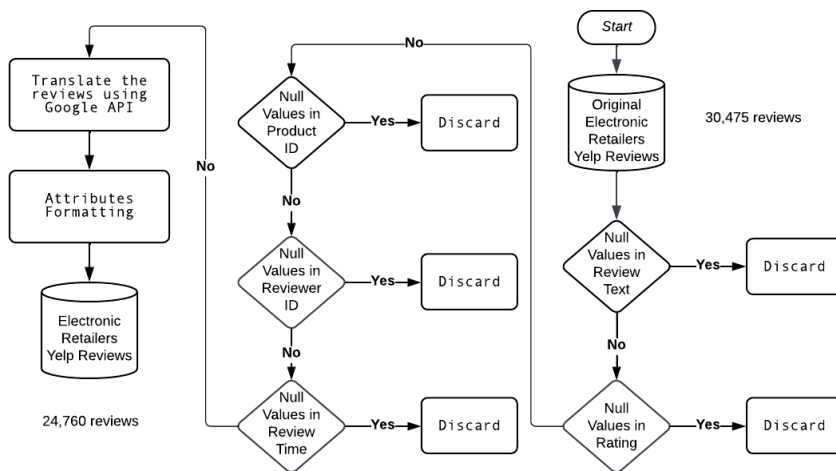


Figure 6: The construction process of Electronic Retailers Yelp dataset.

Table 9: The distribution of Electronic Retailers Yelp dataset.

Num	City	Positive	Neutral	Negative	Fake	Authentic	Review Count
1	New York	3815	275	2399	2052	4437	6489
2	Los Angeles	5968	259	3454	3134	6547	9681
3	Miami	1648	82	992	1145	1577	2722
4	San Francisco	3693	173	2002	1566	4302	5868
Total		15124	789	8847	7897	16863	24760

Table 10: The publicly available and annotated datasets for fake reviews detection.

Ref	Dataset	Domain	Year	Labeling Approach	Volume	Reviewer_Id	Product_Id	Rating	Review_Text	Review_Time	Verified_Purchase
[79]	Yelp	Hotels and Restaurants	2013	Yelp Filter	67 K	✓	✓	✓	✓	✓	X
[108]	DOSC	Hotels	2013	AMT	1,600	X	X	X	✓	X	X
[121]	Amazon	Online Products (Books)	2014	Rule-Based	6,819	X	✓	✓	✓	X	X
[112]	Datatang	Hotels and Restaurants	2014	Yelp Filter	9,765	✓	✓	✓	✓	X	X
[100]	YelpZip	Hotels and Restaurants	2015	Yelp Filter	608 K	✓	✓	✓	✓	✓	X
[100]	YelpCHI	Hotels and Restaurants	2015	Yelp Filter	67 K	✓	✓	✓	✓	✓	X
[100]	YelpNYC	Hotels and Restaurants	2015	Yelp Filter	395 K	✓	✓	✓	✓	✓	X
[124]	Amazon	Online Products	2018	Amazon Filter	21 K	X	✓	✓	✓	X	✓
[125]	Yelp	Electronics Retailers	2019	Yelp Filter	30 K	✓	✓	✓	✓	✓	X
[106]	Amazon	Online Products	2020	Rule-Based	26 M	✓	✓	✓	✓	✓	X

4.2 Sentiment Analysis

To perform sentiment analysis on Arabic reviews using both machine learning and deep learning models, we need to have a well-annotated dataset for training. Our objective is to accurately assess the sentiment of a review and identify any potential discrepancies between the expressed sentiment and other features that contribute to detecting fake reviews. Therefore, the model needs to be trained on a dataset where sentiments are accurately labeled. Unfortunately, the public datasets for online products, LABR [57], BRAD [58], and PROD [59] were labeled based on the rating. So, they are not a well-annotated dataset for training. Regrettably, the Amazon UK dataset lacks sentiment labels. However, the annotation of this dataset considered the review text that contradicted their rating as a fake review. Therefore, we can rely on the ratings of the 10,500 authentic reviews in the dataset to represent the sentiments of the reviews.

The framework for conducting sentiment analysis is illustrated in Figure 7. The following subsections illustrate this framework in detail, starting by preparing the data, then developing and training different machine learning and deep learning models, and finally evaluating the performance of the trained models.

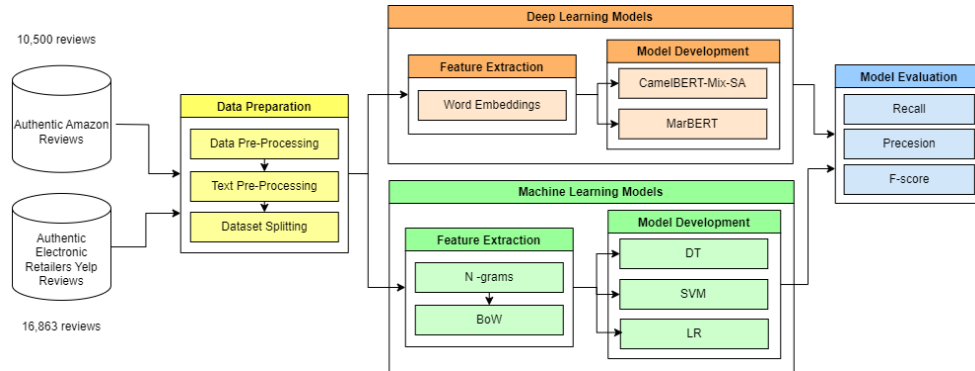


Figure 7: The framework of conducting Arabic sentiment analysis.

4.2.1 Data Preparation

The process of preparing data for training includes pre-processing the dataset, pre-processing the text reviews, and splitting the dataset as illustrated in Figure 8.

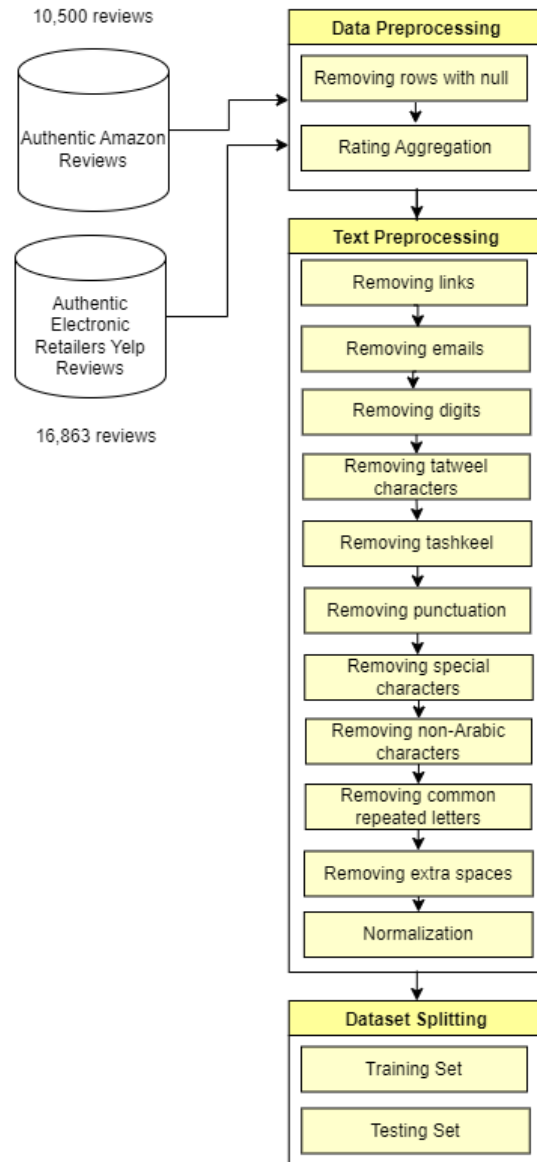


Figure 8: Data preparation for Arabic sentiment analysis.

1) Data Preprocessing

Preprocessing the data involved cleaning the dataset by handling the null values in the columns and performing data formatting to convert some values into a particular data type

and a particular format in the two datasets, the authentic Amazon, and authentic electronic Yelp reviews dataset. No rows have been deleted since we handled the null values in the construction of the datasets at the beginning. So, all the rows contain both the review text and rating fields. However, the values of rating columns have been converted first to numerical values (1, 2, 3, 4, 5) in the two datasets then these values have converted into a different scale. The positive ratings (4 and 5) have been converted to 1 and the negative ratings (1 and 2) have been converted to 3, while the natural ratings (3) have been converted to 2.

2) Text Preprocessing

Preprocessing the review texts undergoes various steps. First, removing links, emails, digits, tatweel and tashkeel characters, punctuation, special characters, non-Arabic characters, common repeated letters such as (ي, و, ا, ر), and extra spaces. After that, applying Arabic character normalization to ensure consistency in the text data by converting (آ, إ, ؤ) to (ا) and converting (ة) to (o).

3) Dataset Splitting

In the task of dataset splitting, the datasets were divided into training and evaluation sets with 80% for training and 20% for testing to facilitate model training and evaluation. Initially, we trained the models with a binary classification approach, focusing on positive and negative labels while excluding reviews with neutral ratings. However, it became evident that fake reviews may exist within the neutral rating category, blurring the line between authentic and fake content. Thus, excluding reviews with a rating of 3 resulted in the loss of numerous reviews containing potentially valuable information. So, we have repeated the experiments by including neutral reviews. Table 11 illustrates the

distributions of the train set and test set for Authentic Amazon Reviews using two labels, while Table 12 presents the distributions for the same dataset using three labels. On the other hand, Table 13 displays the train set and test set distributions for Authentic Electronic Retailers Yelp Reviews using two labels, and Table 14 showcases the distributions using three labels.

Table 11: The training and testing sets of Authentic Amazon Reviews for SA using 2 labels.

Set	Positive	Negative	Total
Training	6509	1133	7642
Testing	1616	295	1911
Total	8125	1428	9553

Table 12 The training and testing sets of Authentic Amazon Reviews for SA using 3 labels.

Set	Positive	Neutral	Negative	Total
Training	6496	764	1135	8400
Testing	1629	178	293	2100
Total	8125	942	1428	10500

Table 12:The training and testing sets of Authentic Electronic Retailers Yelp Reviews for SA using 2 labels.

Set	Positive	Negative	Total
Training	8728	4284	13012
Testing	2135	1118	3253
Total	10863	5402	16265

Table 13: The training and testing sets of Authentic Electronic Retailers Yelp Reviews for SA using 3 labels.

Set	Positive	Neutral	Negative	Total
Training	8652	475	4363	13490
Testing	2211	123	1039	3373
Total	10863	598	5402	16863

In the authentic Amazon dataset, we have used 8,400 reviews to train the model and 2,100 reviews to evaluate the performance of the model using 3 labels. Using 2 labels, we have used 7,646 reviews to train the model and 2,100 reviews to evaluate the performance of the model.

In authentic electronic retailers' Yelp dataset, we have used 13,490 reviews to train the model and 3,373 reviews to evaluate the performance of the model using 3 labels. Using 2 labels, we have used 13,012 reviews to train the model and 3,253 reviews to evaluate the performance of the model.

4.2.2 Training Model

The training has been done using both deep learning models and traditional machine learning. In the following sections, we provide a background on the utilized models in this research and describe the approach utilized in the training for both deep learning and traditional machine learning models on the authentic reviews in the Amazon and Yelp dataset to find the sentiments.

1) Deep Learning SA Models

The two deep learning models utilized for sentiment analysis in this research work are based on a machine learning technique called Transfer Learning. In this technique, models trained on a specific task were adapted to perform another related task. This technique helps to leverage knowledge gained from one task to improve learning and performance on a different, but related, task with less data or computational resources than training a model from scratch for the second task. Figure 9 illustrates the Transfer Learning concept, adopted from [170]. How large or how small the dataset should be varies widely depending on the specific model architecture and the task it was pre-trained on.

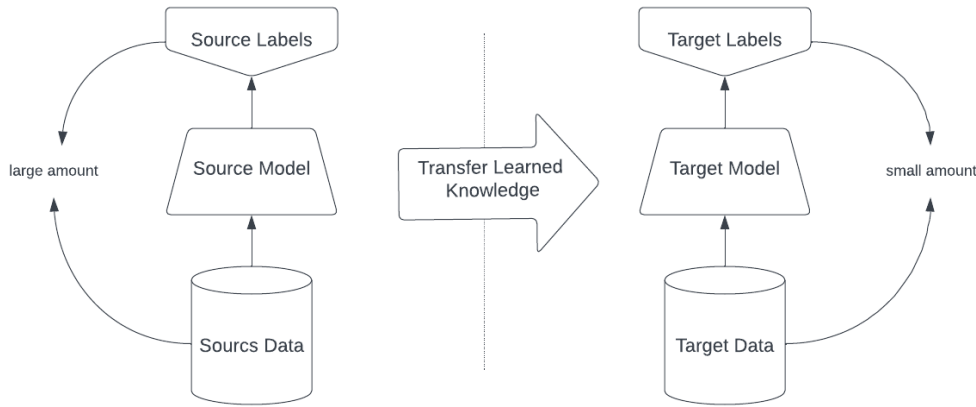


Figure 9: The concept of Transfer Learning.

These two deep learning models are as follows:

a) CAMELBERT-Mix-SA

CAMELBERT-Mix-SA [171] is a pre-trained model that was built by fine-tuning the CAMELBERT-Mix for sentiment analysis tasks using SemEval [75], ArSAS [172], and ASTD [74] datasets. CAMELBERT-Mix is a pre-trained model on a dataset of size 167GB which contained over 17.3 billion words from various sources, such as news articles, books, social media posts, etc., in MSA, DA, and CA. These models are based on the BERT architecture, which uses a bidirectional transformer to encode both the left and the right context of a word. The CAMELBERT-Mix-SA and CAMELBERT-Mix models were developed by the CAMEL-Lab and are available on Hugging Face [173], [174].

b) MARABERT

MARABERT [175] is a pre-trained language model for Arabic that was developed by Abdul-Mageed et al. in 2020. The model was trained on a dataset of size 128GB which contained over 15.6 billion words from various sources, such as news articles, books, Wikipedia pages, etc., in both MSA and DA. This model is also based on the BERT

architecture. However, it differs from CAMELBERT-Mix in the data size, language variant, and fine-tuning task type. The MARABERT model is fine-tuned on two different tasks: sentiment analysis and question answering. MARABERT is also available on Hugging Face [176].

Deep Learning SA Model Development

For deep learning models, two deep learning models were utilized which are, the CAMELBERT-Mix-SA and MARBERT models. Both models are pre-trained language models on sentiment analysis tasks. We fine-tuned these models for sentiment analysis tasks on the authentic reviews from the Amazon dataset. The models were trained using a cross-entropy loss function and Adam optimizer with an input sequence length of 410, 4 epochs, a batch size of 32, and a learning rate of $1e-5$.

2) Traditional Machine Learning SA Models

Three traditional machine-learning models were utilized for sentiment analysis. They are Decision Tree, Support Vector Machine, and Logistic Regression. These models are highlighted in the following subsections.

a) Decision Tree (DT)

Decision Trees are non-linear models used for both classification and regression tasks. They operate by recursively partitioning the input space into regions and assigning a label to each region. The tree structure is built through a process of selecting features that best split the data based on certain criteria, typically aiming to maximize information gain or minimize impurity. Decision Trees are interpretable and intuitive but can be prone to overfitting [177].

b) Support Vector Machine (SVM)

Support Vector Machines are supervised learning models used for classification and regression tasks. SVM aims to find a hyperplane that best separates the data into different classes. It chooses the hyperplane with the maximum margin, which is the distance between the hyperplane and the nearest data points from each class. SVM is effective in high-dimensional spaces and is versatile due to different kernel functions, allowing it to handle non-linear relationships [177].

c) Logistic Regression (LR)

Despite its name, logistic regression is a classification algorithm used for binary and multi-class classification problems. It models the probability that an instance belongs to a particular class using the logistic function. Logistic Regression estimates the coefficients for each feature and combines them linearly, passing the result through the logistic function to obtain probabilities. It is computationally efficient, interpretable, and works well when the relationship between features and classes is approximately linear [177].

3) Feature Extraction

Traditional machine learning models, equipped with a set of features, form the basis of our approach. We incorporated various techniques to ensure a nuanced understanding of the sentiment expressed in the reviews. This relies on extracting different features from the text data and involves transforming raw text into a format that machine learning models can utilize for training and prediction. The employed range of feature extraction techniques to capture different aspects of the text data are Bag of Words and N-grams.

a) Bag of Words (BoW)

Scikit-learn's CountVectorizer [178] is employed for BoW. BoW converts text documents into numerical feature vectors. It represents text data as the frequency of occurrence of each word in the dataset. The minimum number of documents in which a word must appear in the corpus has been set to 5 to ensure that words appearing in fewer than 5 reviews are ignored, reducing noise.

b) N-grams

The `ngram_range = (1,2)` parameter in CountVectorizer allows for the inclusion of both unigrams and bigrams. N-grams enable the model to understand the influence of word combinations on sentiment.

Machine Learning SA Model Development

Three traditional machine learning models, DT, SVM, and LR, were employed for sentiment analysis task using the Scikit-learn library [178]. Each model was trained using a Bag of Words (BoW) approach, specifically utilizing CountVectorizer with unigrams and bigrams and a minimum document frequency of 5. The feature set for each model includes the frequency of occurrence of individual words.

4.2.3 Sentiment Analysis Model Evaluation

For the evaluation of sentiment analysis models, precision (P), recall (R), and F-score (F) are fundamental metrics that provide insights into the model's effectiveness in correctly classifying sentiment categories. These metrics are calculated as follows:

$$P = \frac{TP}{TP + FP} \quad (\text{Eq.1})$$

$$R = \frac{TP}{TP + FN} \quad (\text{Eq.2})$$

$$F = \frac{2PR}{P+R} \quad (\text{Eq.3})$$

Where TN is True Negative, TP is True Positive, FN is False Negative, and FP is False Positive.

Precision (P): Precision assesses the accuracy of positive predictions by measuring the ratio of correctly predicted positive instances to the total instances predicted as positive (see Eq.1). Precision is especially relevant when the cost of false positives is high.

Recall (R): Recall, also known as sensitivity or true positive rate, quantifies the ability of the model to correctly identify all positive instances (see Eq.2).

F-score (F): The F-score, often referred to as the F1 score, is the harmonic mean of precision and recall. It provides a balanced assessment of a model's performance by considering both false positives and false negatives (see Eq.3).

Our task involves imbalanced sentiment analysis, where one sentiment class in the dataset may significantly outnumber the other. To accurately assess model performance, precision, recall, and F-score were calculated using a weighted average. This approach considers the class prevalence in the dataset, assigning more weight to the class with fewer instances. This consideration is crucial for obtaining a balanced view of how the model performs across different sentiment classes.

We trained five models—LR, SVM, DT, MARABERT, and CAMELBERT-Mix-SA—for Arabic sentiment analysis on the test sets of two datasets: authentic Amazon reviews and the electronic retailers Yelp dataset. The analysis encompassed both binary classification with two labels and ternary classification with three labels. The specific outcomes obtained

from the test set of authentic Amazon reviews dataset are showcased in Table 15. In Table 16, the results for the test set of the electronic retailers Yelp dataset are presented.

In general, the models tend to achieve higher precision, recall, and F1 scores in the binary classification setting compared to the ternary classification. This is expected, as neutral reviews often contain a mixture of positive and negative sentiments, posing a challenge for precise classification. Moreover, the pre-trained deep learning models outperformed all traditional machine learning models in all evaluation metrics in authentic Amazon reviews. However, in authentic electronic retailers Yelp reviews, LR and SVM demonstrated strong performance in ternary sentiment analysis with F1-score of 92.36% and 90.35% respectively that surpassed the pre-trained MARABERT with the F1-score of 89.86%. Also, among the traditional machine learning models, LR emerged as the most effective performer. Despite the formidable capabilities showcased by the advanced models like MARABERT, and CAMELBERT-Mix-SA in the sentiment analysis tasks, LR consistently demonstrated competitive precision, recall, and F1 scores. This underscores the robustness of LR in handling sentiment analysis tasks, emphasizing the importance of considering both traditional and state-of-the-art models based on the specific requirements and nuances of the given dataset. Furthermore, among the models, CAMELBERT-Mix-SA consistently demonstrates high performance across all metrics in both binary and ternary classifications. It achieves the highest precision of 92.91%, recall of 93.11%, and F1 score of 92.98% in Electronic Retailers Yelp dataset. However, in Amazon dataset, MARABERT was the best performing model with precision of 83.55%, recall of 86.16%, and F1 score of 84.12%. While DT model tends to have lower scores compared to other models in the two test sets, indicating potential challenges in capturing the complexities of

the sentiment analysis task. Finally, the results suggest that MARABERT and CAMELBERT-Mix-SA are promising models for Arabic sentiment analysis, with CAMELBERT-Mix-SA showing particularly strong performance across both label configurations in Yelp dataset and MARABERT for Amazon dataset.

Table 14: The results of Arabic sentiment analysis on authentic Amazon reviews.

Model	Labels	Precision	Recall	F1
LR	2	0.9002	0.9064	0.898
SVM		0.8947	0.9011	0.8962
DT		0.8368	0.8585	0.8201
MARABERT		0.9335	0.9358	0.934
CAMELBERT-Mix-SA		0.9291	0.9311	0.9298
LR	3	0.8011	0.8242	0.8097
SVM		0.7918	0.8046	0.7971
DT		0.7136	0.7898	0.7292
MARABERT		0.8355	0.8616	0.8412
CAMELBERT-Mix-SA		0.8288	0.8558	0.8335

Table 15: The results of Arabic sentiment analysis on authentic Electronic Retailers Yelp reviews.

Model	Labels	Precision	Recall	F1
LR	2	0.8617	0.8948	0.8778
SVM		0.843	0.8749	0.8586
DT		0.7433	0.7744	0.7582
MARABERT		0.9426	0.9413	0.9404
CAMELBERT-Mix-SA		0.9461	0.9462	0.9459
LR	3	0.9243	0.9244	0.9236
SVM		0.9035	0.9041	0.9035
DT		0.7941	0.7977	0.7946
MARABERT		0.8947	0.9111	0.8986
CAMELBERT-Mix-SA		0.9291	0.9311	0.9298

4.2.4 Annotating the corpora

The best models from the previous stage on the ternary classification of sentiments of the reviews were utilized in this stage to annotate the three corpora, Saudi Amazon,

AliExpress, and Electronic Retailers Yelp. Saudi Amazon and AliExpress corpora were annotated using MARABERT model that trained on authentic Amazon reviews. This choice was based on MARABERT's superior performance on the authentic Amazon reviews dataset, which falls under the same domain as the Saudi Amazon and AliExpress corpora for online products. While Electronic Retailers Yelp corpus was annotated using CAMELBERT-Mix-SA model that trained on authentic Electronic Retailers Yelp dataset, which is part of the same main dataset. However, to make sure that the predicted labels by the models are correct, we performed further investigation and evaluation. We checked only the reviews where the predicted labels were not consistent with their rating scores. To make sure that they are in fact not consistent, or that they are due to a wrong classification by the model. The number of cases in Saudi Amazon corpus that have predicted sentiments not matched with the rating scores was 1530 reviews, 588 were correctly predicted and 942 were incorrectly predicted. While in AliExpress, the number of reviews with mismatching predicted sentiments with the rating scores was 430 reviews, 153 were correctly predicted and 277 were incorrectly predicted. For the Electronic Retailers Yelp corpus, the number of reviews with mismatching predicted sentiments with the rating scores was 1158 reviews, 155 were correctly predicted and 1003 were incorrectly predicted.

We evaluated the correctness of the predicted sentiments on the two corpora before and after relying on the corrected sentiments instead of the rating. In Table 17, the evaluation of the MARABERT model on Saudi Amazon and AliExpress corpora is presented. While Table 18 presents the evaluation of the CAMELBERT-Mix-SA model on Electronic Retailers Yelp corpus.

The results demonstrate that the rating scores are not a good representative of the sentiment of the review. Incorporating corrected sentiments in lieu of the original rating scores leads to notable improvements in model performance across all evaluated metrics. Moreover, this shows the ability of the trained MARABERT and CAMELBERT-Mix-SA models to predict the actual sentiments correctly better than relying on the rating scores.

Table 16: The evaluation of MARABERT sentiment analysis model on Saudi Amazon and AliExpress corpora.

Model	Labels	Corpus	Case	Precision	Recall	F1
MARABERT	3	Saudi Amazon	Before	0.7933	0.8227	0.8015
			After	0.838	0.8586	0.8366
		AliExpress	Before	0.8849	0.8989	0.889
			After	0.8849	0.9036	0.8878

Table 17: The evaluation of CAMELBERT-Mix-SA sentiment analysis model on Electronic Retailers Yelp corpus.

Model	Labels	Corpus	Case	Precision	Recall	F1
CAMELBERT-Mix-SA	3	Electronic Retailers Yelp	Before	0.9532	0.9532	0.9532
			After	0.9563	0.9565	0.9564

4.2.5 Building the lexicons

To construct lexicons from three distinct corpora (Saudi Amazon, AliExpress, and Electronic Retailers Yelp), we employed two LR models. The first model was trained on translated reviews across all three corpora, while the second model was trained exclusively on original Arabic reviews from the Saudi Amazon and AliExpress corpora. The decision to train these models separately stemmed from our observation that certain words in MSA text can convey different sentiments in Classical Arabic (CA) text. The labels used for training these models were derived from corrected sentiments identified in the preceding stage, rather than relying on rating scores. The methodology employed for developing the two models aligns with the approach detailed in ‘*Traditional Machine Learning SA Models*’ section. After model training, we extracted feature names sorted by coefficient

indices. These feature names serve as unigram and bigram lexicons, sorted based on their associated sentiments, from the most positive words to the most negative words. To refine the lexicons, we underwent a thorough examination, discarding neutral entries, validating the accuracy of sentiment assignments, and rectifying misclassified lexicons. The resulting lexicon set comprised 1947 words extracted from the three corpora using the trained LR models. Additionally, 199 lexicons were manually added, bringing the final lexicon count to 2146. The number of collected lexicons is illustrated in Table 19.

Table 18: The collected lexicons.

State/Sentiment	Positive	Negative	Total
Original	205	347	552
Translated	960	634	1594
Total	1165	981	2146

4.3 Fake Reviews Detection

After preparing the dataset and finding the sentiments from the previous stages, this stage aims to perform fake reviews detection on the collected datasets. The framework of performing the fake review detection is illustrated in Figure 10. First, the approach starts by identifying the important features, then training both supervised and semi supervised models, and finally evaluating the trained models. The following sections describe each phase in detail.

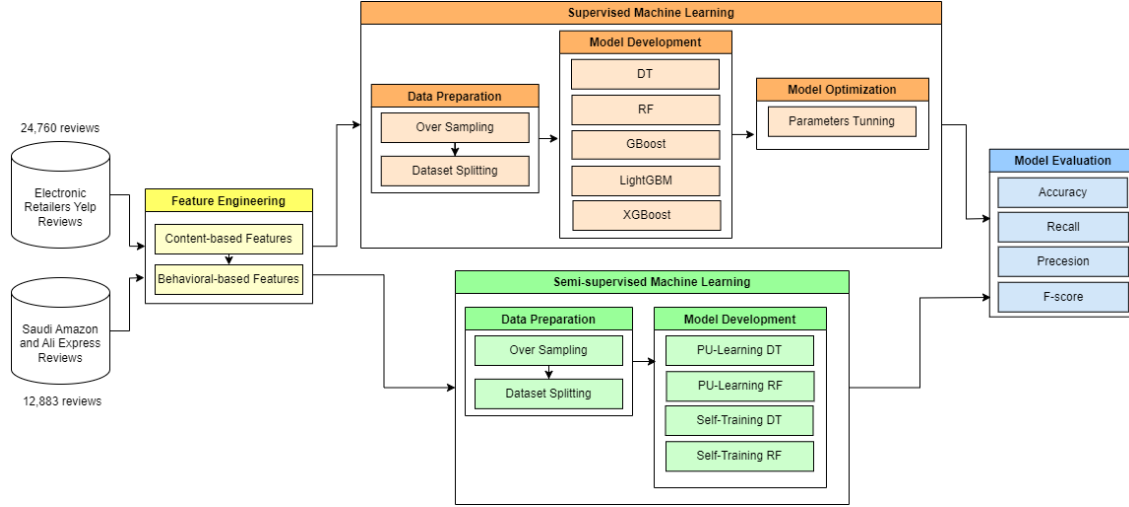


Figure 10: The framework of conducting fake review detection.

4.3.1 Feature Engineering

To detect fake reviews using machine learning models, we computed numerous features. We categorized them into two primary sets: content-based features and behavioral-based features. Subsequently, we further subdivided these sets into sentiment-based features and other features. Table 20 provides a detailed overview of these features, including their descriptions and codes in the dataset. The proposed sentiment-based features are denoted with an asterisk (*) to distinguish them from other features identified by previous studies. Additionally, Table 21 outlines the computation methodology for each of these features. The collected lexicons which mentioned in Section 4.2.5 were utilized here to count the number of positive and negative words in the review. The number of tip words in the review was also calculated using a set of 17 identified words indicative of a tip. Furthermore, regular expressions were employed to determine the number of sentences, words, characters, emails, and links in each review. The categorization of text length was thoughtfully determined based on the distribution of text lengths observed in the dataset. Cosine similarity is utilized to calculate the four content similarity features, CSSRSP, CSSRDP, CSDRSP, and CSDRDP. It serves as a quantitative measure to assess the textual

similarity between different reviews. Specifically, it measures the cosine of the angle between two non-zero vectors in the TF-IDF space, providing a normalized measure of their textual resemblance. The TF-IDF vectors are derived from the review text, capturing the importance of individual words while accounting for their frequency across the entire dataset. Cosine similarity values range from -1 to 1, where a higher value close to 1 indicates a greater similarity between the reviews.

Different proposed equations were employed to calculate the other features, contributing to a comprehensive feature engineering approach.

Table 19: The utilized feature sets for fake review detection.

Type	Sub-type	Feature	Description	Code
Content-based Features	Sentiment-based Features	Sentiment	The sentiment of the review (positive, negative, or neutral) as numbers [154].	S
		Positive Words Count	Number of positive words in the review text [152].	PWC
		Negative Words Count	Number of negative words in the review text [152].	NWC
	Others	Sentence Count	Number of sentences in the review text [152].	SC
		Word Count	Number of words in the review text [152].	WC
		Character Count	Number of characters in the review text [135].	CC
		Tip Count	Number of tip words in the review text [156].	TC
		Email Count	Number of emails in the review text.	EC
		Link Count	Number of links in the review text [158].	LC
		Text Length Category	The text length category of the review text in 5 and 3 scales. (very long, long, medium, short, very short) and (long, medium, short) [179].	TLC
Behavioral-based Features	Sentiment-based Features	Sentiment To Rating Inconsistency	The inconsistency between the review rating and the review text sentiment [179].	SRI
		Product Sentiment Inconsistency *	The inconsistency between the average sentiment of product reviews and the user's review text sentiment.	PSI

		User Sentiments Inconsistency *	The inconsistency between the average sentiment of user's reviews and its review text sentiment.	USI
		Sentiment to Text Length Inconsistency *	The inconsistency between the text length of the review text and the user's review text sentiment.	STLI
		Reviews with Sentiment to Rating Inconsistency Count *	Number of reviews with Sentiment to rating inconsistency.	SRIC
		Reviews with Product Sentiment Inconsistency Count*	Number of reviews with Product sentiment inconsistency	PSIC
	Others	Review Rating	The rating of the review in 5 and 3 scales. (5, 4, 3, 2, 1) and (3, 2, 1) [128].	R
		Reviews Count	Number of user's reviews [156].	RC
		Reviews with Rating Count	Number of user's reviews with a specific rating.	RCR[1-5]
		Maximum Review Count Per Day	Maximum number of user's reviews posted per day [156].	MRC
		Rating Deviation	The deviation of the review rating of a product from the average review ratings of a product [126].	RD
		Content Similarity (same reviewer on the same product) *	The maximum cosine similarity of the review texts of the of same reviewer on the same product.	CSSRSP
		Content Similarity (same reviewer on different products) Content Similarity	The maximum cosine similarity of the review texts of the of same reviewer on different products [7].	CSSRDP
		Content Similarity (different reviewers on the same product)	The maximum cosine similarity of the review texts of the of different reviewers on the same product [7]	CSDRSP
		Content Similarity (different reviewers on different products)	The maximum cosine similarity of the review texts of the of different reviewers on different products [7]	CSDRDP

Table 20: The utilized equations to compute the fake review detection features.

Code	Feature	Equation
S	Sentiment	$S(r) = \begin{cases} 1, & SENT(r) = positive \\ 2, & SENT(r) = neutral \\ 3, & SENT(r) = negative \end{cases}$
PWC	Positive Words Count	$PWC(r) = \sum_{i=1}^n POS(W_i)$ where n represents the number of words in review r

NWC	Negative Words Count	$NWC(r) = \sum_{i=1}^n NEG(W_i)$ where n represents the number of words in review r
SC	Sentence Count	$SC(r) = \sum_{i=1}^n EndOfSentence(C_i)$ where n represents the number of characters in review r
WC	Word Count	$WC(r) = \sum_{i=1}^n (1)$ where n represents the number of words in review r
CC	Character Count	$CC(r) = \sum_{i=1}^n (1)$ where n represents the number of characters in review r
TC	Tips Count	$TC(r) = \sum_{i=1}^n TIP(W_i)$ where n represents the number of words in review r
EC	Email Count	$EC(r) = \sum_{i=1}^n EMAIL(W_i)$ where n represents the number of words in review r
LC	Link Count	$LC(r) = \sum_{i=1}^n LINK(W_i)$ where n represents the number of words in review r
TLC5, TLC3	Text length category	$TLC5(r) = \begin{cases} 1, & WC(r) < 5 \\ 2, & 5 < WC(r) \leq 20 \\ 3, & 20 < WC(r) \leq 50 \\ 4, & 50 < WC(r) \leq 150 \\ 5, & WC(r) > 150 \end{cases}$ $TLC3(r) = \begin{cases} 1, & 5 < WC(r) \leq 20 \\ 2, & 20 < WC(r) \leq 150 \\ 3, & WC(r) > 150 \end{cases}$
SRI	Sentiment to rating inconsistency	$SRI(r) = Abs(R3(r) - S(r))$
PSI	Product sentiment inconsistency*	$AvgPrSen(P) = \left\lfloor \frac{1}{n} \sum_{i=1}^n S(r_i) \right\rfloor$ where n represents the number of reviews on product P $PSI(P, r) = Abs(AvgPrSen(P) - S(r))$
USI	User sentiments inconsistency*	$AvgUserSen(Rvr) = \frac{1}{n} \sum_{i=1}^n S(r_i)$ where n represents the number of reviews by reviewer Rvr $USI(Rvr, r) = Abs(AvgUserSen(Rvr) - S(r))$
STLI	Sentiment to text length inconsistency*	$STLI(r) = Abs(TLC3(r) - S(r))$

SRIC	Reviews with Sentiment to rating inconsistency Count*	$SRIC(Rvr) = \sum_{i=1}^n SRI(r_i) == 2$ where n represents the number of reviews by reviewer Rvr
PSIC	Reviews with Product sentiment inconsistency Count*	$PSSIC(Rvr) = \sum_{i=1}^n \sum_{j=1}^m PSI(P_i, r_j) == 2$ where n represents the number of products the reviewer Rvr posted a review on, and m represent the number of reviews posted by reviewer Rvr on product P_i
R, R3	Review Rating	$R(r) \in \{1,2,3,4,5\}$ $R3(r) = \begin{cases} 1, & R(r) > 3 \\ 2, & R(r) = 3 \\ 3, & R(r) < 3 \end{cases}$
RC	Reviews Count	$RC(Rvr) = \sum_{i=1}^n (r_i)$ where n represents the number of reviews by reviewer Rvr
RCR[1-5]	Reviews with Rating Count	$RCR(Rvr, R) = \sum_{i=1}^n (r_i) \text{ and } R(r_i) == R$ where n represents the number of reviews by reviewer Rvr and R represents a specific rating value
MRC	Maximum Review Count per Day	$RCperDay(Rvr, Day) = \sum_{i=1}^n 1$ where n represents the number of reviews by reviewer Rvr in day Day $RC(Rvr) = \text{Max}_{i=1}^n (RCperDay(Rvr, Day_i))$ where n represents the number of days the reviewer Rvr posted a review
RD	Rating Deviation	$AvgProdR(P) = \left[\frac{1}{n} \sum_{i=1}^n R(r_i) \right]$ where n represents the number of reviews on product P $RD(r) = \text{Abs}(AvgProdR(P) - R(r))$
CSSRSP	Content Similarity (Same reviewer on the same product)	$CSSRSP(Rvr) = \text{Max}_{i,p} \left(\frac{r_{i,p} \cdot r_{j,p}}{\ r_{i,p} \cdot r_{j,p}\ } \right)$ where p represents a specific product, i and j represents the index of different reviews by reviewer Rvr and $i \neq j$
CSSRDP	Content Similarity (Same reviewer on different products) Content Similarity	$CSSRDP(Rvr) = \text{Max}_{i,p} \left(\frac{r_{i,p} \cdot r_{j,q}}{\ r_{i,p} \cdot r_{j,q}\ } \right)$ where p and q represent different products, i and j represents the index of different reviews by reviewer Rvr and $i \neq j$
CSDRSP	Content Similarity (Different reviewers on the same product)	$CSDRSP(Rvr) = \text{Max}_{i,p} \left(\frac{r_{i,p} \cdot r_{j,p}}{\ r_{i,p} \cdot r_{j,p}\ } \right)$

		where p represents a specific product, i and j represents the index of different reviews by different reviewers Rvr and another reviewer
CSDRDP	Content Similarity (Different reviewers on different products)	$CSDRDP(Rvr) = \text{Max}_{i,p} \left(\frac{r_{i,p} \cdot r_{j,q}}{\ r_{i,p} \cdot r_{j,q}\ } \right)$ where p and q represent different products, i and j represents the index of different reviews by different reviewers Rvr and another reviewer

4.3.2 Fake Reviews Detection Approaches

To conduct fake review detection on online reviews, we employed two primary methodologies. The initial approach involved supervised machine learning applied to labeled datasets. In contrast, the second approach leveraged two variants of semi-supervised learning known as Self-Training and Positive Unlabeled Learning (PU-Learning). The subsequent sections will provide a detailed exploration of the experiments conducted within each of these approaches.

1) Supervised Machine Learning (SML)

The general approach of conducting fake review detection using supervised machine learning is illustrated in Figure 11. The following subsections present the data preparation stage, the supervised machine learning model development, model optimization, and model evaluation.

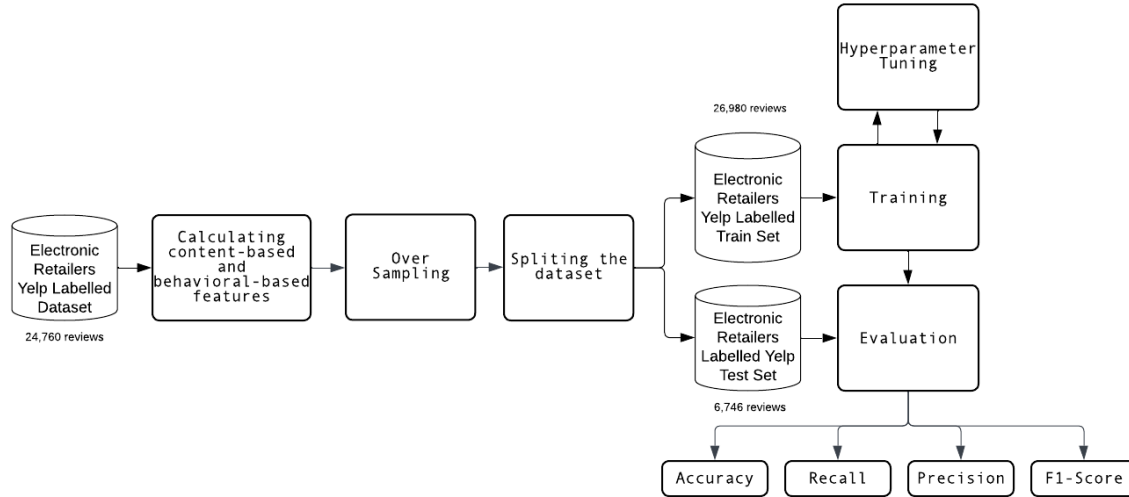


Figure 11: Supervised learning fake review detection approach.

a) Data Preparation for Fake Reviews Detection using SML

The electronic retailers Yelp dataset was utilized for this task. The dataset is labeled with fake and authentic labels along with the sentiments using the previously mentioned approach in Section 4.2.4 to annotate the corpus with the sentiments. After that, all content and behavioral based features proposed in Section 4.3.1 were computed for fake review detection. Then, we split the dataset into train and test sets. The oversampling technique called SMOTE [180] (Synthetic Minority Over-sampling Technique) was applied on the train and test sets to make a balanced distribution among the classes, fake and authentic. As the number of fake reviews is significantly lower than the number of authentic reviews. This imbalance can lead machine learning models to perform poorly, especially on the minority class, fake reviews, as they tend to be biased toward the majority class, authentic reviews. SMOTE works by creating synthetic samples for the minority class. It does this by selecting an instance from the minority class and identifying its nearest neighbors. A random neighbor is then chosen, and a synthetic sample is generated by interpolating between the selected instance and the chosen neighbor. This process is repeated for

multiple instances in the minority class, resulting in an augmented dataset with a more balanced class distribution. The final distribution of the train and test sets is illustrated in Table 22.

Table 21: The distribution of the training and testing set for supervised fake review detection.

Set	Fake	Authentic	Total
Training	13504	13476	26980
Testing	3359	3387	6746
Total	16863	16863	33726

b) FRD Model Development using SML

The balanced train set in this stage is fed into different machine learning models for training. We employed various machine learning models in these experiments, including Decision Tree, Random Forest, Gradient Boosting, Light Gradient-boosting Machine, and Extreme Gradient Boosting. Decision Tree and Random Forest models, renowned for their efficacy in fake review detection, have been extensively utilized in prior studies. The remaining three models, recognized for their effectiveness in diverse domains, were evaluated for their applicability in fake review detection. Subsequent paragraphs will offer a concise overview of the employed models. Note that the background on the Decision Tree model is elaborated in ‘*Traditional Machine Learning SA Models*’ section.

• Random Forest (RF)

RF is an ensemble learning method that builds multiple decision trees and merges their predictions. It introduces randomness by training each tree on a random subset of the features and data points, reducing overfitting and increasing the model's robustness [177].

- **Gradient Boosting (GB)**

GB builds an ensemble of decision trees sequentially, with each tree correcting the errors of the previous one. It minimizes a loss function by adding weak learners [181].

- **Light Gradient-boosting (LGBM)**

LGBM is a gradient boosting framework designed for efficient and distributed training. It uses a histogram-based approach for tree building, making it faster and more memory-efficient than traditional methods [182].

- **Extreme Gradient Boosting (XGB)**

XGB is an optimized and scalable implementation of gradient boosting. It includes features like parallelization, tree pruning, and regularization to enhance performance [183].

c) SML FRD Model Optimization

In order to optimize the performance of the models, we finetuned the models' parameters using Grid Search Cross-Validation (GridSearchCV), provided by Scikit-learn library [178]. GridSearchCV works by defining a grid of hyperparameter values to explore. The algorithm then performs an exhaustive search over this grid, training and evaluating the model for each combination of hyperparameters using cross-validation. The combination of hyperparameters that yields the highest performance score across the cross-validated folds is selected as the optimal set. The optimal hyperparameters help to enhance a model's generalization and predictive capabilities on new or unseen data. The application of grid search in this experiment, detailed in Tables 22-26, involved a careful consideration of various parameters to optimize each model. The exploration encompassed a variety of parameter values, reflecting a deliberate effort to fine-tune and enhance the performance of the models. In Table 23, dedicated to the decision tree model, the choice of 'entropy' as

the criterion was driven by the model's need for better information gain. Opting for 'None' as the splitter parameter indicated that the decision tree performed optimally without specific splitting strategies. The absence of a maximum depth restriction ('None') was preferred to allow the model to naturally adapt to the complexity of the dataset. Notably, no specific values were identified for max features, min samples split, and min samples leaf, suggesting that the default settings or a broader exploration may be suitable for these parameters. Moving to Table 24, which outlines parameter tuning for the random forest model, the selection of 'entropy' as the criterion mirrored the decision tree model. Opting for 'None' in the number of estimators parameter conveyed the model's ability to adapt to the dataset without a predefined fixed number. No specific values were identified for max depth and max features, indicating the robustness of the random forest model across a range of settings. In Table 25, dedicated to the gradient boosting model, the grid search pinpointed 750 as the optimal number of estimators. This choice reflects a careful trade-off between model performance and computational efficiency, ensuring a balance in learning from the training data. A learning rate of '0.1' was chosen to moderate the impact of each tree on the model, preventing overfitting. The selection of a maximum depth of '9' and 'auto' as the max features parameter indicated a preference for a moderately deep tree with automatic determination of feature importance. Table 26, focusing on the light gradient boosting model, reiterated the preference for 750 estimators and a learning rate of '0.1' for similar reasons as the gradient boosting model. Optimal performance was observed with a maximum depth of '9,' emphasizing the model's adaptability to this specific depth. Lastly, in Table 27, outlining the parameter tuning for the extreme gradient boosting model, the choice of '700' as the optimal number of estimators reflected the model's sensitivity to

an adequate but not excessive number of boosting rounds. A learning rate of '1' signaled a preference for aggressive learning from each tree. The absence of a maximum depth restriction ('None') indicated that the model performed optimally without a pre-defined limit on tree depth. The grid search did not identify a specific boosting type, suggesting that the default 'gbtree' boosting type was adequate for this model.

Table 22: The tested values for parameter tuning for decision tree model.

Parameter	Values	Optimal Parameter
criterion	[gini, entropy]	entropy
splitter	[best, random]	None
max_depth	[1, 2, 3, 4, 5, None]	None
max_features	[auto, sqrt, log2, None]	None
min_samples_split	[1, 2, 3, 4, 5]	None
min_samples_leaf	[1, 2, 3, 4, 5]	None

Table 23: The tested values for parameter tuning for random forest model

Parameter	Values	Optimal Parameter
criterion	[gini, entropy]	entropy
n_estimators	[50, 60, 70, 80, 90, 100, 110, 120, 130]	None
max_depth	[None, 1, 2, 3, 4, 5]	None
max_features	[sqrt, log2, None]	None

Table 24: The tested values for parameter tuning for gradient boosting model.

Parameter	Values	Optimal Parameter
n_estimators	[400, 450, 500, 550, 600, 650, 700, 750]	750
learning_rate	[0.001, 0.01, 0.1, 1]	None
max_depth	[None, 1, 2, 3, 4, 5, 6, 7, 8, 9]	9
max_features	[auto, sqrt, log2, None]	auto

Table 25: The tested values for parameter tuning for light gradient boosting model.

Parameter	Values	Optimal Parameter
n_estimators	[400, 450, 500, 550, 600, 650, 700, 750, 1000, 1200]	750
learning_rate	[1, 0.1, 0.01, 0.001]	0.1
max_depth	[5, 6, 7, 8, 9, 10, 15, 20]	9

Table 26: The tested values for parameter tuning for extreme gradient boosting model.

Parameter	Values	Optimal Parameter
n_estimators	[200, 250, 300, 350, 450, 500, 550, 600, 650, 700]	700
learning_rate	[1, 0.1, 0.01, 0.001]	1
max_depth	[-1, 1, 2, 3, 4, 5, 6]	None
boosting_type	[gbdt, dart, goss, rf]	None

d) SML FRD Model Evaluation

The evaluation of fake review detection models was conducted based on the four metrics, accuracy (A), precision (P), recall (R), and F-score (F) using the binary classification on the fake class. The calculation of the precision (P), recall (R), and F-score (F) metrics were previously illustrated in (Eq.1-3) and accuracy (A) is calculated as follows:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (\text{Eq.4})$$

TN: True Negative. TP: True Positive. FN: False Negative FP: False Positive.

Accuracy (A): Accuracy represents the proportion of correctly classified fake instances over the total number of instances (see Eq.4).

The results of the evaluation are illustrated in Table 27 before and after the hyperparameters tuning. Before hyperparameter tuning, each model exhibited specific strengths and weaknesses. Notable results include the DT model, where hyperparameter tuning led to a marginal increase in Precision. The RF model showcased a slight decrease in Precision but an improvement in Recall, resulting in a comparable F1 Score. Contrastingly, the ensemble of gradient boosting models, including GB, XGB, and LGBM, witnessed substantial performance boosts across all metrics following hyperparameter tuning. This unified

improvement underscores the effectiveness of optimization efforts across these gradient boosting algorithms.

Table 27: The supervised fake review detection results.

Model	Parameters Tuning	Precision	Recall	F1	Accuracy
DT	Before	0.8045	0.8998	0.8495	0.8383
	After	0.8069	0.8968	0.8495	0.8387
RF	Before	0.8621	0.8916	0.8766	0.8727
	After	0.8507	0.9056	0.8773	0.8715
GB	Before	0.7333	0.6885	0.7102	0.7149
	After	0.8094	0.8317	0.8204	0.8153
XGB	Before	0.7849	0.8393	0.8112	0.8018
	After	0.8436	0.9158	0.8782	0.8712
LGBM	Before	0.7705	0.8054	0.7875	0.7796
	After	0.8326	0.9037	0.8667	0.8610

Table 29 presents a comparative analysis of supervised fake review detection results, showcasing the performance of models with and without the incorporation of proposed sentiment-based features. Figure 12 visually depicts the models' performance based on the F-score, showing how the inclusion and exclusion of sentiment-based features influence their effectiveness. Across the various machine learning models, when sentiment features are included, there is a general improvement in Precision, Recall, F1, and Accuracy metrics. Notably, the RF model stands out as it experiences substantial enhancements in all these performance metrics. The DT model also demonstrates improvement, albeit marginal, particularly in terms of Precision and Recall. Although GB, XGB, and LGBM, models exhibit consistent performance, the addition of sentiment features generally maintains or improves their overall predictive capabilities. This positive trend emphasizes the value of incorporating sentiment-based features, especially for models like RF, which

benefit significantly across multiple evaluation metrics. These results underscore the potential of sentiment features to contribute positively to the models' ability to classify and predict outcomes.

Table 28: Supervised fake review results with and without sentiment-based features.

Model	Sentiment Features	Precision	Recall	F1	Accuracy
DT- Baseline	With	0.8045	0.8998	0.8495	0.8383
	Without	0.7942	0.8922	0.8404	0.8280
RF- Baseline	With	0.8621	0.8916	0.8766	0.8727
	Without	0.8446	0.8892	0.8663	0.8608
GB- Baseline	With	0.7333	0.6885	0.7102	0.7149
	Without	0.7318	0.6596	0.6938	0.7047
XGB- Baseline	With	0.7849	0.8393	0.8112	0.8018
	Without	0.7822	0.8375	0.8089	0.7993
LGBM- Baseline	With	0.7705	0.8054	0.7875	0.7796
	Without	0.7693	0.7981	0.7834	0.7762

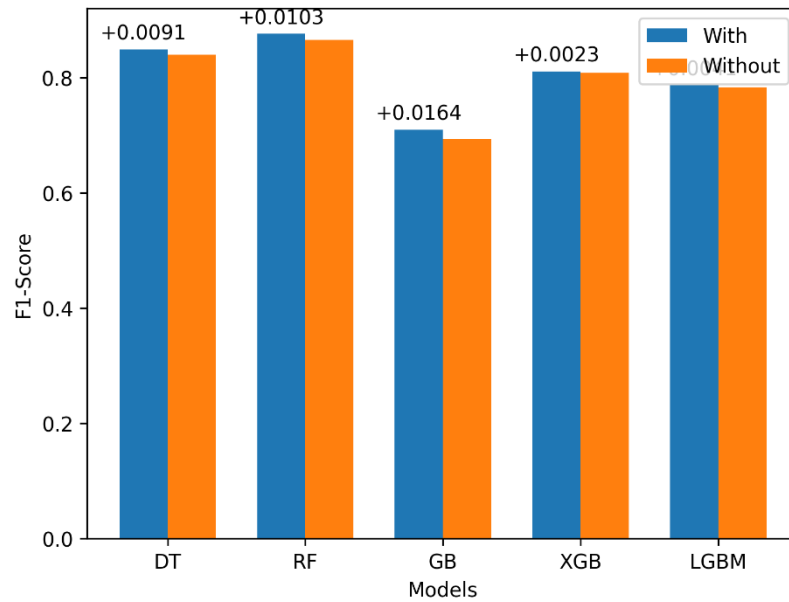


Figure 12: The Performance of the models with and without including the sentiment-based features.

Moreover, to examine the hypothesis asserting that the inclusion of proposed sentiment-based features enhances the model's performance in detecting fake reviews, we conducted a paired sample t-test. This statistical method systematically evaluates the impact of incorporating sentiment-based features into the feature set of five models focused on fake review detection. The paired observations, representing the model's performance with and without sentiment features, were compared using the `ttest_rel()` function from the `scipy.stats` library [184]. The F1 score served as the key metric to quantify performance improvements resulting from the inclusion of sentiment features, with a selected significance level (alpha) of 0.05. The results of the test revealed a t-statistic value of 3.3965 and a P-Value of 0.0274 as illustrated in Table 30. The uniformly low p-value, below the conventional threshold of 0.05, strongly supports the rejection of the null hypothesis. This demonstrates that the inclusion of sentiment-based features significantly enhances the models' capability to accurately identify fake reviews, as reflected in the elevated F1 scores.

Table 29: Paired sample t-test results.

Test Statistic	P-Value	Significance Level	Null Hypothesis
3.3965	0.0274	0.05	Reject

Furthermore, to assess the efficacy of the models in detecting fake reviews with our proposed feature set, we conducted a comparative analysis against a previous study that employed a different feature set on the same dataset [125]. The feature set utilized in the referenced study encompassed content-based and behavioral-based features. Content-based features included TF-IDF, LDA, Word2Vec, bigrams, and sentiments, while

behavioral-based features covered social, personal, reviewer behavior, and trusting features. The study excluded content-based features due to their low performance with baseline models, asserting that textual content does not serve as a reliable indicator for fake reviews. Instead, they relied on behavioral-based features for predictions, employing decision trees, random forest, logistic regression, gaussian naive bayes, and AdaBoost models. The evaluation was based on F-scores, with Table 31 illustrating the outcomes for our proposed feature set compared to theirs. Notably, our proposed feature set, when applied to baseline decision tree and random forest models, outperformed their counterparts using the different feature set and outperformed the best performing model, AdaBoost. Furthermore, all our fine-tuned models exhibited significant improvements surpassing all their models. This underscores the valuable contribution of our proposed feature set in enhancing the prediction of fake reviews on the Electronic Retailers Yelp dataset. Figure 13 provides a visual representation of feature importance, as assessed by the XGB model. Notably, three sentiment-based features, SRIC, SRI, and PSI, emerge within the top 10 most crucial features. Strikingly, SRIC stands out as the single most influential feature in discerning fake reviews. This finding contradicts the results reported in a prior study [125], which asserted that sentiment-based features did not enhance model performance.

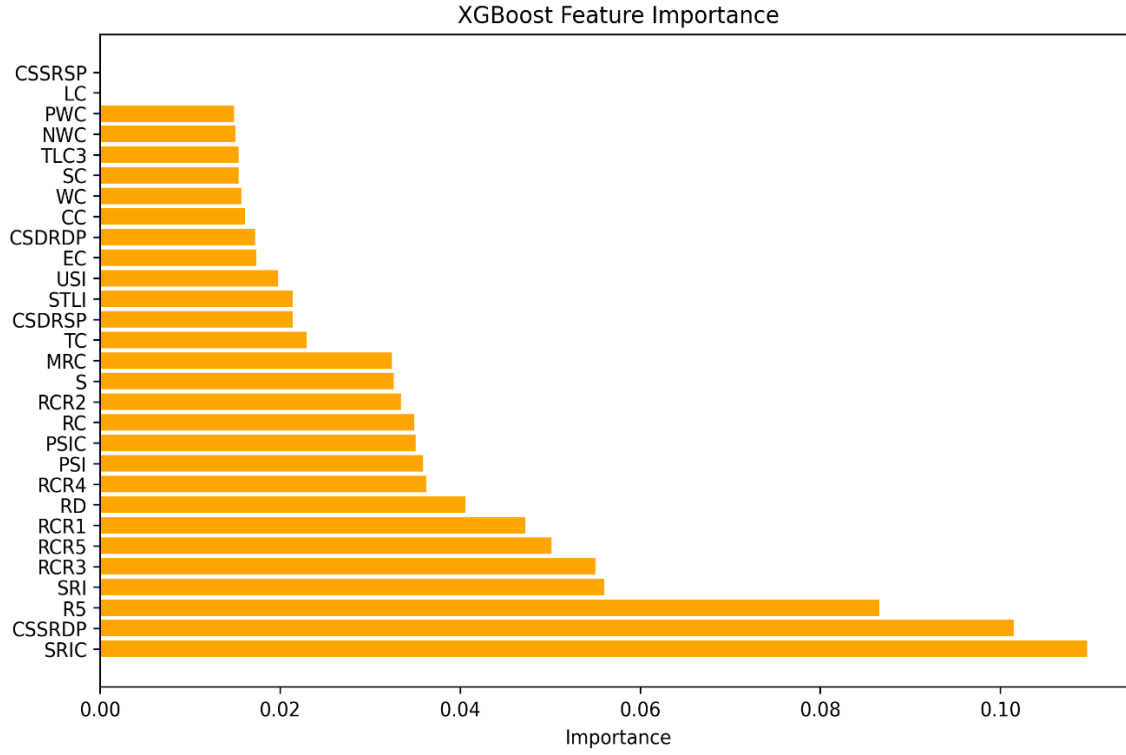


Figure 13: The XGB feature importance.

Table 30: Comparing the results of supervised models from the study with a previous study on the same dataset.

Study	Model	F1-Score
[125]	DT	0.80
	RF	0.80
	LR	0.77
	GNB	0.71
	AB	0.81
Our study	DT-Baseline	0.85
	RF-Baseline	0.88
	GB-Baseline	0.71
	XGB-Baseline	0.81
	LGBM-Baseline	0.79
	DT-Finetuned	0.85
	RF-Finetuned	0.88
	GB-Finetuned	0.82
	XGB-Finetuned	0.88
	LGBM-Finetuned	0.87

In a more extensive evaluation of our proposed feature set across diverse domains, we conducted experiments using the YelpZip dataset, comparing our approach with that of [156], which employed the same dataset for fake review detection. The YelpZip dataset comprises 608 K reviews for hotels and restaurants. This dataset is considered as a benchmark for fake review detection. We randomly selected 23,844 reviews covering 200 products, while [156] chose 24,539 reviews, with 12,898 fake reviews and 11,641 authentic reviews. We applied the same methodology used for preparing the Electronic Retailers Yelp dataset to our extracted YelpZip reviews. Additionally, to ascertain the sentiment of the reviews, we adopted the approach used in [156], leveraging SentiWordNet 3.0 to identify sentiments and calculate the number of positive and negative words. The distribution of the train and test sets for the oversampled YelpZip dataset is presented in Table 32.

Table 31: The distribution of the training and testing set for supervised fake review detection using YelpZip.

Set	Fake	Authentic	Total
Training	16428	16591	33019
Testing	4209	4046	8255
Total	20637	20637	41274

Study [156] achieved the highest accuracy using the full feature set, combining content-based and behavioral-based features, along with three proposed sentiment inconsistency features, employing the Random Forest model. Table 33 presents the results from our proposed feature set and [156] feature set, utilizing the same model on the same dataset, employing the baseline Random Forest without hyperparameter tuning for a more accurate comparison. The outcomes indicated superior performance of the Random Forest model

when equipped with our proposed feature set. This underscores the versatility of our feature set across diverse domains and underscores its robustness in effectively identifying fake reviews.

Table 32: Comparing the results of supervised models from the study with a previous study on the same dataset.

Study	Model	Accuracy	Precession	Recall	F1-Score
[156]	RF	0.929	0.936	0.928	0.932
Our study	RF-Baseline	0.964	0.996	0.931	0.963

2) Semi-supervised Machine Learning (SSML)

Identifying fake reviews involves detecting explicit and implicit instances. Explicit fake reviews are easily discernible through raw data features, while implicit ones pose a greater challenge for detection. In our pursuit of identifying fake reviews within unlabeled datasets from Saudi Amazon and AliExpress, we concentrated in the initial stage on the detection of explicit fake reviews. This involves employing a set of predefined rules and guidelines. Subsequently, in the second stage, our focus shifts to training semi-supervised models to uncover implicit fake reviews that may have eluded detection in the initial stage. This two-stage approach is inspired by a previous study [185], which successfully utilized a dataset of fake reviews in hotels and restaurants to identify fake reviews within unlabeled game reviews.

The upcoming sections will elucidate the detailed methodology followed in each stage to effectively identify and distinguish fake reviews.

- **Stage 1: Detecting explicit fake reviews**

Identification of explicit fake reviews is conducted through a comprehensive analysis across three distinct perspectives:

1. Content Similarity:

A review is considered fake using the content similarity based on the following rules:

- The maximum cosine similarity of the review with the reviews of the same reviewer on the same product exceeds 0.85.
- The maximum cosine similarity of the review with the reviews of the same reviewer on different products exceeds 0.85 and the review consists of more than two words.
- The maximum cosine similarity of the review with the reviews of different reviewers on the same product exceeds 0.85 and the review consists of more than three words or the number of reviews with that similarity was more than 2 reviews.
- The maximum cosine similarity of the reviews of different reviewers on different products exceeds 0.95 and the review consists of more than three words.

The selection of these thresholds is grounded in experiments conducted on the two datasets.

2. Review Content:

The authenticity of a review is assessed by examining its content. A review is deemed fake or non-review if it:

- Contains links or emails, using link and email count feature.
- Appears to be a question, by detecting the reviews that contain question marks or question words and the rating value is not 3.
- Is a review lacking firsthand experience with the product, by searching in the reviews with neutral sentiment and the rating value is not 3.

3. Contradictory Elements:

The presence of contradictory elements to identify fake or untruthful reviews. A review is considered fake if:

- The rating is 5, but the sentiment of the review is negative.
- The rating is 1, but the sentiment of the review is positive.

Every explicit fake review that was identified underwent a manual verification process to ensure its unequivocal classification as a fake review. This manual inspection was conducted to validate and confirm the accuracy of the rule-based identification process. The number of identified explicit fake review in Saudi Amazon dataset was 99 reviews while in AliExpress dataset we detected 227 reviews. The bellow tables, Table 34, Table 35, Table 36, Table 37, Table 38 and Table 38 show some of the identified explicit fake review in each perspective. The reviews in Table 34 reveal instances of similar reviews by the same reviewer on the same product. While instances of duplicate feedback may be expected, the repetition of overly negative reviews by the same reviewer on the same product suggests a deliberate effort by the reviewer to diminish the overall rating of the product. Within Table 35, a pattern emerges as reviews from a single reviewer appear strikingly similar across different products, and the reviews consist of more than two words. This observation indicates a consistent approach or bias in the reviewer's assessments, supporting the decision to classify these reviews as explicit fake reviews. Table 36 brings to light a peculiar phenomenon where different reviewers express remarkably similar reviews on the same product. This discovery strongly suggests a deliberate effort, possibly orchestrated by the product owner, to hire spammers who generate excessively positive reviews with the intention of influencing the overall rating of the product. Table 37 identifies some examples of non-reviews presented as questions and deviating from a neutral rating of 3. These entries, characterized by inquiries rather than authentic reviews, indicate that the reviewer may not have purchased the product and is

seeking information only. Considering these entries as authentic reviews can be misleading, as they contribute to a skewed overall product rating. The departure from a neutral rating of 3, coupled with the absence of genuine product experiences, underscores inaccuracies in the overall assessment of the product. Table 38 presents additional examples of non-reviews where the reviews lack specific product experiences and deviate from the standard neutral rating of 3. Table 39 exposes reviews containing contradictory elements, such as positive ratings accompanied by negative sentiments or vice versa. These inconsistencies raise questions about the reliability of reviewer opinions, and as such, these reviews should not be considered as authentic reviews. No reviews found contain links or emails, and no similar reviews were found of different reviewers on different products in the two datasets.

Table 33: Similar reviews by the same reviewer on the same product.

Reviewer ID	Product ID	Time	Review
2olye3QjUygfeHswjtM2CQ	1.01E+15	2/4/20 23	تحذير الدعاية الكاذبة! المعلن عنها على أنها شحن مجاني. الملصق الذي طبعوه كان احتياليًا واضطرت إلى دفع 5.15 دولار للشحن الذي رفضوا رده! انظر الصور المرفقة.
2olye3QjUygfeHswjtM2CQ	1.01E+15	2/4/20 23	تحذير! لا تشتري! يعلنون عن الشحن المجاني. طبعت ملصق احتيالي واضطرت لدفع 5.15 دولار. لقد رفضوا رد الأموال.
2olye3QjUygfeHswjtM2CQ	1.01E+15	2/4/20 23	تحذير. يعلنون عن الشحن المجاني وقاموا بطباعة ملصق احتيالي. اضطرت لدفع 5.15 دولار لاستلامه. لقد رفضوا رد الأموال! ممارسات تجارية رهيبة! انظر الصور المرفقة.

Table 34: Similar reviews by the same reviewer on different products.

Reviewer ID	Product ID	Time	Review
AHR2HIKVMYSKABVV6FEA2Y7IU	B09ZL5BV9R	6/10/2023	يستحق مقابل السعر
AHR2HIKVMYSKABVV6FEA2Y7IU	B09ZYFS3BD	6/10/2023	يستحق مقابل السعر
AHNXMSOLH2C2UX5KPZOSUPH	B0BXQP3M4G	9/26/2023	احببته منتج رائع شكر الكم
AHNXMSOLH2C2UX5KPZOSUPH	B0BDJMFVWG	9/26/2023	احببته منتج رائع شكر الكم

Table 35: Similar reviews by different reviewers on the same product.

Reviewer ID	Product ID	Time	Review
7JF+yQBcgk7jD67bE1Kfog	1.01E+15	9/18/2022	قمة
X4d/JyzGk50feHswjtM2CQ	1.01E+15	11/1/2022	قمة
1vwD6JjCUdpED/lyya86yw	1.01E+15	11/9/2022	قمة
MI3WKo7fGZIvczbGeFqWDg	1.01E+15	11/14/2022	قمة
ymbUrTufoSEtefdqfJYNWA	1.01E+15	12/22/2022	قمة
sLNFP7aD+J4vczbGeFqWDg	1.01E+15	1/7/2023	قمة
Hi9ismCOrgEwO+DPkZZJkA	1.01E+15	2/2/2023	قمة

Table 36: Reviews with question format and ratings deviating from 3.

Rating	Review
1	ي جماعه بعد اذنكم الموبايل g4 ولا g5 وبالنسباله للي جاله النسخه ال g4 هل تم استبدالها بالفعل؟؟
1	هل يوجد ضمان
2	هل يوجد واي فاي
2	هل يدعم nfc
5	هل توجد ميزة NFC
5	هل الجهاز يدعم الشريحه المدمجه

Table 37: Reviews lacking product experience and deviating from a rating of 3.

Rating	Review
5	لم أتواصل مع البائع ، تم استلامه في نيجيريا لمدة شهر واحد ، ولم أستخذه بعد.
5	انشاء الله يكون طيب
5	انا طلبوا لسه ما جربتوش

Table 38: Reviews with contradictory elements.

Rating	Review	Sentiment
5	الجودة لم تكن على مستوى العلامة	Negative
5	انا موصلش ليا الجهاز واللة الي وصلني الشنطة والشامبو واللة واللة واللة أنا مصدومة من الي حصل	Negative
5	طلبته من أسبوع ولم يصل للان	Negative
5	صراحة رجعتها طلعت خرابانة	Negative
1	حلو وكويس وممتاز زرز	Positive
1	جميل جداً وسعره مناسب هذا المطلوب	Positive
1	حلوة	Positive
1	هاذه النووع ممتاز	Positive

We also implemented our methodology in identifying the explicit fake reviews on the two datasets: the electronic retailers Yelp dataset and YelpZip. The objective was to examine

whether both datasets categorized these reviews as fake or not. In the electronic retailers Yelp dataset, instances where the same reviewer provided similar reviews for the same store were not identified. However, approximately 88 instances of identical reviews, repeated by the same reviewers for different stores, were classified as fake. Intriguingly, 1600 instances of identical reviews under the same circumstances were classified as authentic. Table 40 shows these details. Notably, all instances of similar reviews identified by different reviewers for the same stores were labeled as fake. In cases where similar reviews were repeated by different reviewers for different stores, 3 instances were classified as fake, while 3 others with identical text were erroneously labeled as authentic. Table 41 shows Identical reviews labeled as authentic and fake reviews in the electronic retailers' YELP dataset.

Reviews containing links or emails were consistently classified as fake, with the sole exception detailed in Table 42. Additionally, all reviews exhibiting inconsistencies between the rating and the sentiment were flagged as fake.

Table 39: Identical reviews labeled as authentic reviews in electronic retailers' yelp dataset.

Rating	Reviewer ID	Store ID	Date	Review	Class
1	1.83E+18	2021136	5/26/2016	Seriously, it's not worth the savings. Pay the few extra dollars somewhere else. The time and headache you will save by choosing someone else will be worth it.	Authentic
1	1.83E+18	6674163	5/26/2016	Seriously, it's not worth the savings. Pay the few extra dollars somewhere else. The time and headache you will save by choosing someone else will be worth it.	Authentic
1	1.83E+18	8044395	5/26/2016	Seriously, it's not worth the savings. Pay the few extra dollars somewhere else. The time and headache you will save by	Authentic

Rating	Reviewer ID	Store ID	Date	Review	Class
				choosing someone else will be worth it.	
5	1.63E+17	1234860	3/31/2013	A great store for all of your mobile needs, be it repairs, new phones, used phones, rentals, unlocked needs, international needs etc. The owner and staff are extremely friendly and knowledgeable. There are many shops specializing in wireless needs in SF, but you can stop your search here and you will be in great hands. The team here is great at trouble-shooting and problem-solving, and no mobile phone issue or need is beyond their abilities. I will be back whenever needs arise.	Authentic
5	1.63E+17	2753342	3/31/2013	A great store for all of your mobile needs, be it repairs, new phones, used phones, rentals, unlocked needs, international needs etc. The owner and staff are extremely friendly and knowledgeable. There are many shops specializing in wireless needs in SF, but you can stop your search here and you will be in great hands. The team here is great at trouble-shooting and problem-solving, and no mobile phone issue or need is beyond their abilities. I will be back whenever needs arise.	Authentic

Table 40: Identical reviews labeled as authentic and fake reviews in electronic retailers' yelp dataset.

Rating	Reviewer ID	Store ID	Date	Review	Class
1	2.12E+18	6365431	1/15/2013	WORST customer service ever. Ordered a washing machine with installation for last Sunday, delivery people showed up missing a part (after nonstop complaining that I lived on the 3rd floor - this is New York City, what the hell?) and after I was promised delivery on the next possible date ("guaranteed first stop"), I have now been on the phone all morning because apparently the order never went through. So now I have to take ANOTHER day off work -- and I am a teacher -- for PC Richard's multiple mistakes.	Authentic
1	8.90E+18	8424652	1/15/2013	WORST customer service ever. Ordered a washing machine with installation for last Sunday, delivery people showed up missing a part (after nonstop complaining that I lived on the 3rd floor - this is New York City, what the hell?) and after I was promised delivery on the next possible date ("guaranteed first stop"), I have now been on the phone all morning because apparently the order never went through. So now I have to take ANOTHER day off work -- and I am a teacher -- for PC Richard's multiple mistakes.	Fake
5	7.42E+18	9434397	7/21/2015	Went to PC on Flatbush Avenue for Weber grill needed to take with me for a pool party that night. Since it was right after the 4th of July, none of the stores in my neighborhood had them available and I had tons of food to grill.. I was desperate so I walked in and mgr Steve sold me one right off the floor prebuilt and had it delivered within 2 hrs !What a life saver.. The grill worked perfectly.They have a small but acceptable parking lot and felt a nice vibe from all the workers... I will shop there again and tell my friends that those people care about they're customers.. And they have a nice bedding dept too that I wasn't aware of.	Authentic

Rating	Reviewer ID	Store ID	Date	Review	Class
4	7.51E+17	5376268	10/27/2015	Went to PC on Flatbush Avenue for Weber grill needed to take with me for a pool party that night. Since it was right after the 4th of July, none of the stores in my neighborhood had them available and I had tons of food to grill.. I was desperate so I walked in and mgr Steve sold me one right off the floor prebuilt and had it delivered within 2 hrs !What a life saver.. The grill worked perfectly.They have a small but acceptable parking lot and felt a nice vibe from all the workers... I will shop there again and tell my friends that those people care about they're customers.. And they have a nice bedding dept too that I wasn't aware of.	Fake

Table 41: A review contains an email labeled as authentic review in electronic retailers yelp dataset.

Rating	Reviewer ID	Store ID	Date	Review	Class
1	7.41E+18	6936289	12/17/2015	These people working are full of it , the manager jeremy or jermy I don't remember his exact name , he told me he couldn't help then he took me out side the store and he said we could do something "under the table" , he told me to leave it on him that he would take care of it , I don't trust people so I told him no , I said no I won't give u any cash , he was like "K" can't help you then , pay 460\$ the way u can get a new phone , if Verizon wants to contact me my email is Edgardo.santacruz5@gmail...	Authentic

For the YelpZip dataset, similar patterns were observed. No instances were found where the same reviewer provided similar reviews for the same restaurant or hotel. Similarly, no cases were identified where the same reviewer wrote similar reviews for different restaurants or hotels. Nevertheless, 19 instances of identical reviews identified by different reviewers for the same restaurants or hotels were labeled as fake, while 22 instances under

the same circumstances were classified as authentic (refer to Table 43 for illustration) and the same reviews sometimes classified as authentic and sometimes classified as fake as shown in Table 44. All instances of similar reviews by different reviewers for different restaurants or hotels were labeled as fake. Notably, this dataset does not contain any reviews with links or emails. Regarding reviews with contradictory elements, 253 instances with rating-to-sentiment inconsistencies were classified as authentic. This discrepancy was attributed to an error in sentiment annotation using the lexicon-based method with Senti-Wordnet 3, as the reviews themselves exhibited consistent ratings and sentiments.

Table 42: Identical reviews labeled as authentic reviews in YelpZip dataset.

Rating	Reviewer ID	Restaurant ID	Date	Review	Class
4	10028	3876	1/4/2015	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	82578	3876	12/20/2014	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	82585	3876	12/23/2014	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	82586	3876	12/25/2014	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	82594	3876	12/15/2014	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	82629	3876	12/14/2014	This is the best place in Philadelphia to get a cheesesteak I always go cheese wiz wit onions	Authentic
4	245993	4558	9/26/2014	first time i tasted i was introduce to a new level of steaks.	Authentic
5	245998	4558	9/26/2014	first time i tasted i was introduce to a new level of steaks	Authentic

Rating	Reviewer ID	Restaurant ID	Date	Review	Class
5	246000	4558	9/26/2014	first time i tasted i was introduce to a new level of steaks.	Authentic
5	246009	4558	9/26/2014	first time i tasted i was introduce to a new level of steaks.	Authentic

Table 43: Identical reviews labeled as authentic and fake reviews in YelpZip dataset.

Rating	Reviewer ID	Restaurant ID	Date	Review	Class
4	233216	4630	2/9/2014	Dinner on Saturday night was great! Grabbing a couple delicious margaritas at the bar during happy hour was a great way to start---strong, delicious, inexpensive, and fast!! Dinner was very good, though nothing to rave about. Atmosphere was fun and intimate and the service was great! I'd absolutely return--on a busy weekend night or on a weekday. There are lots of bars nearby as well-good find!!	Fake
4	248216	4630	2/9/2014	Dinner on Saturday night was great! Grabbing a couple delicious margaritas at the bar during happy hour was a great way to start---strong, delicious, inexpensive, and fast!! Dinner was very good, though nothing to rave about. Atmosphere was fun and intimate and the service was great! I'd absolutely return--on a busy weekend night or on a weekday. There are lots of bars nearby as well-good find!!	Authentic
4	118422	3876	1/9/2015	I was there last night and I taste those cheesestake	Fake

Rating	Reviewer ID	Restaurant ID	Date	Review	Class
				and it was super delicious, that I'm going back now. just love it.!!	
4	82658	3876	1/5/2015	I was there last night and I taste those cheesestake and it was super delicious, that I'm going back now. just love it.!!	Fake
4	117999	3876	1/1/2015	I was there last night and I taste those cheesestake and it was super delicious, that I'm going back now. just love it.!!	Authentic

Applying our methodology to identify explicit fake reviews within the two datasets, the electronic retailers Yelp dataset and YelpZip, revealed a significant discrepancy in the classification of certain reviews. Despite being labeled as authentic in the dataset, a closer investigation of these reviews indicated that they were unequivocally explicit fake reviews. This discrepancy underscores the limitations of relying solely on labeled datasets for fake review detection.

Our methodology, designed to meticulously analyze various aspects of reviews, has proven effective in unveiling explicit fake reviews that might otherwise go unnoticed. This finding emphasizes the importance of leveraging advanced methodologies to enhance the accuracy of fake review identification. By employing our approach, we can improve the precision and reliability of detecting explicit fake reviews, providing a valuable contribution to the field of fake review detection.

- **Stage 2: Detecting implicit fake reviews**

To detect the implicit fake reviews that were not detected in the first stage on the unlabeled datasets, we utilized two techniques of semi-supervised learning, self-training, and positive unlabeled learning. The following sections will describe the workflow of each technique.

- a) Self-training FRD Model using SSML**

Self-training [186] is a semi-supervised learning approach that iteratively improves model performance by using its own predictions to label unlabeled data. Initially, as illustrated in in Figure 14, the model is trained on an oversampled train set of the labeled data that compromises the labeled reviews from electronic retailers' yelp dataset and the explicit fake reviews from Saudi Amazon and AliExpress datasets after calculating the content and behavioral-based features. Then, it makes predictions on the unlabeled data of Saudi Amazon and AliExpress datasets, and the most confident predictions are added to the labeled set for the next iteration. This iterative process continues, with the model being retrained on the updated labeled dataset in each iteration. Finally, the model is evaluated on the testing set of the labeled using four evaluation metrics, accuracy, precision, recall, and F1 score. The following sections will illustrate the detailed training process of the models using this approach.

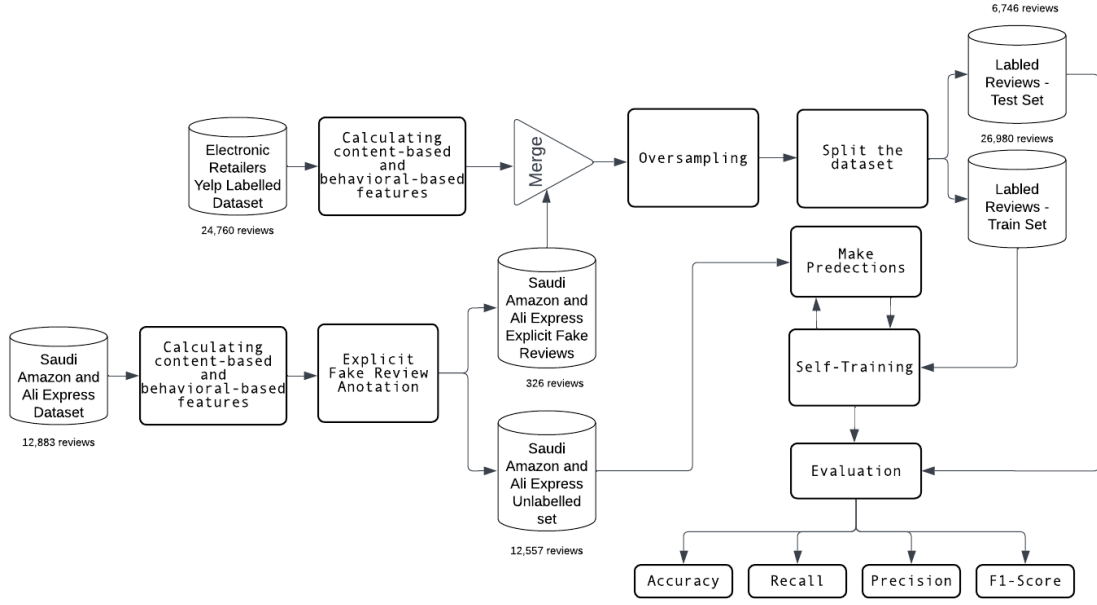


Figure 14: Fake review detection using self-training approach.

i. Data Preparation

The same process undertook to prepare the oversampled electronic retailers yelp dataset in the supervised learning that discussed in ‘*Data Preparation for Fake Reviews Detection using SML*’ section also applied here. Additionally, the unlabeled reviews from the two datasets, Saudi Amazon and AliExpress, after removing the explicit fake reviews from stage 1 were combined to form one dataset of 12,557 unlabeled reviews.

ii. Self-training FRD Model Development using SSML

The two baselines machine learning models, Decision Trees and Random Forest, provided by Scikit-learn library [178] were employed for this task. The initial stage of the model development involves training each model on the oversampled training set of the labeled dataset. Following this, three self-training iterations unfold, during which the model generates predictions on the unlabeled dataset. High-confidence predictions, determined by a confidence threshold of 0.8, are filtered. The selection of this threshold was informed by a grid search over a range of threshold values. The threshold that yielded the optimal

performance, measured by the F-score value, was chosen. Instances meeting this confidence criterion, where the predicted confidence is 0.8 or higher, are then added to the labeled dataset. After each self-training iteration, the model is retrained on the expanded labeled dataset, and its performance is assessed on the oversampled labeled test set.

iii. Self-training FRD Model Evaluation

The final evaluation of the model is conducted on the labeled test set after completing the self-training iterations.

b) PU Learning FRD Model using SSML.

Positive Unlabeled Learning, often referred to as PU learning, is a semi-supervised approach designed for binary classification tasks where the majority of the data is unlabeled, and only positive examples are labeled [187]. The PU learning goal is to learn a classifier that can distinguish between positive and unlabeled instances. In scenarios where obtaining negative examples is challenging or expensive, PU learning becomes valuable. It assumes that the unlabeled data contains both negative and positive instances, but only the positive instances are identified. PU learning can be applied using various techniques such as standard classifier, PU bagging, and two step approach. We utilized PU bagging techniques to train the model to detect the fake reviews. PU bagging works by creating a bootstrapped dataset and applying the standard classifier to out-of-bag samples to finally arrive at the estimated probability of unlabeled samples to be positive. So, the goal of PU Bagging is to improve the performance of a classifier on positive instances by iteratively training multiple classifiers on different subsets of the data. Figure 15 illustrates the approach of the PU-learning models which shows similar steps to the self-training approach, however, here the PU-bagging models utilize only the fake reviews from the

train set. The following four subsections address the PU Learning FRD Model using SSML through data preparation, model development, experiments, and model evaluation.

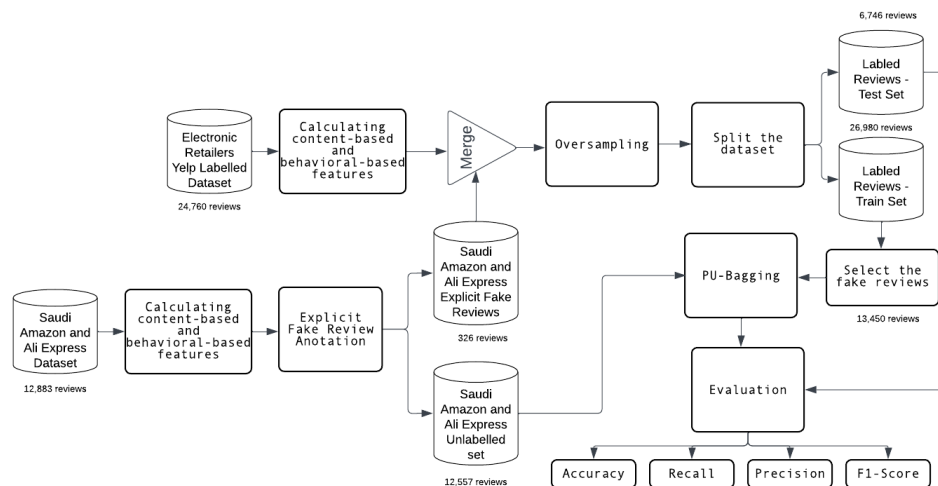


Figure 15: Fake review detection using PU-learning approach.

I. Data Preparation

In Positive-Unlabeled (PU) learning, training exclusively incorporates positive instances from the labeled set, while all unlabeled reviews are treated as negative instances. The positive instances in our case are the fake reviews and the negative instances are unlabeled reviews. To achieve this, we initially merge the unlabeled reviews from two datasets, Saudi Amazon and AliExpress, following the exclusion of explicit fake reviews identified in the initial stage with the labeled dataset. The combined dataset is then shuffled to ensure randomization and all instances in this dataset are initially assigned a class label of 0, indicating a negative instance. In another hand, all the positive instances has a class label equal to 1, representing the fake reviews.

II. PU-learning FRD Model Development using SSML

Model development and training phase initiate with the setup of the classifier. Like the two classifiers employed in the experiments of semi-supervised self-training, identical

classifiers are utilized for PU Bagging. A sequence of 15 iterations is executed to train the model on diverse subsets of the training data. This number of iterations was chosen after training the models using three different numbers of iterations, namely 5, 10, and 15, and the best performance on the testing datasets for the experiment discussed in the *‘Additional Experiments’* section was achieved by using 15. In each iteration, a new model is trained on the set of positive instances (instances belonging to the fake class) along with an equal number of randomly sampled unlabeled instances (instances with unknown labels). After training, the model is used to make predictions on the remaining unlabeled instances that were not included in the training subset for that iteration. These predictions on the unlabeled instances are aggregated or combined across multiple iterations. A threshold is set to 80% as determined in self-training model. If the mean prediction for an instance is equal to or exceeds the threshold, it is considered a positive prediction. If the mean prediction is below the threshold, it is considered a negative prediction. Instances that consistently receive positive predictions (exceed the threshold) across multiple iterations are considered likely positives.

III. PU-learning FRD Model Evaluation

To evaluate the performance of the PU-bagging models, the same number of iterations were used to allow the 15 pretrained models to make a prediction on the testing set. Then, based on the determined threshold (80%) and the mean predictions from the 15 models the model for each instance of the testing set assigns the predicted value.

- **Additional Experiments**

As additional experiments, we trained self-training and PU learning models to detect the implicit fake reviews from Saudi Amazon and AliExpress using YelpZip dataset instead of

utilizing Electronic Retailer Yelp dataset with the same approach and settings. Furthermore, we also trained self-training and PU learning models on the two datasets, Electronic Retailer Yelp dataset and YelpZip, purely without including any reviews from any other dataset. For each of the two datasets, the original labeled dataset was divided into two sets of equal size. The first set was utilized as unlabeled dataset by setting the class labels to 0 to all the instances. While the second set considered as the labeled dataset for training and testing the model. This labeled dataset is then oversampled to create a balanced dataset for training and testing. After that, these labeled and unlabeled datasets are used as inputs for each approach as discussed in the previous sections.

- **SSML FRD Models Evaluation**

For the evaluation of fake review detection models using semi-supervised approach, the same four metrics utilized for fake review detection in supervised learning were used here. Table 45, Table 46, Table 47, and Table 47 present results of training the model to detect the implicit fake reviews from Saudi Amazon and AliExpress by utilizing the two labeled datasets, Electronic Retailer Yelp dataset and YelpZip, using the two semi-supervised approaches.

Although the domain of YelpZip dataset is not online products, the results showed that utilizing YelpZip, enhanced the performance of the prediction of fake reviews by reaching an F1-score value of 96.92% on the random forest self-training model. In the case of the Electronic Retailer Yelp dataset, Self-Training with Random Forests achieved the highest overall performance with an F1-score of 86.52%. Additionally, the models considerably maintained the same performance levels across the two approaches when utilizing the Electronic Retailer Yelp dataset. However, we can notice the huge shift in the performance

across the two approaches while utilizing YelpZip dataset. PU-Learning, particularly with Decision Trees, exhibited lower performance on the YelpZip dataset, suggesting that the effectiveness of semi-supervised learning approaches may vary based on the characteristics of the datasets used.

Table 44: Results of SSML training using the labeled Yelp dataset, unlabeled Saudi Amazon and AliExpress datasets.

Approach	Model	Accuracy	Precession	Recall	F1-Score
Self-Training	DT-Baseline	0.8226	0.7822	0.8919	0.8335
	RF- Baseline	0.8601	0.8315	0.9018	0.8652
PU-Learning	DT- Baseline	0.8251	0.7860	0.8913	0.8354
	RF- Baseline	0.8654	0.8430	0.8967	0.8690

Table 45: Results of SSML training using the labeled YelpZip dataset, unlabeled Saudi Amazon and AliExpress datasets.

Approach	Model	Accuracy	Precession	Recall	F1-Score
Self-Training	DT- Baseline	0.9294	0.8760	0.9985	0.9333
	RF- Baseline	0.9686	0.9420	0.9980	0.9692
PU-Learning	DT- Baseline	0.5260	0.5108	0.9927	0.6745
	RF- Baseline	0.4993	0.4970	0.9966	0.6632

Tables 44 and 45 present the results of SSML training using different datasets for the FRD task. In Table 44, the SSML training incorporated the labeled Yelp dataset in conjunction with unlabeled data sourced from Saudi Amazon and AliExpress. On the other hand, Table 45 focuses on SSML training using the labeled YelpZip dataset alongside unlabeled data from the same sources, Saudi Amazon and AliExpress. The two tables, Table 46 and Table 47, show that utilizing the YelpZip dataset resulted in improved performance for PU-learning models when compared to the inclusion of unlabeled instances from other datasets. Conversely, PU-learning models trained exclusively on the Yelp dataset demonstrated a decrease in performance compared to models incorporating unlabeled instances from the Saudi Amazon and AliExpress datasets alongside the Yelp dataset. In summary, the choice

of dataset and the inclusion of unlabeled instances from relevant sources play a crucial role in the performance of PU-learning models.

Table 46: Results of Semi supervised training using the labeled Yelp dataset only.

Approach	Model	Accuracy	Precession	Recall	F1-Score
Self-Training	DT- Baseline	0.8218	0.7799	0.8846	0.8289
	RF- Baseline	0.8512	0.8250	0.8821	0.8526
PU-Learning	DT- Baseline	0.7954	0.8292	0.7315	0.7773
	RF- Baseline	0.8159	0.9794	0.6361	0.7713

Table 47: Results of Semi supervised training using the labeled YelpZip dataset only.

Approach	Model	Accuracy	Precession	Recall	F1-Score
Self-Training	DT- Baseline	0.9319	0.8821	0.9976	0.9363
	RF- Baseline	0.9700	0.9463	0.9966	0.9708
PU-Learning	DT- Baseline	0.9731	0.9944	0.9517	0.9726
	RF- Baseline	0.9697	0.9860	0.9531	0.9693

For further evaluation, we compared our semi-supervised fake reviews detection models with the best performing models from the previous studies as shown in Table 48. Notably, our proposed decision trees PU-Learning model, utilizing our feature set exclusively with the YelpZip dataset, achieved the highest accuracy and F1 score. Furthermore, our random forest Self-Training model outperformed all previously reported semi-supervised models when incorporating the labeled YelpZip dataset along with two unlabeled datasets from Saudi Amazon and AliExpress. Additionally, the decision trees PU-Learning with our feature sets outperformed SpamGAN-GPT2 from a study [188] that trained on the same dataset. However, the study [188] relied only on word embeddings to identify the fake reviews. In contrast, to employing a diverse domain dataset for labeling unlabeled data, our random forest self-training model, equipped with our suggested feature set and trained on either the Yelp dataset or YelpZip dataset combined with unlabeled online product reviews,

demonstrated superior performance compared to the decision trees PU-Learning model in study [185]. This study also utilized different sentiment-based features along with stylistic, part of speech, doc2vec, and raw data features, for fake reviews identification. 50 shows that while our PU-Learning models demonstrated competitive performance, further evaluation of using Tripadvisor with our unlabeled dataset is recommended for a comprehensive assessment.

Table 48: Best-performing models using Semi-supervised approach.

Study	Dataset	Model	F1	Accuracy
[188]	YelpZip	SpamGAN-GPT2	0.698	0.673
	Tripadvisor	SpamGAN-GPT2	0.902	0.875
[126]	YelpChi	Co-training (Multi-Feature Fusion)	0.819	-
[185]	Tripadvisor Game Reviews	PU-Learning-DT	0.798	-
[189]	Tripadvisor	SpamGAN	0.868	-
[190]	Tripadvisor	Self-Training-NB	0.930	0.930
	Tripadvisor-Positive Polarity	Self-Training-NB	0.940	0.940
	Tripadvisor-Negative Polarity	Self-Training-NB	0.960	0.950
[191]	Yelp	Ramp-OCSVM	0.725	0.744
	Tripadvisor Positive Polarity	Ramp-OCSVM	0.924	0.921
	Tripadvisor Negative Polarity	Ramp-OCSVM	0.906	0.903
[192]	Tripadvisor	PU-Learning-NB	0.837	-
Our Study	ERYelp SA Amazon AliExpress	Self-Training-RF	0.876	0.872
	YelpZip SA Amazon AliExpress	Self-Training-RF	0.969	0.969
	ERYelp	Self-Training-RF	0.852	0.851
	YelpZip	PU-Learning-DT	0.973	0.973

Table 49: Best performing PU-learning models using Semi-supervised approach.

Study	Dataset	Model	F1	Accuracy
[188]	Tripadvisor	PU-Learning-NB	0.852	0.843
	YelpZip	PU-Learning-NB	0.656	0.609
[185]	Tripadvisor Game Reviews	PU-Learning-DT	0.798	-
[192]	Tripadvisor	PU-Learning-NB	0.837	-
Our Study	ERYelp SA Amazon AliExpress	PU-Learning-RF	0.675	0.526
	YelpZip SA Amazon AliExpress	PU-Learning-DT	0.674	0.526
	ERYelp	PU-Learning-DT	0.777	0.800
	YelpZip	PU-Learning-DT	0.973	0.973

CHAPTER 5

CONCLUSION

Online shopping has become a trendy way for people to purchase their products. In online shopping, customers heavily rely on the product reviews provided by other customers to adjust their purchasing choices. However, these reviews may be fake and created by spammers to mislead customers and to let them make wrong purchase decisions. Fake reviews have a serious impact on the customer shopping experience on e-commerce applications. Therefore, there is a need to automatically detect fake reviews in e-commerce applications to protect customers from being exploited by them. Although many researchers have focused on detecting fake reviews in English, there are few studies focused on the detection of fake reviews in Arabic. This thesis introduces an approach for detecting Arabic fake reviews in online product evaluations, leveraging machine and deep learning techniques coupled with sentiment analysis.

The methodology is initiated by aggregating reviews from prominent e-commerce platforms, namely Saudi Amazon and AliExpress, resulting in a dataset exceeding 25,000 reviews. Through several preprocessing steps, the corpus is refined to 12,865 reviews in both original Arabic and translated English text. After that, the sentiment analysis phase is conducted. It involves employing three traditional machine learning models, Logistic Regression, Decision Tree, and Support Vector Machine, as well as two pre-trained deep learning models, MARABERT and CAMELBERT-Mix-SA. The conducted sentiment analysis task encompasses binary and ternary classifications using a set of 10,500 authentic

reviews from Amazon after translation. The MARABERT model, exhibiting optimal performance in ternary classifications, with an 84.12% F-score value, is then employed to annotate the collected reviews with sentiments. Validation of the predicted sentiments is conducted through semi-manual verification based on review ratings. Additionally, the trained Logistic Regression model was utilized to establish lexicons comprising 2,146 terms for Arabic online product reviews.

The subsequent phase involves computing proposed sentiment-based features alongside content and behavior-based features for the collected reviews, contributing to the identification of fake reviews.

In the final phase, two annotated datasets, Electronic Retailers Yelp, and YelpZip, were used for training machine learning models for fake review detection. The Electronic Retailers Yelp dataset was translated first and then labeled with the sentiment using the same approach utilized in annotating the collected reviews, while the YelpZip dataset was annotated through a lexicon-based approach using Senti-Wordnet 3 lexicons. Supervised learning employs five machine learning models, Decision Tree, Random Forest, Gradient Boosting, Light Gradient-boosting Machine, and Extreme Gradient Boosting, on the two annotated datasets to evaluate the effectiveness of the proposed feature set. Hyperparameter tuning was also applied to enhance model performance. Several evaluations demonstrated the effectiveness of the proposed feature set in detecting fake reviews compared to prior studies. The semi-supervised approach was introduced with a rule-based approach to annotate the collected reviews with fake and authentic labels. The rule-based approach was used to identify explicit fake reviews initially. The remaining unlabeled reviews are subsequently annotated with the implicit fake reviews using two

semi-supervised models, Decision Tree and Random Forest, trained on the labeled datasets using Self-Training and PU-Learning variants. Results showcase the Random Forest Self-Training model, trained on the YelpZip dataset, achieving the highest F1 score of 96.9%, surpassing existing semi-supervised models in the literature when integrated with the proposed feature set.

5.1 Remarkable Findings

After completing the goal of the study which is conducting the fake reviews detection using sentiment analysis and machine learning techniques, this section focuses on the remarkable findings of this thesis as follows:

- The publicly labeled dataset with the sentiments for online product reviews are labeled using the rating of the review. However, there are many reviews that have contradictory sentiment with their rating.
- Most of the studies that focused on the fake review detection utilized binary sentiment classification. This loss of a lot of important aspects that help in identifying the fake reviews. Adding the neutral sentiment class helped in identifying the non-reviews.
- The conventional logistic regression and support vector machine bigram models demonstrated robust performance in ternary sentiment analysis, surpassing the pre-trained MARABERT on the Electronic Retailers Yelp dataset.
- The models generally showed higher evaluation scores in binary classification settings compared to ternary classifications. The neutral reviews frequently encompass a blend of positive and negative sentiments, introducing a challenge for precise categorization.
- There are no public labeled dataset for fake reviews in Arabic for any domain.

- There are no public labeled dataset for fake reviews in English for the online products with complete feature set for the reviewer, product and time-based features.
- Oversampling is a crucial step in fake review detection to address the limited number of instances in the fake class.
- Gradient Boosting classifiers are indeed sensitive to their hyperparameters, and selecting appropriate parameter values is crucial for achieving good performance. However, decision trees and random forest classifiers demonstrated lower sensitivity to hyperparameters compared with gradient boosting classifiers.
- The inclusion of the sentiment-based features enhanced the performance of the prediction.
- Our proposed feature set is applicable to online products, electronic stores, hotels and restaurants reviews.
- Utilizing our proposed feature set using supervised approach with the same models utilized from the previous studies, on the same dataset and without parameter optimization outperformed their models using their different feature sets.
- When applying our rule-based approach to identify the fake reviews in the public labeled datasets, Electronic Retailers Yelp and YelpZip, we found many explicit fake reviews considered as not fake in the two datasets.
- Although the number of collected reviews from Saudi Amazon was greater than the number of collected reviews from AliExpress, but the detected explicit fake reviews from AliExpress was more.
- No reviews found in both AliExpress and Saudi Amazon contains links or emails.

- The semi-supervised fake review detection models outperformed the supervised one when testing on the same dataset.
- Leveraging a labeled dataset from a different domain to annotate the unlabeled dataset using a semi-supervised approach can serve as a viable solution to address the scarcity of available data for that particular domain.
- Utilizing our proposed feature set with a semi-supervised approach outperformed the best-performing semi-supervised models from the previous studies.
- The performance of PU-learning models is significantly influenced by the selection of datasets and the incorporation of labeled instances from different sources.
- Positive-unlabeled learning is particularly helpful in scenarios where obtaining negative instances for training is challenging or expensive.

5.2 Future Work

We foresee several avenues for future research, as follows:

- Identifying potential spammers using our proposed feature set.
- Experimenting both the supervised and semi-supervised approach with our proposed feature set with the public datasets, TripAdvisor, YelpChi, and YelpNew datasets.
- Correct the incorrectly labeled reviews from the public dataset using our proposed rule-based approach to identify the explicit reviews.
- Training the labeled Saudi Amazon and AliExpress for fake review detection by adding the word embeddings, and POS features.
- Utilizing our constructed lexicons to perform sentiment analysis and compare the results with the trained sentiment analysis model,

- Designing a tool to identify fake reviews on a website and show the refined rating of the products after removing the potential fake review and their effects.

References

- [1] J. Li, P. Lv, W. Xiao, L. Yang, and P. Zhang, “Exploring groups of opinion spam using sentiment analysis guided by nominated topics,” *Expert Syst. Appl.*, vol. 171, p. 114585, Jun. 2021, doi: 10.1016/J.ESWA.2021.114585.
- [2] T. Ermakova, M. Henke, and B. Fabian, “Commercial Sentiment Analysis Solutions: A Comparative Study,” *Proc. 17th Int. Conf. Web Inf. Syst. Technol.*, pp. 103–114, Oct. 2021, doi: 10.5220/0010709400003058.
- [3] S. Pedipina, S. Sankar, and R. Dhanalakshmi, “Sentimental analysis on twitter data of political domain,” *Lect. Notes Data Eng. Commun. Technol.*, vol. 66, pp. 205–216, 2021, doi: 10.1007/978-981-16-0965-7_17/COVER.
- [4] K. Mite-Baidal, C. Delgado-Vera, E. Solís-Avilés, A. H. Espinoza, J. Ortiz-Zambrano, and E. Varela-Tapia, “Sentiment analysis in education domain: A systematic literature review,” *Commun. Comput. Inf. Sci.*, vol. 883, pp. 285–297, 2018, doi: 10.1007/978-3-030-00940-3_21/COVER.
- [5] “E-commerce market in Saudi Arabia - statistics & facts | Statista.” <https://www.statista.com/topics/7723/e-commerce-in-saudi-arabia/> (accessed Dec. 24, 2022).
- [6] “Why Do People Shop Online? [Updated Nov 2022] - Oberlo.” <https://www.oberlo.com/statistics/why-do-people-shop-online> (accessed Dec. 24, 2022).
- [7] N. Jindal and B. Liu, “Opinion Spam and Analysis,” *WSDM’08 - Proc. 2008 Int. Conf. Web Search Data Min.*, pp. 219–229, 2008, doi: 10.1145/1341531.1341560.
- [8] L. Akoglu, R. Chandy, and C. Faloutsos, “Opinion Fraud Detection in Online Reviews by Network Effects,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 7, no. 1, pp. 2–11, 2013, doi: 10.1609/ICWSM.V7I1.14380.
- [9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artif. Intell. Rev.* 2022 557, vol. 55, no. 7, pp. 5731–5780, Feb. 2022, doi: 10.1007/S10462-022-10144-1.
- [10] M. Birjali, M. Kasri, and A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends,” *Knowledge-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/J.KNOSYS.2021.107134.
- [11] S. Tammina and S. Annareddy, “Sentiment analysis on customer reviews using convolutional neural network,” *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, Jan. 2020, doi: 10.1109/ICCCI48352.2020.9104086.
- [12] S. Behdenna, F. Barigou, and G. Belalem, “Document Level Sentiment Analysis:

- A survey,” *EAI Endorsed Trans. Context. Syst. Appl.*, vol. “4,” no. 13, p. 154339, Mar. 2018, doi: 10.4108/EAI.14-3-2018.154339.
- [13] F. Alshuwaier, A. Areshey, and J. Poon, “Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review,” *Intell. Syst. with Appl.*, vol. 15, p. 200090, Sep. 2022, doi: 10.1016/J.ISWA.2022.200090.
 - [14] A. Ghallab, A. Mohsen, and Y. Ali, “Arabic Sentiment Analysis: A Systematic Literature Review,” *Appl. Comput. Intell. Soft Comput.*, vol. 2020, 2020, doi: 10.1155/2020/7403128.
 - [15] G. Choi, S. Oh, and H. Kim, “Improving Document-Level Sentiment Classification Using Importance of Sentences,” *Entropy*, vol. 22, no. 12, pp. 1–11, Mar. 2021, doi: 10.3390/e22121336.
 - [16] B. Liu, “Sentiment Analysis and Opinion Mining,” in *Synthesis lectures on human language technologies*, 1st ed., vol. 5, Cham: Springer International Publishing, 2012, pp. 1–167.
 - [17] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/J.ASEJ.2014.04.011.
 - [18] K. Schouten and F. Frasincar, “Survey on Aspect-Level Sentiment Analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016, doi: 10.1109/TKDE.2015.2485209.
 - [19] K. Amarouche, H. Benbrahim, and I. Kassou, “Product Opinion Mining for Competitive Intelligence,” *Procedia Comput. Sci.*, vol. 73, pp. 358–365, Jan. 2015, doi: 10.1016/J.PROCS.2015.12.004.
 - [20] T. Tran, H. Ba, and V. N. Huynh, “Measuring hotel review sentiment: An aspect-based sentiment analysis approach,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11471 LNAI, pp. 393–405, 2019, doi: 10.1007/978-3-030-14815-7_33/COVER.
 - [21] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, “Aspect-based sentiment analysis using adaptive aspect-based lexicons,” *Expert Syst. Appl.*, vol. 148, p. 113234, Jun. 2020, doi: 10.1016/J.ESWA.2020.113234.
 - [22] “Amazon Mechanical Turk.” Accessed: Sep. 10, 2023. [Online]. Available: <https://www.mturk.com/>.
 - [23] R. M. Duwairi, R. Marji, N. Sha’Ban, and S. Rushaidat, “Sentiment analysis in arabic tweets,” *2014 5th Int. Conf. Inf. Commun. Syst. ICICS 2014*, 2014, doi: 10.1109/IACS.2014.6841964.
 - [24] “The effects of Pre-Processing Techniques on Arabic Text Classification,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 1, pp. 41–48, Feb. 2021, doi:

10.30534/IJATCSE/2021/061012021.

- [25] I. Abu Farha and W. Magdy, “A comparative study of effective approaches for Arabic sentiment analysis,” *Inf. Process. Manag.*, vol. 58, no. 2, p. 102438, Mar. 2021, doi: 10.1016/J.IPM.2020.102438.
- [26] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, “A review of sentiment analysis research in Arabic language,” *Futur. Gener. Comput. Syst.*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/J.FUTURE.2020.05.034.
- [27] M. Avinash and E. Sivasankar, “A Study of Feature Extraction techniques for Sentiment Analysis,” *Adv. Intell. Syst. Comput.*, vol. 814, pp. 475–486, Jun. 2019, doi: 10.1007/978-981-13-1501-5_41.
- [28] H. S. Ibrahim, S. M. Abdou, and M. Gheith, “MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis,” *2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.*, pp. 353–358, Aug. 2015, doi: 10.1109/RETIS.2015.7232904.
- [29] S. Kübler and E. Mohamed, “Part of speech tagging for Arabic,” *Nat. Lang. Eng.*, vol. 18, no. 4, pp. 521–548, Oct. 2012, doi: 10.1017/S1351324911000325.
- [30] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, “OCA: Opinion corpus for Arabic,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 10, pp. 2045–2054, 2011, Accessed: Sep. 10, 2023. [Online]. Available: https://www.academia.edu/15155829/OCA_Opinion_corpus_for_Arabic.
- [31] R. Duwairi and M. El-Orfali, “A study of the effects of preprocessing strategies on sentiment analysis for Arabic text,” *J. Inf. Sci.*, vol. 40, no. 4, pp. 501–513, 2014, doi: 10.1177/0165551514534143.
- [32] M. Abdul-Mageed, M. Diab, and S. Kübler, “SAMAR: Subjectivity and sentiment analysis for Arabic social media,” *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, Jan. 2014, doi: 10.1016/J.CSL.2013.03.001.
- [33] N. Elhassan *et al.*, “Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning,” *Comput. 2023, Vol. 12, Page 126*, vol. 12, no. 6, p. 126, Jun. 2023, doi: 10.3390/COMPUTERS12060126.
- [34] S. Alhumoud, T. Albuhairi, and M. Altuwaijri, “Arabic sentiment analysis using WEKA a hybrid learning approach,” in *IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2015, vol. 1, pp. 402–408, doi: 10.5220/0005616004020408.
- [35] S. Naeem, D. Logofătu, and F. Muharemi, “Sentiment Analysis by Using Supervised Machine Learning and Deep Learning Approaches,” *Commun. Comput. Inf. Sci.*, vol. 1287, pp. 481–491, 2020, doi: 10.1007/978-3-030-63119-2_39/COVER.

- [36] B. Agarwal and N. Mittal, "Prominent Feature Extraction for Sentiment Analysis," 2016, doi: 10.1007/978-3-319-25343-5.
- [37] G. Yoo and J.-S. Nam, "A Hybrid Approach to Sentiment Analysis Enhanced by Sentiment Lexicons and Polarity Shifting Devices," 2018.
- [38] L. Lei and D. Liu, "Conducting Sentiment Analysis," *Elem. Corpus Linguist.*, Sep. 2021, doi: 10.1017/9781108909679.
- [39] W. Jiang, K. Zhou, C. Xiong, G. Du, C. Ou, and J. Zhang, "KSCB: a novel unsupervised method for text sentiment analysis," *Appl. Intell.*, vol. 53, no. 1, pp. 301–311, Jan. 2023, doi: 10.1007/S10489-022-03389-4/METRICS.
- [40] Y. Han, Y. Liu, and Z. Jin, "Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 5117–5129, May 2020, doi: 10.1007/S00521-018-3958-3/METRICS.
- [41] V. L. Shan Lee, K. H. Gan, T. P. Tan, and R. Abdullah, "Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data," *Procedia Comput. Sci.*, vol. 161, pp. 577–584, Jan. 2019, doi: 10.1016/J.PROCS.2019.11.159.
- [42] H. Zou and Z. Wang, "A semi-supervised short text sentiment classification method based on improved Bert model from unlabelled data," *J. Big Data*, vol. 10, no. 1, pp. 1–19, Dec. 2023, doi: 10.1186/S40537-023-00710-X/TABLES/4.
- [43] Y. Pan, Z. Chen, Y. Suzuki, F. Fukumoto, and H. Nishizaki, "Sentiment analysis using semi-supervised learning with few labeled data," *Proc. - 2020 Int. Conf. Cyberworlds, CW 2020*, pp. 231–234, Sep. 2020, doi: 10.1109/CW49994.2020.00044.
- [44] T. J. Lu, "Semi-supervised microblog sentiment analysis using social relation and text similarity," *2015 Int. Conf. Big Data Smart Comput. BIGCOMP 2015*, pp. 194–201, Mar. 2015, doi: 10.1109/35021BIGCOMP.2015.7072831.
- [45] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: A survey," *Neurocomputing*, vol. 531, pp. 195–216, Apr. 2023, doi: 10.1016/J.NEUCOM.2023.02.015.
- [46] A. Hamdi, K. Shaban, and A. Zainal, "CLASENTI: A class-specific sentiment analysis framework," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 4, Jul. 2018, doi: 10.1145/3209885.
- [47] A. Kumar and G. Garg, "The Multifaceted Concept of Context in Sentiment Analysis," *Adv. Intell. Syst. Comput.*, vol. 1040, pp. 413–421, 2020, doi: 10.1007/978-981-15-1451-7_44/COVER.
- [48] J. Mielke *et al.*, "Understanding dynamic complexity in context—Enriching contextual analysis in implementation science from a constructivist perspective,"

- Front. Heal. Serv.*, vol. 2, p. 953731, Jul. 2022, doi: 10.3389/FRHS.2022.953731.
- [49] A. Kotelnikova, D. Paschenko, K. Bochenina, and E. Kotelnikov, "Lexicon-based Methods vs. BERT for Text Sentiment Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13217 LNCS, pp. 71–83, Nov. 2021, doi: 10.1007/978-3-031-16500-9_7.
 - [50] V. Lin and L.-P. Morency, "SenteCon: Leveraging Lexicons to Learn Human-Interpretable Language Representations," May 2023, Accessed: Nov. 15, 2023. [Online]. Available: <https://arxiv.org/abs/2305.14728v2>.
 - [51] M. S. Hossen and N. R. Dev, "An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data," *Wirel. Pers. Commun.*, vol. 120, no. 1, pp. 535–544, Sep. 2021, doi: 10.1007/S11277-021-08474-4/METRICS.
 - [52] P. Thangavel and R. Lourdusamy, "A lexicon-based approach for sentiment analysis of multimodal content in tweets," *Multimed. Tools Appl.*, vol. 82, no. 16, pp. 24203–24226, Jul. 2023, doi: 10.1007/S11042-023-14411-3/METRICS.
 - [53] Q. He, "Recent Works for Sentiment Analysis using Machine Learning and Lexicon Based Approaches," *Proc. - 2022 5th Int. Conf. Adv. Electron. Mater. Comput. Softw. Eng. AEMCSE 2022*, pp. 422–426, 2022, doi: 10.1109/AEMCSE55572.2022.00090.
 - [54] G. K. Shinde, "Sentiment Analysis Using Hybrid Approach," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 12, pp. 282–285, Dec. 2021, doi: 10.22214/IJRASET.2021.39202.
 - [55] A. L. Ilao and A. C. Fajardo, "Sentiment analysis of tweet messages using hybrid approach algorithm," *Int. Conf. ICT Knowl. Eng.*, vol. 2019-November, Nov. 2019, doi: 10.1109/ICTKE47035.2019.8966887.
 - [56] B. Keith and C. Meneses, "A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews," 2017.
 - [57] M. Aly and A. Atiya, "LABR: A Large Scale Arabic Book Reviews Dataset." pp. 494–498, 2013, Accessed: Nov. 18, 2023. [Online]. Available: <https://aclanthology.org/P13-2088>.
 - [58] A. Elnagar and O. Einea, "BRAD 1.0: Book reviews in Arabic dataset," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 0, Jul. 2016, doi: 10.1109/AICCSA.2016.7945800.
 - [59] H. ElSahar and S. R. El-Beltagy, "Building Large Arabic Multi-domain Resources for Sentiment Analysis," *Conf. Intell. Text Process. Comput. Linguist.*, vol. 9042, pp. 23–34, 2015, doi: 10.1007/978-3-319-18117-2_2.
 - [60] A. Elnagar, "Investigation on sentiment analysis for Arabic reviews," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 0, Jul. 2016, doi: 10.1109/AICCSA.2016.7945623.

- [61] N. Ali Hakami and H. A. Hosni Mahmoud, "Deep Learning Analysis for Reviews in Arabic E-Commerce Sites to Detect Consumer Behavior towards Sustainability," *Sustain.* 2022, Vol. 14, Page 12860, vol. 14, no. 19, p. 12860, Oct. 2022, doi: 10.3390/SU141912860.
- [62] M. Al-Ayyoub, A. Nuseir, G. Kanaan, and R. Al-Shalabi, "Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, 2016, doi: 10.14569/IJACSA.2016.070269.
- [63] M. A. Sghaier and M. Zrigui, "Sentiment analysis for Arabic e-commerce websites," *Proc. - 2016 Int. Conf. Eng. MIS, ICEMIS 2016*, Nov. 2016, doi: 10.1109/ICEMIS.2016.7745323.
- [64] S. Alotaibi and C. Anderson, "Word Clustering as a Feature for Arabic Sentiment Classification," *Int. J. Educ. Manag. Eng.*, vol. 7, no. 1, p. 1, Jan. 2017, doi: 10.5815/IJEME.2017.01.01.
- [65] P. Sasikala and L. Mary Immaculate Sheela, "Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS," *J. Big Data*, vol. 7, no. 1, pp. 1–20, Dec. 2020, doi: 10.1186/S40537-020-00308-7/FIGURES/4.
- [66] M. E. M. Abo *et al.*, "A Multi-Criteria Approach for Arabic Dialect Sentiment Analysis for Online Reviews: Exploiting Optimal Machine Learning Algorithm Selection," *Sustain.* 2021, Vol. 13, Page 10018, vol. 13, no. 18, p. 10018, Sep. 2021, doi: 10.3390/SU131810018.
- [67] L. Bai *et al.*, "Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning," *Comput.* 2023, Vol. 12, Page 126, vol. 12, no. 6, p. 126, Jun. 2023, doi: 10.3390/COMPUTERS12060126.
- [68] A. El-Halees and A. Al-Asmar, "Ontology Based Arabic Opinion Mining," <https://doi.org/10.1142/S0219649217500289>, vol. 16, no. 3, Jul. 2017, doi: 10.1142/S0219649217500289.
- [69] L. Abd-Elhamid, D. Elzanfaly, and A. S. Eldin, "Feature-based sentiment analysis in online Arabic reviews," *Proc. 2016 11th Int. Conf. Comput. Eng. Syst. ICCES 2016*, pp. 260–265, Jan. 2017, doi: 10.1109/ICCES.2016.7822011.
- [70] S. Ismail, A. Alsammak, and T. Elshishtawy, "A generic approach for extracting aspects and opinions of arabic reviews," *ACM Int. Conf. Proceeding Ser.*, vol. 09-11-May-2016, pp. 173–179, May 2016, doi: 10.1145/2908446.2908467.
- [71] M. S. Al-Batah, S. Mrayyen, and M. Alzaqebah, "Investigation of Naive Bayes combined with Multilayer Perceptron for Arabic sentiment analysis and opinion mining," *J. Comput. Sci.*, vol. 14, no. 8, pp. 1104–1114, 2018, doi: 10.3844/JCSSP.2018.1104.1114.
- [72] H. Al-Rubaiee, R. Qiu, and D. Li, "Identifying Mubasher software products

- through sentiment analysis of Arabic tweets,” *2016 Int. Conf. Ind. Informatics Comput. Syst. CIICS 2016*, Apr. 2016, doi: 10.1109/ICCSII.2016.7462396.
- [73] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, “AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons,” *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap.*, vol. 2, pp. 697–705, 2016, doi: 10.18653/V1/P16-1066.
- [74] M. Nabil, M. Aly, and A. F. Atiya, “ASTD: Arabic Sentiment Tweets Dataset,” *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.*, pp. 2515–2519, 2015, doi: 10.18653/V1/D15-1299.
- [75] S. Rosenthal, N. Farra, and P. Nakov, “SemEval-2017 Task 4: Sentiment Analysis in Twitter,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 502–518, 2017, doi: 10.18653/V1/S17-2088.
- [76] F. Alqasemi, A. Abdelwahab, and H. Abdelkader, “An enhanced feature extraction technique for improving sentiment analysis in Arabic language,” *Colloq. Inf. Sci. Technol. Cist*, vol. 0, pp. 380–384, Jul. 2016, doi: 10.1109/CIST.2016.7805075.
- [77] A. Alduailej and A. Alothaim, “AraXLNet: pre-trained language model for sentiment analysis of Arabic,” *J. Big Data*, vol. 9, no. 1, pp. 1–21, Dec. 2022, doi: 10.1186/S40537-022-00625-Z/TABLES/12.
- [78] H. Paul and A. Nikolaev, “Fake review detection on online E-commerce platforms: a systematic literature review,” *Data Min. Knowl. Discov.*, vol. 35, no. 5, pp. 1830–1881, Sep. 2021, doi: 10.1007/S10618-021-00772-6/TABLES/3.
- [79] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp Fake Review Filter Might Be Doing?,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 7, no. 1, pp. 409–418, 2013, doi: 10.1609/ICWSM.V7I1.14389.
- [80] N. Jindal and B. Liu, “Review Spam Detection,” in *The 16th international conference on World Wide Web*, 2007, pp. 1189–1190, Accessed: Dec. 24, 2022. [Online]. Available: <http://www.r-project.org/>.
- [81] S. P. Algur, A. P. Patil, P. S. Hiremath, and S. Shivashankar, “Conceptual level similarity measure based review spam detection,” *Proc. 2010 Int. Conf. Signal Image Process. ICSIP 2010*, pp. 416–423, 2010, doi: 10.1109/ICSIP.2010.5697509.
- [82] Yinqing Xu, Bei Shi, Wentao Tian, and Wai Lam, “A unified model for unsupervised opinion spamming detection incorporating text generality | Proceedings of the 24th International Conference on Artificial Intelligence,” in *The 24th International Conference on Artificial Intelligence*, 2015, pp. 725–731, Accessed: Jan. 17, 2023. [Online]. Available: <https://dl.acm.org/doi/10.5555/2832249.2832350>.
- [83] V. W. Feng and G. Hirst, “Detecting Deceptive Opinions with Profile

- Compatibility,” in *The Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 338–346, Accessed: Jan. 17, 2023. [Online]. Available: <https://aclanthology.org/I13-1039>.
- [84] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding Deceptive Opinion Spam by Any Stretch of the Imagination,” *ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 1, pp. 309–319, Jul. 2011, doi: 10.48550/arxiv.1107.4557.
 - [85] S. Banerjee, A. Y. K. Chua, and J. J. Kim, “Using supervised learning to classify authentic and fake online reviews,” in *The 9th International Conference on Ubiquitous Information Management and Communication*, Jan. 2015, pp. 1–7, doi: 10.1145/2701126.2701130.
 - [86] S. Banerjee and A. Y. K. Chua, “Authentic versus fictitious online reviews: A textual analysis across luxury, budget, and mid-range hotels,” <http://dx.doi.org/10.1177/0165551515625027>, vol. 43, no. 1, pp. 122–134, Feb. 2016, doi: 10.1177/0165551515625027.
 - [87] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “Fake Review Detection : Classification and Analysis of Real and Pseudo Reviews,” 2013.
 - [88] R. Y. K. Lau, S. Y. Liao, R. Chi-Wai Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 4, Jan. 2012, doi: 10.1145/2070710.2070716.
 - [89] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li, and D. Song, “High-order concept associations mining and inferential language modeling for online review spam detection,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1120–1127, 2010, doi: 10.1109/ICDMW.2010.30.
 - [90] G. Wu, D. Greene, and P. Cunningham, “Merging multiple criteria to identify suspicious reviews,” in *The 4th ACM Conference on Recommender Systems*, 2010, pp. 241–244, doi: 10.1145/1864708.1864757.
 - [91] R. P. Kashti and D. Prasad, “Enhancing NLP Techniques for Fake Review Detection,” *Int. Res. J. Eng. Technol.*, 2019.
 - [92] A. Mukherjee *et al.*, “Spotting opinion spammers using behavioral footprints,” in *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2013, vol. Part F128815, pp. 632–640, doi: 10.1145/2487575.2487580.
 - [93] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, “Detection of opinion spam based on anomalous rating deviation,” *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8650–8657, Aug. 2016, doi: 10.48550/arxiv.1608.00684.
 - [94] G. Wu, D. Greene, B. Smyth, and P. Cunningham, “Distortion as a validation

- criterion in the identification of suspicious reviews,” in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 2010, pp. 10–13, doi: 10.1145/1964858.1964860.
- [95] F. Li, M. Huang, Y. Yang, and X. Zhu, “Learning to identify review spam,” in *The Twenty-Second international joint conference on Artificial Intelligence*, Jul. 2011, pp. 2488–2493, Accessed: Dec. 24, 2022. [Online]. Available: <https://dl.acm.org/doi/10.5555/2283696.2283811>.
 - [96] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, “Uncovering crowdsourced manipulation of online reviews,” in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2015, pp. 233–242, doi: 10.1145/2766462.2767742.
 - [97] S. Kumar, M. Kumar, B. Hooi, C. Faloutsos, D. Makhija, and V. S. Subrahmanian, “REV2: Fraudulent user prediction in Rating platforms,” *WSDM 2018 - Proc. 11th ACM Int. Conf. Web Search Data Min.*, vol. 2018-February, pp. 333–341, Feb. 2018, doi: 10.1145/3159652.3159729.
 - [98] X. Wang, K. Liu, S. He, and J. Zhao, “Learning to Represent Review with Tensor Decomposition for Spam Detection,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 866–875, 2016, doi: 10.18653/V1/D16-1083.
 - [99] Y. Lu, L. Zhang, Y. Xiao, and Y. Li, “Simultaneously detecting fake reviews and review spammers using Factor Graph model,” *Proc. 5th Annu. WebSci’13*, vol. volume, pp. 225–233, 2013, doi: 10.1145/2464464.2464470.
 - [100] S. Rayana and L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata,” in *The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2015, vol. 2015-August, pp. 985–994, doi: 10.1145/2783258.2783370.
 - [101] D. Martens and W. Maalej, “Towards understanding and detecting fake reviews in app stores,” *Empir. Softw. Eng.*, vol. 24, no. 6, pp. 3316–3355, Dec. 2019, doi: 10.1007/S10664-019-09706-9/FIGURES/18.
 - [102] Z. You, T. Qian, and B. Liu, “An Attribute Enhanced Domain Adaptive Model for Cold-Start Spam Review Detection,” in *The 27th international conference on computational linguistics*, 2018, pp. 1884–1895, Accessed: Jan. 21, 2023. [Online]. Available: <https://aclanthology.org/C18-1160>.
 - [103] Y. Liu and B. Pang, “A unified framework for detecting author spamicity by modeling review deviation,” *Expert Syst. Appl.*, vol. 112, pp. 148–155, Dec. 2018, doi: 10.1016/J.ESWA.2018.06.028.
 - [104] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, “Detecting group review spam,” in *Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011*, 2011, pp. 93–94, doi: 10.1145/1963192.1963240.

- [105] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S. M. R. Beheshti, E. Bertino, and N. Foo, "Collusion detection in online rating systems," in *APWeb 2013: Web Technologies and Applications*, 2013, vol. 7808 LNCS, pp. 196–207, doi: 10.1007/978-3-642-37401-2_21/COVER.
- [106] N. Hussain, H. Turab Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam Review Detection Using the Linguistic and Spammer Behavioral Methods," *IEEE Access*, vol. 8, pp. 53801–53816, 2020, doi: 10.1109/ACCESS.2020.2979226.
- [107] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," *52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf.*, vol. 1, pp. 1566–1576, 2014, doi: 10.3115/V1/P14-1147.
- [108] M. Ott, C. Cardie, and J. T. Hancock, "Negative Deceptive Opinion Spam," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2013, pp. 497–501, Accessed: Dec. 24, 2022. [Online]. Available: <http://plagiarisma.net>.
- [109] A. A. Hammad and A. El-Halees, "An Approach for Detecting Spam in Arabic Opinion Reviews," *Int. Arab J. Inf. Technol.*, vol. 12, no. 1, 2015.
- [110] Y. Ren, D. Ji, and H. Zhang, "Positive Unlabeled Learning for Deceptive Reviews Detection," in *Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 488–498.
- [111] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 2015-August, pp. 985–994, Aug. 2015, doi: 10.1145/2783258.2783370.
- [112] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting Fake Reviews via Collective Positive-Unlabeled Learning," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2015-January, no. January, pp. 899–904, Jan. 2014, doi: 10.1109/ICDM.2014.47.
- [113] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1242–1247, 2011, doi: 10.1109/ICDM.2011.124.
- [114] S. K. Fayazbakhsh and J. Sinha, "Spam Detection : A Network-based Approach," 2012.
- [115] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," *WWW'12 - Proc. 21st Annu. Conf. World Wide Web*, pp. 191–200, 2012, doi: 10.1145/2187836.2187863.
- [116] Q. Peng, "Store review spammer detection based on review relationship," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8697, pp. 287–298, 2014, doi: 10.1007/978-3-319-14139-8_30/COVER.

- [117] B. Jiang, R. hao Cao, and B. Chen, “Detecting Product Review Spammers using Activity Model,” pp. 650–653, Aug. 2013, doi: 10.2991/ICACSEI.2013.155.
- [118] J. Huang, T. Qian, G. He, M. Zhong, and Q. Peng, “Detecting professional spam reviewers,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8347 LNAI, no. PART 2, pp. 288–299, 2013, doi: 10.1007/978-3-642-53917-6_26/COVER.
- [119] J. Wang and X. Liang, “Discovering the rating pattern of online reviewers through data coclustering,” *IEEE ISI 2013 - 2013 IEEE Int. Conf. Intell. Secur. Informatics Big Data, Emergent Threat. Decis. Secur. Informatics*, pp. 374–376, 2013, doi: 10.1109/ISI.2013.6578862.
- [120] C. M. Aye and K. M. Oo, “Review Spammer Detection by using Behaviors Based Scoring Methods,” Mar. 2014, doi: 10.15242/IIE.E0314158.
- [121] T. Fornaciari and M. Poesio, “Identifying fake Amazon reviews as learning from crowds,” *14th Conf. Eur. Chapter Assoc. Comput. Linguist. 2014, EACL 2014*, pp. 279–287, 2014, doi: 10.3115/V1/E14-1030.
- [122] Z. Hai, P. Zhao, P. Cheng, P. Yang, X. L. Li, and G. Li, “Deceptive Review Spam Detection via Exploiting Task Relatedness and Unlabeled Data,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 1817–1826, 2016, doi: 10.18653/V1/D16-1187.
- [123] E. Choo, T. Yu, and M. Chi, “Detecting Opinion Spammer Groups Through Community Discovery and Sentiment Analysis,” pp. 170–187, 2015, doi: 10.1007/978-3-319-20810-7_11i.
- [124] “Amazon Reviews 2018 (Full Dataset) | Kaggle.”
<https://www.kaggle.com/datasets/roga16/amazon-reviews-2018-full-dataset>
 (accessed Dec. 24, 2022).
- [125] R. Barbado, O. Araque, and C. A. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1234–1244, Jul. 2019, doi: 10.1016/J.IPM.2019.03.002.
- [126] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, “Fake review detection based on multiple feature fusion and rolling collaborative training,” *IEEE Access*, vol. 8, pp. 182625–182639, 2020, doi: 10.1109/ACCESS.2020.3028588.
- [127] R. Mohawesh, S. Xu, M. Springer, Y. Jararweh, M. Al-Hawawreh, and S. Maqsood, “An explainable ensemble of multi-view deep learning model for fake review detection,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 8, p. 101644, Sep. 2023, doi: 10.1016/J.JKSUCI.2023.101644.
- [128] S. N. Alsubari, T. H. H. Aldhyani, S. N. Deshmukh, M. Maashi, S. Alharbi, and H. H. Al-Baity, “Computational Intelligence Based Recurrent Neural Network for Identification Deceptive Review in the E-Commerce Domain,” *Comput. Intell.*

Neurosci., vol. 2022, pp. 1–14, Nov. 2022, doi: 10.1155/2022/4656846.

- [129] C. G. Harris, “Combining Linguistic and Behavioral Clues to Detect Spam in Online Reviews,” *Proc. - 2022 IEEE Int. Conf. E-bus. Eng. ICEBE 2022*, pp. 38–44, 2022, doi: 10.1109/ICEBE55470.2022.00017.
- [130] J. Kumar, “Fake Review Detection Using Behavioral and Contextual Features,” Feb. 2020, Accessed: Sep. 15, 2023. [Online]. Available: <https://arxiv.org/abs/2003.00807v1>.
- [131] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. Binte Hassan, “Spam Review Detection Using Deep Learning,” *2019 IEEE 10th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2019*, pp. 27–33, Oct. 2019, doi: 10.1109/IEMCON.2019.8936148.
- [132] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting Burstiness in Reviews for Review Spammer Detection,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 7, no. 1, pp. 175–184, 2013, doi: 10.1609/ICWSM.V7I1.14400.
- [133] M. Fazzolari, F. Buccafurri, G. Lax, and M. Petrocchi, “Improving Opinion Spam Detection by Cumulative Relative Frequency Distribution,” Dec. 2020, Accessed: Sep. 15, 2023. [Online]. Available: <https://arxiv.org/abs/2012.13905v1>.
- [134] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, “Ensemble machine learning model for classification of spam product reviews,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8857570.
- [135] F. Khurshid, Y. Zhu, Z. Xu, M. Ahmad, and M. Ahmad, “Enactment of Ensemble Learning for Review Spam Detection on Selected Features,” *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, pp. 387–394, Jan. 2019, doi: 10.2991/IJCIS.2019.125905655.
- [136] S. Mani, S. Kumari, A. Jain, and P. Kumar, “Spam review detection using ensemble machine learning,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10935 LNAI, pp. 198–209, 2018, doi: 10.1007/978-3-319-96133-0_15/COVER.
- [137] J. Yao, Y. Zheng, and H. Jiang, “An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization,” *IEEE Access*, vol. 9, pp. 16914–16927, 2021, doi: 10.1109/ACCESS.2021.3051174.
- [138] N. Cao, S. Ji, D. K. W. Chiu, M. He, and X. Sun, “A deceptive review detection framework: Combination of coarse and fine-grained features,” *Expert Syst. Appl.*, vol. 156, p. 113465, Oct. 2020, doi: 10.1016/J.ESWA.2020.113465.
- [139] F. Khurshid, Y. Zhu, C. W. Yohannese, and M. Iqbal, “Recital of supervised learning on review spam detection: An empirical analysis,” *Proc. 2017 12th Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2017*, vol. 2018-January, pp. 1–6, Jul. 2017,

doi: 10.1109/ISKE.2017.8258755.

- [140] S. Noekhah, N. binti Salim, and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," *Inf. Process. Manag.*, vol. 57, no. 1, p. 102140, Jan. 2020, doi: 10.1016/J.IPM.2019.102140.
- [141] Z. Wang, R. Hu, Q. Chen, P. Gao, and X. Xu, "ColluEagle: collusive review spammer detection using Markov random fields," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1621–1641, Nov. 2020, doi: 10.1007/S10618-020-00693-W/METRICS.
- [142] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, Apr. 2019, doi: 10.1109/ECACE.2019.8679186.
- [143] H. Deng *et al.*, "Semi-supervised learning based fake review detection," *Proc. - 15th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 16th IEEE Int. Conf. Ubiquitous Comput. Commun. ISPA/IUCC 2017*, pp. 1278–1280, May 2018, doi: 10.1109/ISPA/IUCC.2017.00195.
- [144] S. Zhao, Z. Xu, L. Liu, M. Guo, and J. Yun, "Towards Accurate Deceptive Opinion Spam Detection based on Word Order-preserving CNN," *Math. Probl. Eng.*, vol. 2018, Nov. 2017, doi: 10.48550/arxiv.1711.09181.
- [145] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manag.*, vol. 54, no. 4, pp. 576–592, Jul. 2018, doi: 10.1016/J.IPM.2018.03.007.
- [146] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, "An Inferable Representation Learning for Fraud Review Detection with Cold-start Problem," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2019-July, Jul. 2019, doi: 10.1109/IJCNN.2019.8852437.
- [147] C. C. Wang, M. Y. Day, C. C. Chen, and J. W. Liou, "Detecting spamming reviews using long short-term memory recurrent neural network framework," *ACM Int. Conf. Proceeding Ser.*, Jun. 2018, doi: 10.1145/3234781.3234794.
- [148] W. Liu, W. Jing, and Y. Li, "Incorporating feature representation into BiLSTM for deceptive review detection," *Computing*, vol. 102, no. 3, pp. 701–715, Mar. 2020, doi: 10.1007/S00607-019-00763-Y/METRICS.
- [149] Z. Y. Zeng, J. J. Lin, M. S. Chen, M. H. Chen, Y. Q. Lan, and J. L. Liu, "A Review Structure Based Ensemble Model for Deceptive Review Spam," *Inf. 2019, Vol. 10, Page 243*, vol. 10, no. 7, p. 243, Jul. 2019, doi: 10.3390/INFO10070243.
- [150] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17259–17274, Dec. 2020, doi:

10.1007/S00521-020-04757-2/METRICS.

- [151] R. Narayan, J. K. Rout, and S. K. Jena, "Review spam detection using opinion mining," *Adv. Intell. Syst. Comput.*, vol. 519, pp. 273–279, 2018, doi: 10.1007/978-981-10-3376-6_30/COVER.
- [152] S. Jamshidi Nejad, F. Ahmadi-Abkenari, and P. Bayat, "Opinion Spam Detection based on Supervised Sentiment Analysis Approach," *2020 10th Int. Conf. Comput. Knowl. Eng. ICCKE 2020*, pp. 209–214, Oct. 2020, doi: 10.1109/ICCKE50421.2020.9303677.
- [153] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17259–17274, Dec. 2020, doi: 10.1007/S00521-020-04757-2/METRICS.
- [154] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making," *Ind. Mark. Manag.*, vol. 90, pp. 523–537, Oct. 2020, doi: 10.1016/J.INDMARMAN.2019.08.003.
- [155] P. P. Patil, S. Phansalkar, S. Ahirrao, and A. Pawar, "Alosi: aspect-level opinion spam identification," *Lect. Notes Networks Syst.*, vol. 132, pp. 135–147, 2021, doi: 10.1007/978-981-15-5309-7_14/COVER.
- [156] G. Shan, L. Zhou, and D. Zhang, "From conflicts and confusion to doubts: Examining review inconsistency for fake review detection," *Decis. Support Syst.*, vol. 144, p. 113513, May 2021, doi: 10.1016/J.DSS.2021.113513.
- [157] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel arabic-reviews dataset construction for sentiment analysis applications," *Stud. Comput. Intell.*, vol. 740, pp. 35–52, 2018, doi: 10.1007/978-3-319-67056-0_3/COVER.
- [158] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "SPAR: A system to detect spam in Arabic opinions," *2013 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. AEECT 2013*, 2013, doi: 10.1109/AEECT.2013.6716442.
- [159] H. Najadat, M. A. Alzubaidi, and I. Qarqaz, "Detecting Arabic Spam Reviews in Social Networks Based on Classification Algorithms," *Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 1, Nov. 2021, doi: 10.1145/3476115.
- [160] R. M. K. Saeed, S. Rady, and T. F. Gharib, "An ensemble approach for spam detection in Arabic opinion texts," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1407–1416, Jan. 2022, doi: 10.1016/J.JKSUCI.2019.10.002.
- [161] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 1, pp. 309–319, Jul. 2011, doi:

10.48550/arxiv.1107.4557.

- [162] I. M. Alwayle *et al.*, “Automated Spam Review Detection Using Hybrid Deep Learning on Arabic Opinions,” *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, pp. 2947–2961, Apr. 2023, doi: 10.32604/CSSE.2023.034456.
- [163] A. Ziani *et al.*, “Deceptive Opinions Detection Using New Proposed Arabic Semantic Features,” *Procedia Comput. Sci.*, vol. 189, pp. 29–36, Jan. 2021, doi: 10.1016/J.PROCS.2021.05.067.
- [164] A. Ziani, C. A. K. A. Kounta, N. Azizi, and S. Cheriguene, “A New Hybrid Classification System for Subjectivity Detection | Request PDF,” Dec. 2016.
- [165] samaher alharthi, R. Siddiq, and H. Alghamdi, “Detecting Arabic Fake Reviews in E-commerce Platforms Using Machine and Deep Learning Approaches,” *J. King Abdulaziz Univ. Comput. Inf. Technol. Sci.*, vol. 11, no. 1, pp. 27–34, Sep. 2022, doi: 10.4197/Comp.11-1.3.
- [166] R. AlYami and R. Al-Zaidy, “Weakly and Semi-Supervised Learning for Arabic Text Classification using Monodialectal Language Models,” *WANLP 2022 - 7th Arab. Nat. Lang. Process. - Proc. Work.*, pp. 260–272, 2022, doi: 10.18653/V1/2022.WANLP-1.24.
- [167] L. Richardson, “Beautiful Soup Documentation.” 2021, Accessed: May 21, 2023. [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.
- [168] “googletrans.” <https://pypi.org/project/googletrans/> (accessed Dec. 12, 2023).
- [169] “Selenium.” Accessed: May 21, 2023. [Online]. Available: <https://www.selenium.dev/>.
- [170] G. Chandrarathne, K. Thanikasalam, and A. Pinidiyaarachchi, “A Comprehensive Study on Deep Image Classification with Small Datasets,” *Lect. Notes Electr. Eng.*, vol. 619, pp. 93–106, 2020, doi: 10.1007/978-981-15-1289-6_9/COVER.
- [171] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, “The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models,” *WANLP 2021 - 6th Arab. Nat. Lang. Process. Work. Proc. Work.*, pp. 92–104, Mar. 2021, Accessed: Nov. 21, 2023. [Online]. Available: <https://arxiv.org/abs/2103.06678v2>.
- [172] A. A. Elmadany, H. Mubarak, and W. Magdy, “An Arabic Speech-Act and Sentiment Corpus of Tweets.” European Language Resources Association (ELRA), May 12, 2018, Accessed: Nov. 21, 2023. [Online]. Available: <https://www.research.ed.ac.uk/en/publications/an-arabic-speech-act-and-sentiment-corpus-of-tweets>.
- [173] “CAMEL-Lab/bert-base-arabic-camelbert-mix · Hugging Face.” <https://huggingface.co/CAMEL-Lab/bert-base-arabic-camelbert-mix> (accessed Nov. 21, 2023).

- [174] “CAMEL-Lab/bert-base-arabic-camelbert-mix-sentiment · Hugging Face.” <https://huggingface.co/CAMEL-Lab/bert-base-arabic-camelbert-mix-sentiment> (accessed Nov. 21, 2023).
- [175] M. Abdul-Mageed, A. R. Elmadany, and E. M. B. Nagoudi, “ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic,” *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 7088–7105, Dec. 2020, doi: 10.18653/v1/2021.acl-long.551.
- [176] “UBC-NLP/MARBERT · Hugging Face.” <https://huggingface.co/UBC-NLP/MARBERT> (accessed Nov. 21, 2023).
- [177] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques*. O’Reilly Media, Incorporated, 2019.
- [178] “scikit-learn: machine learning in Python.” Accessed: Dec. 12, 2023. [Online]. Available: <https://scikit-learn.org/stable/>.
- [179] R. P. Kashti and D. Prasad, “Enhancing NLP Techniques for Fake Review Detection,” 2019.
- [180] “SMOTE — Version 0.11.0.” https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html (accessed Dec. 12, 2023).
- [181] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” <https://doi.org/10.1214/aos/1013203451>, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/AOS/1013203451.
- [182] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” doi: 10.5555/3294996.3295074.
- [183] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.
- [184] “Statistical functions (scipy.stats) .” <https://docs.scipy.org/doc/scipy/reference/stats.html> (accessed Dec. 13, 2023).
- [185] P. Bian, L. Liu, and P. Sweetser, “Detecting Spam Game Reviews on Steam with a Semi-Supervised Approach,” *ACM Int. Conf. Proceeding Ser.*, Aug. 2021, doi: 10.1145/3472538.3472547.
- [186] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1995-June, pp. 189–196, 1995, doi: 10.3115/981658.981684.
- [187] J. Bekker and J. Davis, “Learning from positive and unlabeled data: a survey,” *Mach. Learn.*, vol. 109, no. 4, pp. 719–760, Apr. 2020, doi: 10.1007/S10994-020-05877-5/FIGURES/6.

- [188] A. A. Irissappane, H. Yu, Y. Shen, A. Agrawal, and G. Stanton, “Leveraging GPT-2 for Classifying Spam Reviews with Limited Labeled Data via Adversarial Training,” Dec. 2020, Accessed: Sep. 16, 2023. [Online]. Available: <https://arxiv.org/abs/2012.13400v1>.
- [189] G. Stanton and A. A. Irissappane, “GANs for Semi-Supervised Opinion Spam Detection,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-August, pp. 5204–5210, Mar. 2019, doi: 10.24963/ijcai.2019/723.
- [190] A. Ligthart, C. Catal, and B. Tekinerdogan, “Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification,” *Appl. Soft Comput.*, vol. 101, p. 107023, Mar. 2021, doi: 10.1016/J.ASOC.2020.107023.
- [191] Y. Tian, M. Mirzabagheri, P. Tirandazi, and S. M. H. Bamakan, “A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM,” *Inf. Process. Manag.*, vol. 57, no. 6, p. 102381, Nov. 2020, doi: 10.1016/J.IPM.2020.102381.
- [192] D. H. Fusilier, R. G. Cabrera, M. Montes-Y-Gómez, and P. Rosso, “Using PU-Learning to Detect Deceptive Opinion Spam.” *Association for Computational Linguistics*, pp. 38–45, 2013, Accessed: Dec. 07, 2023. [Online]. Available: <https://aclanthology.org/W13-1606>.

Vitae

Name : Reem Abdulrahman AlJunaid

Nationality : Saudi

Date of Birth : 15/August/1994

Email : raaljunaid@iau.edu.sa

Address : Kaab Alahbara Street,
Postal Code: 32415, Dammam, Saudi Arabia

Academic Background : Reem earned her bachelor's degree in computer science from Imam Abdulrahman bin Faisal University in 2017. Currently, she is undertaking her master's degree in computer science at KFUPM, with a research focus on Arabic Natural Language Processing (NLP) and the application of Machine Learning and Deep Learning techniques.

Publications:

- Yamani, K. Bajbaa and R. Aljunaid, "Web Application Security Threats and Mitigation Strategies when Using Cloud Computing as Backend," 2022 14th International Conference on Computational Intelligence and Communication

Networks (CICN), Al-Khobar, Saudi Arabia, 2022, pp. 811-818, doi: 10.1109/CICN56167.2022.10008368.

- Mezhoudi, N., Alghamdi, R., Aljunaid, R. et al. Employability prediction: a survey of current approaches, research challenges and applications. J Ambient Intell Human Comput 14, 1489–1505 (2023). <https://doi.org/10.1007/s12652-021-03276-9>

Awards:

- Durah Title of College of Computer Science and Information Technology, 2015 Imam Abdulrahman Bin Faisal University.
- Third Durah Title of Imam Abdulrahman Bin Faisal University, 2015.
- First Best Project in Database Design & OOP Courses, 201, Imam Abdulrahman Bin Faisal University.
- Ambassador of the College of Computer Science and Information Technology, 2015, Imam Abdulrahman Bin Faisal University Career forum.