JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Abuhammad, A. S., & Ahmed, M. A. (2024). Automatic negation detection for semantic analysis in Arabic hotel reviews through lexical and structural features: A supervised classification. *Journal of Information and Communication Technology, 23*(4), 709-744. https://doi.org/10.32890/jict2024.23.4.5

Automatic Negation Detection for Semantic Analysis in Arabic Hotel Reviews Through Lexical and Structural Features: A Supervised Classification

[1]Ahmed Suliman Abuhammad & [2]Mahmoud Ali Ahmed
[1]Department of Computer Science and Information Technology, University College of Science and Technology, Palestine
[1]Department of Information Technology, University of the Holy Quran and Taseel of Science, Sudan
[2]Department of Computer Science, University of Khartoum, Sudan

*[1]asj.hammad@ucst.edu.ps
[2]mali@uofk.edu
*Corresponding author

Received: 23/5/2024 Revised: 16/10/2024 Accepted: 17/10/2024 Published: 28/10/2024

ABSTRACT

One significant challenge in sentiment analysis is the presence of negation, which reverses the meanings of sentences, transforming positive statements into negative ones and impacting the sentiment conveyed in the text. This issue is particularly pronounced in Arabic, a language known for its complex morphology. Detecting negation

is crucial for enhancing sentiment analysis performance and various natural language processing applications. This paper presents an approach for automatically detecting negation in user-generated Arabic hotel reviews through lexical and structural features. It comprises several stages: data collection, text pre-processing, feature extraction, supervised learning classification, and evaluation. The study employed multiple supervised classification techniques, including naïve Bayes, random forest, logistic regression, support vector machines, and deep learning, to analyse lexical and structural features extracted from the dataset. The results of the experiments yielded promising outcomes, demonstrating the feasibility of the approach for practical applications. The classifiers exhibited highly comparable performance in identifying negation, with only marginal deviations in their performance metrics. Notably, the deep learning classifier consistently emerged as the top performer, achieving an exceptionally high overall accuracy rate of 99.24 percent, surpassing established benchmarks in Arabic text processing and underscoring its potential for practical applications. These findings hold significant implications for advancing Arabic text processing, particularly in sentiment analysis and related NLP tasks. The high accuracy of 99.24 percent achieved by the deep learning classifier highlights its robustness in accurately detecting negation, a critical challenge in sentiment analysis. This classifier performance demonstrates the potential to be integrated into real-world applications, such as automated review systems and opinion mining tools, where accurate sentiment interpretation is essential.

Keywords: Arabic hotel reviews, lexical features, negation detection, semantic analysis, structural features, supervised classification.

INTRODUCTION

Online platforms and social networks have become integral to contemporary communication, with billions of users sharing opinions and content on a wide range of topics. This user-generated content—spanning comments, videos, and images—holds substantial value for businesses and organisations seeking insights into public opinion regarding political events, brand perception, products, and customer service (Burbach et al., 2020; Genadi & Khodra, 2022). Sentiment analysis (SA) has emerged as a critical tool for harnessing these insights. SA is designed to discern the semantic orientation of text,

categorising it as positive, negative, or neutral. Its applications extend across various industries, including business, education, commerce, and healthcare.

Negation, a fundamental aspect of natural language, plays a pivotal role in sentiment analysis. It alters the meaning of sentences by reversing their polarity, transforming affirmative statements into negative ones (Abuhammad & Ahmed, 2023; Burbach et al., 2020). Dictionary.com (n.d.) defines negation as "the exact opposite of something; the act of causing something not to exist or to become its opposite". Similarly, Collins Dictionary (2023) describes it as "the opposite or absence of something". In SA, negation can disrupt sentiment classification by changing the polarity of words within a text, leading to potential misinterpretations and inaccuracies (Eremyan, 2023; Hussein, 2018; Mohammad, 2016). The following are examples of negation sentences:

1. "لا أنصح بالإقامة هنا. الأسرة غير مريحة والموظفون غير متعاونين."
(I do not recommend staying here. The beds are uncomfortable and the staff is unhelpful.)

2. "المسبح ماكو نظيف والماء متسخ."
(The pool is not clean, and the water is dirty.)

3. "الخدمة مو مرضية على الإطلاق."
(The service is not satisfactory at all.)

4. "مش عاجبني الفندق ده، المكان مليان عيوب وخدماته سيئة جدًا."
(I don't like this hotel; the place is full of flaws, and its services are very bad.)

5. "المدير مهوش ودود."
(The manager is not friendly.)

Negation can appear in various forms, such as explicit and implicit negations, diminutives, and other subtle linguistic patterns (Farooq, 2017). These forms can be either morphological or syntactic, with syntactic negations including fake and double negations (Councill et al., 2010; Mukherjee et al., 2021). Each type of negation impacts sentiment polarity differently, making accurate detection crucial (Alotaibi, 2015; El-Dine & El-Zahraa, 2013).

Automated detection of negation involves computational techniques to determine whether a review expresses a 'negated positive' sentiment or remains positive. This task is particularly challenging due to the diverse ways in which negation can be expressed, especially in languages with complex morphology like Arabic.

Consider the following examples from the corpus:

- Positive: "المطعم رائع" ("The restaurant is wonderful")
- Negated Positive: "المطعم ليس رائعاً" ("The restaurant is not wonderful")

In the first example, the sentiment is clearly positive. In the second example, the sentiment is negated positive, indicating that while the sentiment term "رائع" ("wonderful") is positive, the negation "ليس" ("not") changes the overall sentiment to negative.

While research on automatic negation detection has largely focused on English, there is a notable gap in studies addressing this issue in Arabic. Arabic's intricate morphology and syntax present unique challenges that have not been thoroughly explored in existing research. Arabic, spoken by over 422 million people worldwide, poses significant hurdles for natural language processing (NLP) due to its complexity, distinguishing it from languages like English (Wikipedia, 2023a; 2023b). Despite advancements in Arabic NLP tools, such as morphological analysers and syntactical parsers, significant challenges remain, particularly in areas like text classification and sentiment analysis (World Internet Users' Statistics and 2023 World Population Stats., n.d.).

An approach for automated negation detection in Arabic reviews, leveraging both lexical and structural features, is proposed in this study. It entails utilising various supervised classification techniques, including Naive Bayes (NB), Random Forest (RF), Logistic Regression (LogR), Support Vector Machine (SVM), and Deep Learning (DL), applied to a collection of lexical and structural features extracted from the dataset. Data was collected from prominent online Arabic economic websites hosting opinion reviews, resulting in a corpus of 84,000 Arabic opinion reviews evenly divided between 'negated positive' and positive reviews. This paper has five main sections. The first section introduces the topic, followed by a section

on related work. The third section outlines the proposed approach for negation detection in Arabic reviews, while the fourth section details the experiments conducted and analyses their outcomes. Finally, the paper concludes and outlines directions for future research.

RELATED WORKS

Automatic negation detection has become an important area of research in text mining, particularly due to its impact on SA in online platforms and social media. While much of the existing research has concentrated on English, there is a growing interest in expanding these methods to other languages. This section focuses on studies related to Arabic online reviews, highlighting the specific challenges of negation, double negation, and implicit negation in semantic analysis. Mukherjee et al. (2021) investigated the integration of negation handling in SA by developing a negation marking algorithm to identify explicit negation. They applied various classifiers, including NB, SVM, artificial neural networks (ANNs), and recurrent neural networks (RNNs), to a dataset of 75,000 Amazon product reviews. Their approach improved performance significantly, with RNNs achieving the highest accuracy of 95.67 percent. However, their study did not address double or implicit negation, which may limit the comprehensiveness of their results.

Alharbi (2020) proposed a method to enhance sentiment classification in consumer reviews by tackling negation through machine learning. Their approach used a sentiment lexicon, defined rules, and linguistic knowledge implemented in Python 3.0. The dataset comprised 2,400 annotated reviews in Modern Standard Arabic (MSA) and Jordanian colloquial language. They identified 50 common negation terms and evaluated classifiers such as SVM, k-nearest neighbors (KNN), NB, and (LR). SVM achieved an accuracy of 89.17 percent. However, the algorithm did not address implicit negation or account for intensifiers and diminishers, which could affect sentiment polarity classification.

Funkner et al. (2020) conducted SA focused on Russian medical reports, using multi-class classification to identify negations. Their dataset included 3,434 electronic medical records (EMRs), with XGBoost, RF, and KNN classifiers. They demonstrated that integrating negation detection improved predictive model performance, with

F-scores ranging from 81.00 percent to 93.00 percent. While their study provides insights into negation detection, it is limited to medical texts and does not address double or implicit negation. Mahany et al. (2020) emphasised the importance of detecting negation in MSA and Classical Arabic (CA) texts. They used a manually annotated dataset from King Saud University Corpus of Classical Arabic (KSUCCA) and Wikipedia, focusing on six negative particles. Their experiments employed word embedding models and classification techniques. For word embedding, they used Word2Vec and FastText. For classification, they utilised both classical machine learning and deep learning approaches, including SVM for classical machine learning and BiLSTM for deep learning. They achieved an F1-score of 89.00 percent and an accuracy of 93.00 percent in negation scope detection. However, their research did not provide detailed information on implicit negation or fake inverters.

The reviewed studies indicate a growing interest in automatic negation detection, particularly in Arabic texts. However, there is a notable research gap in addressing complex negation scenarios such as double and implicit negation in Arabic online reviews. This study aims to fill this gap by focusing on these specific challenges, thereby contributing to more accurate SA in Arabic text processing. Table 1 summarises the existing works related to automatic negation detection, highlighting the corpus, features, models used, best results obtained, and gaps.

Table 1

A Summary of the Existing Works Related to Automatic Negation Detection

Studies	Features	Model Used	Corpus	Best Result	Gaps
Mukherjee et al. (2021)	Syntactic	Machine Learning (NB, SVM, ANN, and RNN)	Product Review	Accuracy RNN 95.67%	In their investigation of sentiment polarity detections, this method did not consider implicit negations and double negations. Experiments were based on Amazon Product Reviews, specifically on cell phones, and not tested in the general domain.
Alharbi (2020)	Syntactic	Rule-Based and Machine Learning (SVM, NB, KNN, and LR)	Review	F1-Score SVM 89.17%	Implicit negation, which can also have a negative impact on polarity classification, is ignored by the algorithm. The use of intensifiers and diminishers, which can alter the polarity of words or phrases, is something that the proposed method does not address.
Funkner et al. (2020)	Syntactic	Machine Learning (XGBoost, RF, and KNN)	EMR	F1-Score RF %93.00	The work does not deal with odd negation, fake inverters, complex negation, and implicit negation.

(continued)

Studies	Features	Model Used	Corpus	Best Result	Gaps
Mahany et al. (2020)	Word Embedding	SVM and Deep Learning (BiLSTM)	KSUCCA and Wikipedia Sentence	F1-Score FastText+ BiLSTM %86.0	Only two genres (KSUCCA and Wikipedia) have been considered, and further testing on other genres is required. The negation cues in the entire corpus have only six negative particles. They did not explain how to deal with implicit negation and fake inverters through their proposed system.

THE PROPOSED APPROACH

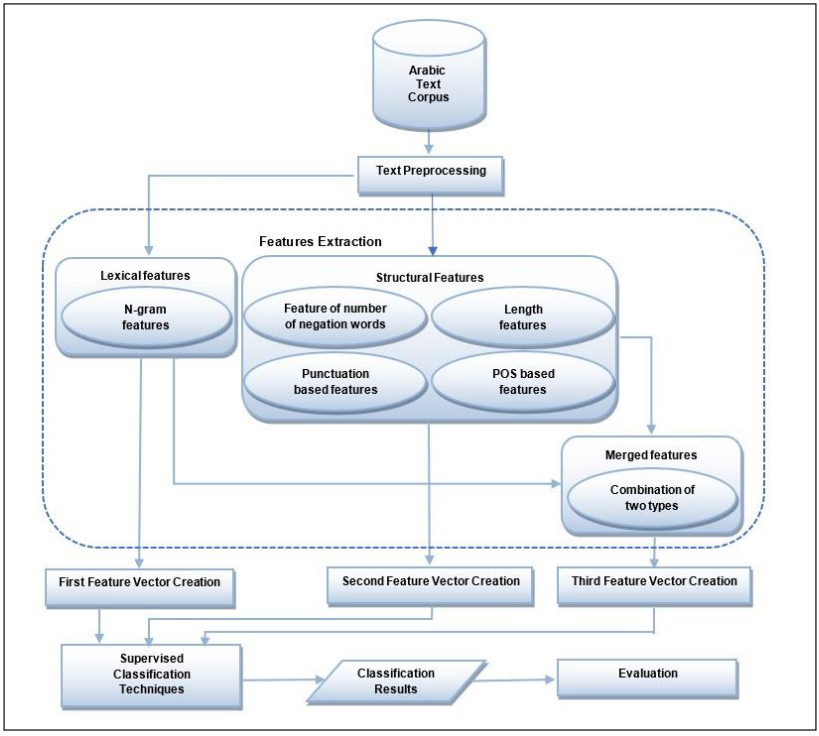
This section outlines the approach for detecting negation in Arabic opinion reviews, which consists of five key components: Arabic Text Corpus Collection, Text Processing, Feature Extraction, Supervised Classification, and Model Evaluation.

1) Arabic text corpus - This step involves gathering Arabic opinion reviews from various online platforms, forming the corpus for analysis.
2) Text processing - The collected reviews undergo pre-processing, which includes cleaning the data by removing noise, such as irrelevant characters and formatting inconsistencies.
3) Feature extraction - In this phase, relevant linguistic features, such as negation cues and sentiment-bearing words, are identified and extracted from the processed text.
4) Supervised classification techniques - Using a supervised learning approach, the reviews are classified by applying specific algorithms such as includin NB, RF, LogR, SVM, and DL. These classifiers analyze features such as the presence or absence of negation cues in the text to determine the sentiment of the reviews. The technique leverages labeled training data to accurately predict the sentiment in new reviews based on whether negation is present or not.
5) Evaluation - The performance of the negation detection is assessed using standard evaluation metrics.

The overall process is illustrated in Figure 1, which visualises these components and their interactions within the negation detection framework.

Figure 1

The Entire Process of the Proposed Approach



Text Data Collection

The study collected Arabic opinion reviews from three major economic websites from June 2013 to June 2023: (1) TripAdvisor (https://www.tripadvisor.com), (2) Booking.com (https://www.booking.com), and (3) Agoda (https://www.agoda.ae). The dataset comprises 84,000 reviews, evenly split between 42,000 'negated positive' and 42,000 positives.

Data Collection Process

The data collection process for Arabic opinion reviews from selected websites involves several steps:

1) Target domain selection- The study chose tourism domains like hotels, resorts, and vacation rentals.

2) Crawling and scraping- The study used web crawling and scraping to collect review text, ratings, and metadata.
3) Language filtering- Arabic reviews are filtered using language identification techniques.
4) Negation annotation- To create a labelled dataset of positive and 'negated positive' reviews for negation detection, the study relied on annotations provided by reviewers based on their ratings. Reviews rated 4 or 5 stars were classified as positive, while those receiving 1 or 2 stars were classified as negative. The study assumed that the most reliable judgment of whether a review is positive or negative comes from the review author. Utilising reviews labelled by their authors' ratings allowed us to establish a high-quality gold standard and build extensive datasets. This process, known as automatic annotation, was followed by manual annotation. During manual annotation, experts familiar with Arabic linguistics meticulously examined each review to identify the presence or absence of specific linguistic indicators known as negation cues. Annotators thoroughly reviewed the text to identify these cues, which included negation words like 'لا' (Laa, meaning 'no') and negation phrases such as 'ليس' (laysa, meaning 'not'). Specifically, three annotators were involved in this process. This meticulous examination allowed annotators to mark and categorise reviews based on whether they contained these identified negation cues. To validate the correctness of the resultant dataset, the study employed a multi-step process:
    - Expert review- The process involved a comprehensive review by domain experts in ANLP. Three experts meticulously examined the identified negation cues and the resultant dataset, providing valuable feedback to refine the approach. Their involvement was crucial in validating the accuracy of the negation detection.
    - Annotation guidelines- The annotation process adhered to well-established guidelines specifically developed for negation detection. These guidelines were formulated in consultation with experts to standardise the annotation process and enhance reliability. This study detailed these guidelines in the proposed approach section of this paper to ensure transparency.
    - Inter-annotator agreement- To measure the consistency and reliability of the annotations, the study conducted

an inter-annotator agreement assessment. Multiple annotators independently identified negation cues in a subset of the dataset, and the study calculated the agreement rates using Cohen's Kappa metric. The high agreement rates demonstrate the annotation process's robustness and consistency.

5) Data storage- filtered and annotated reviews are stored in structured formats like Excel or text files.

This process yields a substantial collection of Arabic opinion reviews, covering various forms of negation and totalling 84,000 reviews split evenly between 'negated positive' and positive reviews.

Data Set Size and Composition

The resulting dataset comprises 84,000 Arabic opinion reviews, evenly split between 42,000 'negated positive' reviews and 42,000 positive reviews, all acquired from TripAdvisor (https://www.tripadvisor.com), Booking.com (https://www.booking.com), and Agoda (https://www.agoda.ae). Table 2 summarises the distribution of reviews and the key dataset characteristics. The dataset includes reviews written by users in Arabic (MSA and DA or a combination of both), specifically focusing on tourism domains such as hotels, accommodations, and related services. The study expects the dataset to exhibit variations in the lengths of reviews, writing styles, and the sentiments expressed. It encompasses a variety of opinions and experiences shared by users, providing a representative sample of Arabic opinion reviews in the tourism domain.

Table 2

Statistics of the Data Set

	Negated Positive	Positive	Total
Number of documents	42,000	42,000	84,000
Number of sentences	84,149	608,571	692,720
Number of words	1,208,061	3,588,204	4,796,265
Average length documents (in sentences)	20.04	14.49	17.27
Average length documents (in words)	28.76	85.43	57.10
Average length sentences (in words)	14.36	5.90	10.13

The dataset contains around 150 negation words and phrases. Table 3 displays a list of commonly used negation cues found in Arabic texts (MSA and DA) and their frequency.

Table 3

The Common Negation Cues and Their Frequencies

Negation	Frequency	Negation Cue	Frequency
لا	91,267	مو	23,723
لم	66,056	مب - موب	6,413
ما	39,893	مش	4,210
غير	25,969	مفي - مافي - ما في	2,198
لن	18,888	مفيه - مافيه - ما فيه	897
عدم – عديم	4,948	محد	671
ليس	6,892	منو – مانو – ماني	356
دون	3,923	ماش – بلاش – كنش	228
لما	3,345	ماهو - ماهي - مهي	201
لات	2,809	ماكو	133
عدا – ماعدا	413	مافيها - ما فيها	120
بتاتا – البتة	400	لامبالاة – لاإرادي -	106
Negation words also take place by: • (م) or (ما) as a prefix • (ش) as a suffix • ((م) or (ما) as a prefix) and (ش) as a suffix			2,319

Text Pre-processing

Valuable textual data on web pages is often unstructured, and directly applying negation detection to such data may yield poor results. Therefore, pre-processing techniques are crucial to enhance data quality and aid in negation detection (Aldayel & Azmi, 2016). To reduce feature dimensions, the study implemented key pre-processing steps in the Arabic text corpus, including text cleaning, tokenisation, stopword removal, term stemming, and pruning steps. Term stemming reduces words to their base form, while pruning reduces data

dimensionality. Table 4 summarises the Text Pre-processing Steps using the RapidMiner tool (Altair, n.d.).

Table 4

Text Pre-processing Steps

Step	Description
Text Cleaning	Using regular expressions, remove irrelevant elements (e.g., usernames, hashtags, URLs).
Tokenisation	Divide the text into tokens (words or sentences) to identify boundaries.
Stopwords Removal	Eliminate non-discriminatory terms like articles and conjunctions to reduce feature space.
Stemming	Reduce words to their base form using light stemming to preserve meanings.
Pruning	Remove infrequent words (occurring fewer than 15 times) to reduce dimensionality.

Features Extraction

Features extraction in machine learning involves selecting appropriate features to effectively detect negation in text. Two types of features are extracted: lexical and structural. Lexical features- like unigrams, are crucial for analysing word usage and frequency and are essential for negation detection. The unigram model represents individual words in a review, and the study utilises the TF-IDF model to assess word importance (Alotaibi 2015). It included words occurring over 15 times to manage dimensionality, resulting in 2,079 distinct features. On the other hand, structural features - aim to understand the structure of Arabic opinion reviews and identify negation presence. These features include the number of negation words, length-based, punctuation-based, and PoS-based features. A total of 17 structural features were extracted using the Python programming language. Table 5 summarises these structural features.

Table 5

Description of the Structural Feature

Group	Features	Description
Feature of the Number of Negation Words	No. of NW	The total number of negation words in the review.
Length Features	LengthWords LengthChars LengthSentences	The total number of sentences, words and characters in the review, respectively.
Punctuation-Based Features	Question Exclamation Colons Semicolons Commas Full stops Quotation Ellipsis	The number of each punctuation mark in the review.
PoS Based-Features	No. of PM Nouns Adjectives Adverbs Verbs	The number of each PoS-tag in the review.

From Table 5, the feature of the number of negation words indicates the presence and frequency of negation within a text, which often correlates with more negative sentiment (El-Dine & El-Zahraa, 2013). Higher occurrences of negation words suggest 'negated positive' sentiments, aiding in more accurate SA. The length features measure sentence, word, and character lengths in reviews, assuming that negated positive reviews might be longer due to creative language use. The punctuation-based features represent punctuation marks, including question marks, exclamation marks, colons, semicolons, commas, full stops, quotation marks, and ellipses, which play a role in text readability and message conveyance (Farra et al., 2010; Reitan et al., 2015). Higher usage of certain punctuation marks, like multiple exclamations or ellipses, can indicate 'negated positive' content. The study collected punctuation-related properties. These features include the number of question marks, exclamation marks, colons, semicolons, commas, full stops, quotation marks, ellipses, and overall punctuation marks in a review.

Finally, the part-of-speech (PoS)-based features involve assigning PoS tags to tokens in a text, such as nouns, adjectives, adverbs, and verbs (Farra et al., 2010; Reitan et al., 2015). Specific linguistic patterns, like excessive use of intensifiers or reduced occurrence of verbs, in negated positive contexts can be identified through PoS tagging, aiding in the detection of negation (El-Dine & El-Zahraa, 2013; Farra et al., 2010; Patodkar & Sheikh, 2016). Features related to the number of nouns, adjectives, adverbs, and verbs in reviews are extracted to pinpoint certain word types in negation utterances.

Classification

For classifying reviews, the study selected five classifiers, namely NB, RF, LogR, SVM, and DL implemented using H2O.ai, based on their effectiveness in text classification tasks, as supported by previous research (See Table 6). The study employed the RapidMiner tool to execute these classifiers, which offers a wide range of machine learning algorithms, including NB, RF, LogR, SVM, DL, and others. RapidMiner also offers diverse testing methodologies, like split validation and cross-validation. The study harnessed feature vectors derived from the review data to train these algorithms and construct the classification model (Altair, n.d.).

Naïve Bayes (NB)

Using training data, the NB classifier estimates the probabilities of variable values within a class and applies these probabilities to classify new entities (Duda et al., 2012; Han et al., 2011; Witten & Frank, 2009). It relies on Bayes' theorem and assumes independence between features within a class (Han et al., 2011; Larkey et al., 2002; Silva & Ribeiro, 2003). This simplicity and efficiency make it suitable for high-dimensional datasets without complex parameter estimation methods (Han et al., 2011; Larkey et al., 2002; Silva & Ribeiro, 2003). NB is widely used in document classification due to its consistently outstanding performance.

Random Forest (RF)

The RF classifier, an ensemble learning approach, enhances model performance by combining multiple classifiers. It constructs numerous decision trees on different subsets of the dataset, averages their predictions, and improves accuracy. RF aggregates predictions from each decision tree and relies on the forecast with the most votes.

With more trees in the forest, this ensemble method achieves higher accuracy and mitigates overfitting compared to individual decision trees (Javatpoint, n.d.).

Logistic Regression (LogR)

LogR is a mathematical modelling technique that describes the relationship between independent variables and a binary response variable (Martín-Valdivia et al., 2011). It builds a probabilistic model using data, fitting a logistic function to represent the class distribution (Hosmer et al., 2013). Each training instance is assigned a weight vector and processed through the logistic function, often depicted as a sigmoid function (Raeder, 2016).

Support Vector Machines (SVM)

SVM are robust classifiers widely used for binary classification tasks. They analyse data and identify patterns by creating a discriminative classifier with a separating hyperplane (Duda et al., 2012; Han et al., 2011; Witten & Frank, 2009). SVMs excel in learning tasks due to their fast algorithm and proven effectiveness. In SVM, examples are represented as points in space, separated by a substantial gap, ensuring instances from different categories are distinctly classified based on their position relative to the gap (Duda et al., 2012; Han et al., 2011; Witten & Frank, 2009).

Deep Learning (DL)

DL is a powerful classification technique that leverages artificial neural networks with multiple layers to extract intricate patterns from data. Unlike traditional models, DL excels in capturing complex relationships within data, making it ideal for challenging classification tasks. DL autonomously learns valuable features from raw data, eliminating the need for manual feature engineering. Its scalability and adaptability make it suitable for tasks with extensive datasets. DL models consist of interconnected neurons organised into layers, processing different aspects of input data. Through iterative training, these networks adjust internal parameters to minimise prediction errors, continually improving accuracy and generalisation to new data. DL has demonstrated remarkable performance in various applications like image recognition, speech recognition, and natural language processing, establishing itself as a crucial tool in modern machine learning (Deng & Yu, 2014; Goodfellow et al., 2016; Lee, 2018).

Table 6

Classifiers and Their Rationale

Classifier	Rationale
NB	Proven effective for high-dimensional datasets; assumes feature independence (Duda et al., 2012; Han et al., 2011).
RF	Combines multiple decision trees to enhance accuracy and reduce overfitting (Javatpoint, n.d.).
LogR	Models of binary outcomes with a probabilistic approach are well-suited for classification tasks (Martín-Valdivia et al., 2011; Hosmer et al., 2013).
SVM	It creates a hyperplane for optimal separation of classes and is effective for binary classification (Duda et al., 2012; Han et al., 2011).
DL	It captures complex patterns through neural networks and excels in large, intricate datasets (Deng & Yu, 2014; Goodfellow et al., 2016).

EVALUATION AND RESULTS

This section explains the experiments carried out to investigate and test machine learning classifiers for negation detection. It showcases experimental outcomes, their evaluation, and a discussion of the results to support the feasibility of the proposed approach.

Evaluation

The evaluation of classifiers involves assessing performance using specific metrics. A common method is the confusion matrix, which helps evaluate classification accuracy by comparing actual and predicted classifications. Four key metrics derived from the confusion matrix are accuracy, precision, recall, and the F-measure. Accuracy gauges overall classification correctness, and precision measures the relevance of 'negated positive' reviews. Recall assesses the ability to identify relevant 'negated positive' reviews, and the F-measure combines precision and recall for a standardised evaluation of classifier performance (Witten & Frank, 2009).

Experimental Setup

This section outlines the experimental setup used to assess the proposed approach for identifying 'negated positive' reviews. The corpus

consisted of 84,000 Arabic opinion reviews, evenly split between 'negated positive' and positive reviews, with 70 percent allocated for training and 30 percent for testing. Feature extraction generated vectors used to train supervised machine learning algorithms, including NB, RF, LogR, SVM, and DL. Three sets of experiments were conducted using different feature sets: lexical features, structural features, and a combination of both. A baseline experiment using simple lexical features, specifically the unigram model, was established. Performance metrics such as accuracy, precision, recall, and F-measure were computed to evaluate the classifiers. To conduct the evaluation for identifying 'negated positive' reviews, the following tools and platforms were employed:

1) Python was the primary programming language used for executing the experimental setup, including data processing, feature extraction, and the evaluation of classification models.
2) NLTK was used for text pre-processing tasks, such as tokenisation, stopword removal, and text cleaning. It also facilitated the extraction of lexical features, like unigrams, and the preparation of the dataset for classification.
3) RapidMiner was used to implement machine learning algorithms and run the experiments. It provided support for classifiers such as NB, RF, LogR, SVM, and DL, and offered a user-friendly interface for feature extraction, cross-validation, and performance metric calculations, including accuracy, precision, recall, and F-measure.

Together, these tools ensured a smooth and efficient workflow for training and evaluating the classifiers, enabling the extraction and analysis of both lexical and structural features.

Results

This section displays the results and analysis of a series of experiments that were conducted. The main objective of these experiments was to determine the feature sets and machine learning classifiers that are most efficient in detecting negation in Arabic reviews.

Experiments with The Lexical Features (Baseline Experiments)

In these experiments, baseline results were established for comparison with subsequent experiments. The chosen baseline provides

fundamental knowledge about the text and preserves its primary semantic features. These experiments involved various machine learning classifiers and consisted of 5 trials. The feature set used in this baseline model comprises 2,079 distinct features.

Table 7 summarises the performance metrics for the 5 classifiers of the baseline experiments. The study calculated accuracy, precision, recall, and F-measure using split validation. The boldest values represent the top outcomes achieved across all feature sets and classifiers, whereas the underlined values signify the best results attained using the baseline experiments.

Table 7

Performance Metrics for the 5 Classifiers of the Baseline Experiments

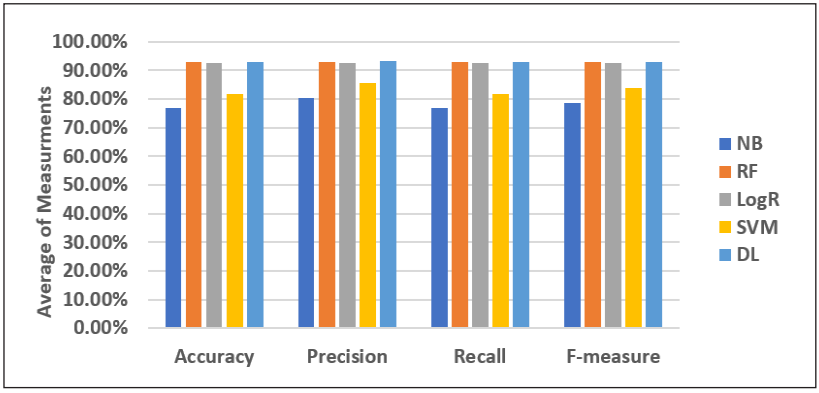
			NB	
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	94.65%	94.66%	94.66%	94.66%
			RF	
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	96.68%	96.68%	96.68%	96.68%
			LogR	
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	97.50%	97.50%	97.51%	97.50%
			SVM	
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**98.10%**	**98.10%**	**98.10%**	**98.10%**
			DL	
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	98.05%	98.06%	98.05%	98.05%

Figure 2 visually represents the performance metrics associated with the baseline experiments. As indicated in Table 5 and Figure 2, the accuracy across experiments ranges from 94.65 percent to 98.10 percent, with the F-measure falling within the same range, which is considered quite good. The SVM classifier achieved the highest overall accuracy at 98.10 percent, accompanied by an F-measure of

98.10 percent. On the other hand, the NB, RF, LogR, and DL classifiers exhibited similar performance in the task, with slight variations, although all of them fell short of the SVM classifier's performance. This underscores the suitability of employing a machine learning approach for negation identification.

Figure 2

Performance Metrics for the 5 Classifiers in Experiments Using the Baseline



Furthermore, these outcomes imply that fundamental lexical features, constituting the baseline, offer valuable information for the task of negation identification. These findings will be utilised as a baseline to evaluate different feature sets in the upcoming experiments. Additionally, they highlight the classifier's ability to acquire new knowledge from additional features.

Experiments with The Structural Features

In these experiments, various structural features were constructed, such as the number of negation words in the review, length features, punctuation-based features, and PoS-based features. The objective was to assess their impact on various machine learning classifiers for Arabic negation detection. These experiments aimed to evaluate how structural information from the reviews affected the feature model. Additionally, this analysis helped us explore the effect of using these features independently for the first time in Arabic negation detection. The study conducted 5 different machine learning classifiers in these evaluations, resulting in a feature set comprised of 17 various features.

Table 8 summarises the performance metrics for the 5 classifiers of both the baseline and structural features. The study calculated accuracy, precision, recall, and F-measure using split validation. The boldest values indicate the top outcomes achieved across all feature sets and classifiers, whereas the underlined values signify the best results attained using a specific feature model for each classifier.

Table 8

Performance Metrics for the 5 Classifiers of Both the Baseline and Structural Features

	NB			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**94.65%**	**94.66%**	**94.66%**	**94.66%**
Structural Features	76.82%	80.49%	76.91%	78.66%
	RF			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**96.68%**	**96.68%**	**96.68%**	**96.68%**
Structural Features	92.85%	92.86%	92.86%	92.86%
	LogR			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**97.50%**	**97.50%**	**97.51%**	**97.50%**
Structural Features	92.50%	92.66%	92.52%	92.59%
	SVM			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**98.10%**	**98.10%**	**98.10%**	**98.10%**
Structural Features	81.78%	85.73%	81.87%	83.76%
	DL			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	**98.05%**	**98.06%**	**98.05%**	**98.05%**
Structural Features	93.01%	93.16%	93.03%	93.09%

Figure 3 visually represents the performance metrics associated with the structural features. Table 8 and Figure 3 show that the accuracy ranged from 76.82 percent to 93.01 percent, and the F-measure ranged from 94.65 percent to 98.10 percent in the structural features classification

experiments. These results are relatively lower compared to those obtained in the classification experiments using lexical features. For example, in the case of the NB classifier, the accuracy and F-measure were 94.65 percent and 94.66 percent, respectively, in experiments with lexical features, whereas with structural features, the accuracy dropped to 76.82 percent and the F-measure was 78.66 percent. The highest overall accuracy achieved using structural features was 93.01 percent, obtained using the DL classifier, with an F-measure of 93.09 percent. The NB, RF, LogR, and SVM classifiers demonstrated similar classification performance in experiments with structural features, with slight differences. The lowest overall accuracy using structural features was 76.82 percent, achieved with the NB classifier, and the F-measure was 78.66 percent.

Figure 3

Performance Metrics for the 5 Classifiers in Experiments Using Structural Features

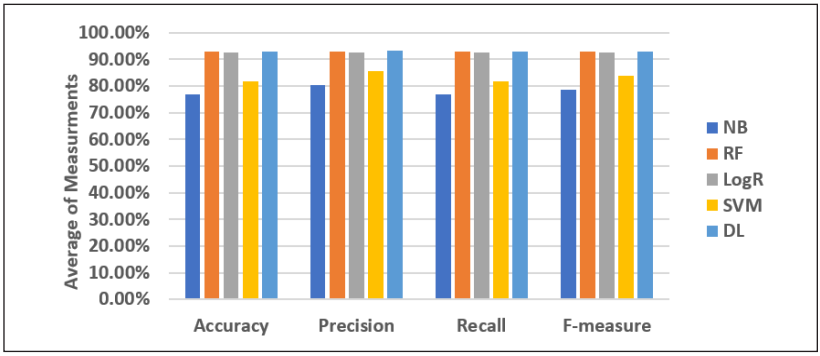


Table 8 demonstrate that the baseline uni-gram model consistently outperforms the structural features alone across all classifiers. This suggests that the uni-gram model remains a powerful choice for Arabic negation detection and is considered the most effective model for machine learning classification. The uni-gram model offers comprehensive data coverage by capturing the fundamental unit of sentences: individual words.

Experiments with a Combination of Lexical Features and the Types of Structural Features

In these experiments, various feature combinations were explored, including lexical and structural features like the number of negation

words, length features, punctuation-based features, and PoS-based features. The objective was to evaluate how these feature combinations impacted the performance of various machine learning classifiers in the context of Arabic negation detection.

15 feature sets were created by combining various features to conduct these experiments. The sets are listed below:

1) Uni-gram + Negation Words- This set included the uni-gram model and the total number of negation words in the review.
2) Uni-gram + Length Features- The study added the length-related features (total sentences, words, and characters) to the uni-gram model.
3) Uni-gram + Punctuation Features- 9 punctuation-related features, representing the number of each punctuation mark, were added to the uni-gram model.
4) Uni-gram + PoS Features- This set incorporated 4 features representing the number of each PoS tag into the uni-gram model.
5) Uni-gram + Negation Words + Length Features.
6) Uni-gram + Negation Words + Punctuation Features.
7) Uni-gram + Negation Words + PoS Features.
8) Uni-gram + Length Features + Punctuation Features.
9) Uni-gram + Length Features + PoS Features.
10) Uni-gram + Punctuation Features + PoS Features.
11) Uni-gram + Negation Words + Length Features + Punctuation Features.
12) Uni-gram + Negation Words + Length Features + PoS Features.
13) Uni-gram + Negation Words + Punctuation Features + PoS Features.
14) Uni-gram + Length Features + Punctuation Features + PoS Features.
15) Uni-gram + All Structural Features- This set combined all structural features, including negation words, length, punctuation, and PoS features.

The main objectives were to assess how incorporating structural knowledge into the uni-gram baseline model influenced its performance, identify the most effective feature combinations, and evaluate the impact of these features on Arabic negation detection. These experiments involved multiple machine learning classifiers, resulting in 75 experiments.

Table 9 summarises the performance metrics for the 5 classifiers of both the baseline trials and the experiments with the 15 feature

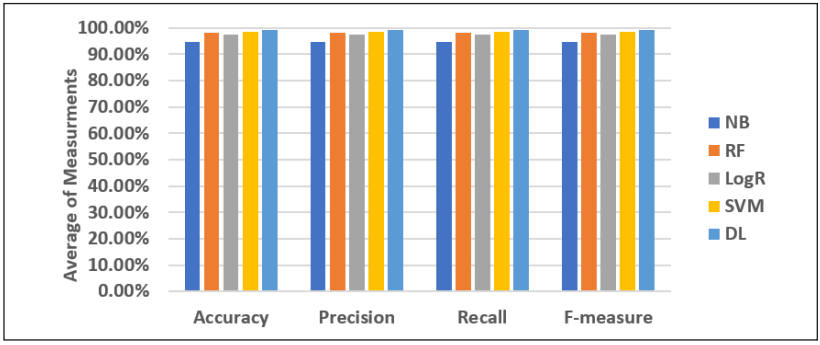
sets. Metrics such as accuracy, precision, recall, and F-measure were calculated using split validation. In Tables 9a, 9b, 9c, 9d, and 9e, bold values highlight the highest performance across all feature sets and classifiers, while underlined values denote the best results achieved with each specific feature model for each classifier.

As indicated in Tables 9a, 9b, 9c, 9d, and 9e, the accuracy and F-measure in classification experiments using a combination of lexical and structural features ranged from 94 percent to 99 percent. This represented a notable improvement compared to experiments using lexical features alone. For instance, the RF classifier realised an accuracy and F-measure of 96.68 percent with lexical features, but these numbers increased to 98.62 percent when combining lexical and structural features.

Figure 4 provides a visual representation of the performance metrics associated with the 'uni-gram + negation words + length + punctuation + PoS' feature set, which achieved the best results among all feature sets.

Figure 4

Performance Metrics for the 5 Classifiers in Experiments Utilising the 'Uni-gram +Number of Negation Words + Length + Punctuation + PoS' Feature Set



As shown in Figure 4, the DL classifier consistently outperformed other classifiers across all feature sets. The DL and SVM classifiers achieved the highest overall accuracies of 99.24 percent and 98.67 percent, respectively, both corresponding to F-measures equal to their accuracies. While the NB, RF, and LogR classifiers exhibited similar performance, they generally scored lower than the DL and SVM classifiers.

In conclusion, the DL classifier consistently outperformed other classifiers across various feature sets. The DL classifier, in combination with the 'uni-gram + negation words + length + punctuation + PoS' feature set, delivered the top performance, achieving an exceptional overall accuracy and F-measure of 99.24 percent. Adding structural features to the uni-gram baseline model led to a notable enhancement in classification performance across all classifiers. This emphasises the appropriateness of utilising a machine learning approach to detect negation in Arabic texts and underscores the effectiveness of feature combinations that encompass both fundamental lexical features and structural characteristics for this purpose.

Table 9a

Performance Metrics for the NB Classifier of the Baseline and Various Feature Combinations Experiments

	NB			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	94.65%	94.65%	94.65%	94.65%
Uni-gram + No. of negation words	**94.69%**	**94.69%**	**94.69%**	**94.69%**
Uni-gram + Length	94.66%	94.66%	94.66%	94.66%
Uni-gram + Punctuation	94.64%	94.64%	94.64%	94.64%
Uni-gram + PoS	94.65%	94.65%	94.65%	94.65%
Uni-gram + No. of negation words + Length	**94.69%**	**94.69%**	**94.69%**	**94.69%**
Uni-gram + No. of negation words + Punctuation	94.68%	94.68%	94.68%	94.68%
Uni-gram + No. of negation words + PoS	**94.69%**	**94.69%**	**94.69%**	**94.69%**
Uni-gram + Length + Punctuation	94.63%	94.63%	94.63%	94.63%
Uni-gram + Length + PoS	94.65%	94.65%	94.65%	94.65%
Uni-gram + Punctuation + PoS	94.63%	94.63%	94.63%	94.63%
Uni-gram + No. of negation words + Length + Punctuation	94.67%	94.67%	94.67%	94.67%
Uni-gram + No. of negation words + Length + PoS	**94.69%**	**94.69%**	**94.69%**	**94.69%**
Uni-gram + No. of negation words + Punctuation + PoS	94.67%	94.67%	94.67%	94.67%
Uni-gram + Length + Punctuation + PoS	94.63%	94.63%	94.63%	94.63%
Uni-gram + No. of negation words + Length + Punctuation + PoS	94.67%	94.67%	94.67%	94.67%

Table 9b

Performance Metrics for the RF Classifier of the Baseline and Various Feature Combinations Experiments

	RF			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	96.68%	96.68%	96.68%	96.68%
Uni-gram + No. of negation words	98.38%	98.38%	98.38%	98.38%
Uni-gram + Length	96.73%	96.73%	96.73%	96.73%
Uni-gram + Punctuation	96.85%	96.85%	96.85%	96.85%
Uni-gram + PoS	96.58%	96.58%	96.58%	96.58%
Uni-gram + No. of negation words + Length	98.52%	98.52%	98.52%	98.52%
Uni-gram + No. of negation words + Punctuation	98.57%	98.57%	98.57%	98.57%
Uni-gram + No. of negation words + PoS	**98.62%**	**98.62%**	**98.62%**	**98.62%**
Uni-gram + Length + Punctuation	96.83%	96.83%	96.83%	96.83%
Uni-gram + Length + PoS	96.66%	96.66%	96.66%	96.66%
Uni-gram + Punctuation + PoS	96.84%	96.84%	96.84%	96.84%
Uni-gram + No. of negation words + Length + Punctuation	98.34%	98.34%	98.34%	98.34%
Uni-gram + No. of negation words + Length + PoS	98.60%	98.60%	98.60%	98.60%
Uni-gram + No. of negation words + Punctuation + PoS	98.35%	98.35%	98.35%	98.35%
Uni-gram + Length + Punctuation + PoS	96.60%	96.60%	96.60%	96.60%
Uni-gram + No. of negation words + Length + Punctuation + PoS	98.25%	98.25%	98.25%	98.25%

Table 9c

Performance Metrics for the LogR Classifier of the Baseline and Various Feature Combinations Experiments

	LogR			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	97.50%	97.50%	97.50%	97.50%
Uni-gram + No. of negation words	**98.08%**	**98.08%**	**98.08%**	**98.08%**
Uni-gram + Length	97.70%	97.70%	97.70%	97.70%
Uni-gram + Punctuation	97.60%	97.60%	97.60%	97.60%
Uni-gram + PoS	97.55%	97.55%	97.55%	97.55%
Uni-gram + No. of negation words + Length	97.80%	97.80%	97.80%	97.80%
Uni-gram + No. of negation words + Punctuation	98.06%	98.06%	98.06%	98.06%
Uni-gram + No. of negation words + PoS	97.62%	97.62%	97.62%	97.62%
Uni-gram + Length + Punctuation	97.58%	97.58%	97.58%	97.58%
Uni-gram + Length + PoS	97.66%	97.66%	97.66%	97.66%
Uni-gram + Punctuation + PoS	97.51%	97.51%	97.51%	97.51%
Uni-gram + No. of negation words + Length + Punctuation	97.51%	97.51%	97.51%	97.51%
Uni-gram + No. of negation words + Length + PoS	97.72%	97.72%	97.72%	97.72%
Uni-gram + No. of negation words + Punctuation + PoS	97.67%	97.67%	97.67%	97.67%
Uni-gram + Length + Punctuation + PoS	97.61%	97.61%	97.61%	97.61%
Uni-gram + No. of negation words + Length + Punctuation + PoS	97.61%	97.61%	97.61%	97.61%

Table 9d

Performance Metrics for the SVM Classifier of the Baseline and Various Feature Combinations Experiments

	SVM			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	98.10%	98.10%	98.10%	98.10%
Uni-gram + No. of negation words	98.60%	98.60%	98.60%	98.60%
Uni-gram + Length	98.22%	98.22%	98.22%	98.22%
Uni-gram + Punctuation	98.19%	98.19%	98.19%	98.19%
Uni-gram + PoS	98.12%	98.12%	98.12%	98.12%
Uni-gram + No. of negation words + Length	98.67%	98.67%	98.67%	98.67%
Uni-gram + No. of negation words + Punctuation	98.67%	98.67%	98.67%	98.67%
Uni-gram + No. of negation words + PoS	98.63%	98.63%	98.63%	98.63%
Uni-gram + Length + Punctuation	98.21%	98.21%	98.21%	98.21%
Uni-gram + Length + PoS	98.17%	98.17%	98.17%	98.17%
Uni-gram + Punctuation + PoS	98.19%	98.19%	98.19%	98.19%
Uni-gram + No. of negation words + Length + Punctuation	98.68%	98.68%	98.68%	98.68%
Uni-gram + No. of negation words + Length + PoS	98.64%	98.64%	98.64%	98.64%
Uni-gram + No. of negation words + Punctuation + PoS	**98.70%**	**98.70%**	**98.70%**	**98.70%**
Uni-gram + Length + Punctuation + PoS	98.24%	98.24%	98.24%	98.24%
Uni-gram + No. of negation words + Length + Punctuation + PoS	98.67%	98.67%	98.67%	98.67%

Table 9e

Performance Metrics for the DL Classifier of the Baseline and Various Feature Combinations Experiments

	DL			
	Accuracy	Precision	Recall	F-measure
Uni-gram Baseline	98.05%	98.05%	98.05%	98.05%
Uni-gram + No. of negation words	98.94%	98.94%	98.94%	98.94%
Uni-gram + Length	98.10%	98.10%	98.10%	98.10%
Uni-gram + Punctuation	97.97%	97.97%	97.97%	97.97%
Uni-gram + PoS	98.17%	98.17%	98.17%	98.17%
Uni-gram + No. of negation words + Length	99.19%	99.19%	99.19%	99.19%
Uni-gram + No. of negation words + Punctuation	99.15%	99.15%	99.15%	99.15%
Uni-gram + No. of negation words + PoS	99.11%	99.11%	99.11%	99.11%
Uni-gram + Length + Punctuation	98.05%	98.05%	98.05%	98.05%
Uni-gram + Length + PoS	97.98%	97.98%	97.98%	97.98%
Uni-gram + Punctuation + PoS	98.32%	98.32%	98.32%	98.32%
Uni-gram + No. of negation words + Length + Punctuation	99.13%	99.13%	99.13%	99.13%
Uni-gram + No. of negation words + Length + PoS	99.11%	99.11%	99.11%	99.11%
Uni-gram + No. of negation words + Punctuation + PoS	99.21%	99.21%	99.21%	99.21%
Uni-gram + Length + Punctuation + PoS	98.38%	98.38%	98.38%	98.38%
Uni-gram + No. of negation words + Length + Punctuation + PoS	**99.24%**	**99.24%**	**99.24%**	**99.24%**

Discussions

Performance of Classifiers

The experiments demonstrate that using a combination of lexical and structural features significantly improves the performance of classifiers in detecting negated positive reviews in Arabic. The DL and SVM classifiers consistently outperformed others, with the DL classifier achieving the highest accuracy and F-measure of 99.24 percent when using the combined feature set of uni-gram, number of negation words, length, punctuation, and PoS features. The performance metrics indicate that the inclusion of structural features alongside lexical features enhances the classifiers' ability to detect negation. Specifically, the DL classifier's remarkable performance suggests that deep learning models can effectively leverage complex feature interactions, which traditional machine learning classifiers may not as effectively utilise.

The Importance of Feature Sets

The baseline experiments with uni-gram lexical features provided a strong foundation, achieving up to 98.10 percent accuracy with the SVM classifier. However, structural features alone did not perform as well, with the highest accuracy of 93.01 percent achieved by the DL classifier. This indicates that while structural features contribute valuable information, they are most effective when combined with lexical features. The experiments also highlight the importance of specific feature combinations. For instance, combining uni-gram with negation words and length features consistently improved performance across all classifiers. This suggests that certain structural features, like the count of negation words and review length, are particularly influential in detecting negated sentiments.

Limitations

While the experiments demonstrate the effectiveness of various machine learning classifiers and feature sets for negation detection in Arabic reviews, the following limitations should be considered. First, the experiments were conducted on a specific dataset of 84,000 Arabic opinion reviews. The performance metrics might not generalise to other datasets with different characteristics or domains.

Future work should test the classifiers on diverse datasets to validate their robustness. Second, although the study explored a variety of lexical and structural features, there may be other useful features that were not considered in this study. For example, semantic features or contextual embeddings from models like BERT could improve performance further. The limited scope of feature types might have restricted the full potential of the classifiers. Finally, the deep learning model achieved the highest accuracy but also requires significant computational resources for training and inference compared to traditional machine learning models like NB or LogR. This might limit the practicality of deploying such models in resource-constrained environments. By addressing these limitations, future work can aim to develop more robust, generalisable, and interpretable models for negation detection in Arabic reviews, potentially exploring additional features and validating across various datasets.

CONCLUSIONS

This study proposed an approach for detecting negation in user-generated Arabic hotel reviews through lexical and structural features. It encompassed several stages: data collection, pre-processing, feature extraction, supervised learning classification, and evaluation. The dataset utilised in the experiments was sourced from prominent online platforms specialising in Arabic economic content, namely TripAdvisor, Booking.com, and Agoda. This extensive corpus consisted of 84,000 Arabic opinion reviews, evenly divided between 42,000 'negated positive' reviews and 42,000 positive reviews. The reviews encompassed both Modern Standard Arabic (MSA) and Dialectal Arabic (DA), as well as a mixture of MSA and DA. Throughout the research, a series of experiments were conducted aimed at determining the most effective feature sets and machine learning classifiers for detecting negation in Arabic hotel reviews. Five machine learning classifiers—Naive Bayes (NB), Random Forest (RF), Logistic Regression (LogR), Support Vector Machine (SVM), and Deep Learning (DL)—were evaluated. This approach leveraged both lexical and structural features, including the number of negation words, length, punctuation-based features, and Part-of-Speech (PoS) tags, in the supervised machine-learning process. The study employed key metrics such as accuracy, precision, recall, and F-measure to assess classifier performance. The results of the experiments have

yielded promising outcomes, demonstrating the feasibility of the approach for practical applications. The classifiers exhibited highly comparable performance in identifying negation, with only marginal deviations in their performance metrics. Particularly noteworthy, the DL classifier consistently emerged as the top performer, achieving an exceptionally high overall accuracy rate of 99.24 percent, surpassing established benchmarks in Arabic text processing and underscoring its potential for practical applications. These findings hold significant implications within the realm of Arabic text processing.

Concerning the upcoming research directions, the study intends to implement this approach on social networking platforms like Twitter and Facebook. This choice stems from the striking resemblance in the fundamental structure of these platforms. Furthermore, the study plans to explore the utilisation of lexicons and incorporate new features to enhance the capability to distinguish more effectively between 'negated positive' reviews and positive reviews. In the forthcoming investigations, the study is expected to assess the effectiveness of the approach using datasets that exhibit imbalance and real-world data, alongside an evaluation of its performance on more extensive corpora spanning various domains.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Abuhammad, S., & Ahmed, M. A. (2023). Negation detection techniques in sentiment analysis: A survey. *International Journal of Applied Science and Engineering, 20*(2), 1–7. https://doi.org/10.6703/ijase.202306_20(2).003

Aldayel, H. K., & Azmi, A. M. (2016). Arabic tweets sentiment analysis: A hybrid scheme. *Journal of Information Science, 42*(6), 782–797. https://doi.org/10.1177/0165551515610513

Alharbi, O. (2020). Negation handling in machine learning-based sentiment classification for colloquial Arabic. *International Journal of Operations Research and Information Systems, 11*(4), 33–45. https://doi.org/10.4018/ijoris.2020100102

Alotaibi, S. S. (2015). *Sentiment analysis in the Arabic language using machine learning* [Master's thesis, Colorado State University]. Mountain Scholar. https://mountainscholar.org/handle/10217/167091

Altair. (n.d.). In *RapidMiner: Data analytics and AI platform***.** http://rapidminer.com/

Burbach, L., Halbach, P., Ziefle, M., & Calero Valdez, A. (2020). Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. *Frontiers in Artificial Intelligence, 3*, 45. https://doi.org/10.3389/frai.2020.00045

Collins Dictionary (2023). Negation definition & meaning. In *Collins Dictionary*. https://www.collinsdictionary.com/dictionary/english/negation

Councill, I. G., McDonald, R., & Velikovich, L. (2010). What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP)* (pp. 51–59). Association for Computational Linguistics. https://aclanthology.org/2010.nespnlp-1.7

Deng, L., & Yu, D. (2014). *Deep learning*. Springer.

Dictionary.com. (n.d.). Negation. In *Lexico*. https://www.lexico.com/definition/negation

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification***.** John Wiley & Sons.

El-Dine, A., & El-Zahraa, F. (2013). Sentiment analyser for Arabic comments system. *International Journal of Advanced Computer Science and Applications, 4*(3). https://doi.org/10.14569/ijacsa.2013.040317

Eremyan, R. (2023). Four pitfalls of sentiment analysis accuracy. *Toptal Engineering Blog*. https://www.toptal.com/deep-learning/4-sentiment-analysis-accuracy-traps

Farooq, U. (2017). Negation handling in sentiment analysis at sentence level. *Journal of Computers, 12*(5), 470–478. https://doi.org/10.17706/jcp.12.5.470-478

Farra, N., Challita, E., Assi, A., & Hajj, H. (2010). Sentence-level and document-level sentiment mining for Arabic texts. In *Data Mining Workshops (ICDMW) – IEEE International Conference Proceedings* (pp. 1114–1119).

Funkner, A., Balabaeva, K., & Kovalchuk, S. (2020). Negation detection for clinical text mining in Russian. *Studies in Health Technology and Informatics, 270*, 43–47. https://doi.org/10.3233/SHTI200889

Genadi, R. A., & Khodra, M. L. (2022). Opinion Triplet Extraction for Aspect-Based Sentiment Analysis Using Co-Extraction Approach. *Journal of Information and Communication Technology, 21*(2), 255–277. https://doi.org/10.32890/jict2022.21.2.5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://books.google.ie/books?id=omivDQAAQBAJ&printsec=frontcover&dq=Deep+Learning&hl=&cd=1&source=gbs_api

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://books.google.com/books?id=bRoxQBIZRd4C

Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges**.** *Journal of King Saud University - Engineering Sciences, 30*(4), 330–338. https://doi.org/10.1016/j.jksues.2016.04.002

Javatpoint. (n.d.). *Machine learning random forest algorithm*. https://www.javatpoint.com/machine-learning-random-forest-algorithm

Larkey, L., Ballesteros, L., & Connell, M. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275-282). ACM.

Lee, S. (2018). Application of artificial neural networks in geoinformatics. *MDPI*. http://books.google.ie/books?id=2MFUDwAAQBAJ&printsec=frontcover&dq=Artificial+Neural+Networks+Handbook+of+Research+on+Geoinformatics&hl=&cd=2&source=gbs_api

Mahany, A., Fouad, M. M., Aloraini, A., Khaled, H., Nawaz, R., Aljohani, N. R., & Ghoniemy, S. (2021). Supervised learning for negation scope detection in Arabic texts. In *IEEE 10th International Conference on Intelligent Computing and Information Systems (ICICIS) – Conference Proceedings* (pp. 177–182). Cairo, Egypt.

Martín-Valdivia, M., Montejo-Ráez, A., Ureña-López, L. A., & Saleh, M. (2012). Learning to classify neutral examples from positive and negative opinions. *Journal of Universal Computer Science, 18*(16), 2319–2333.

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 174–179). San Diego, California.

Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S., & Sharma, R. (2021). Effect of negation in sentences on sentiment analysis and polarity detection**.** *Procedia Computer Science, 185***,** 370–379. https://doi.org/10.1016/j.procs.2021.05.038

Patodkar, V. N., & Sheikh I. R. (2016). Twitter as a corpus for sentiment analysis and opinion mining. *International Journal of Advanced Research in Computer and Communication Engineering, 5*(12), 320–322. https://doi.org/10.17148/ijarcce.2016.51274

Raeder, J. (2016). *Automatic sarcasm detection in Twitter messages* [Master's thesis, Norwegian University of Science and Technology, Department of Computer and Information Science].

Reitan, J., Faret, J., Gambäck, B., & Bungum, L. (2015). Negation scope detection for Twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 99–108).

Saad, M. (2010). *The impact of text pre-processing and term weighting on Arabic text classification* [Master's thesis, Department of Computer Engineering, The Islamic University-Gaza].

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorisation. In *Proceedings of the International Joint Conference on Neural Networks* (Vol. 3, pp. 1661-1666). Portland, OR, USA.

Witten, I. H., & Frank, E. (2009). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann. http://books.google.ie/books?id=ic9MPgAACAAJ&dq=Data+Mining:+Practical+Machine+Learning+Tools+and+Techniques&hl=&cd=9&source=gbs_api

Wikipedia (2023a). *Arabic*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Arabic

Wikipedia (2023b). *Lexical analysis*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization

World Internet Users' Statistics and 2023 World Population Stats. (n.d.). *Internet World Stats***.** https://www.internetworldstats.com/stats.htm