Arabic Sentiment Analysis using Dependency-Based Rules and Deep Neural Networks

Arwa Diwali [a], Kia Dashtipour [b], Kawther Saeedi [c] , Mandar Gogate [d] , Erik Cambria [e] , Amir Hussain [f]

[a] School of Computing Edinburgh Napier University, Edinburgh, UK, arwa.diwali@napier.ac.uk

[a] Faculty of Computing and Information Technology King Abdulaziz University, Jeddah , KSA, adiwali@kau.edu.sa

[b] School of Computing Edinburgh Napier University, Edinburgh, UK, k.dashitpour@napier.ac.uk

[c] Faculty of Computing and Information Technology King Abdulaziz University, KSA, ksaeedi@kau.edu.sa

[d] School of Computing Edinburgh Napier University, Edinburgh, UK, m.gogate@napier.ac.uk

[e] Nanyang Technological University, Singapore, cambria@ntu.edu.sg

[f] School of Computing Edinburgh Napier University, UK, a.hussain@napier.ac.uk

Abstract

With the growth of social platforms in recent years and the rapid increase in the means of communication through these platforms, a significant amount of textual data is available that contains an abundance of individuals' opinions. Sentiment analysis is a task that supports companies and organizations to evaluate this textual data with the intention of understanding people's thoughts concerning services or products. Most previous research in Arabic sentiment analysis relies on word frequencies, lexicons, or black box methods to determine the sentiment of a sentence. It should be noted that these approaches do not take into account the semantic relations and dependencies between words. In this work, we propose a framework that incorporates Arabic dependency-based rules and deep learning models. Dependency-based rules are created by using linguistic patterns to map the meaning of words to concepts in the dependency structure of a sentence. By examining the dependent words in a sentence, the general sentiment is revealed. In the first stage of sentiment classification, the dependency grammar rules are used. If the rules are unsuccessful in classifying the sentiment, the algorithm then applies deep neural networks (DNNs). Three DNN models were employed, namely LSTM, BiLSTM, and CNN, and several Arabic benchmark datasets were used for sentiment analysis. The performance results of the proposed framework show a greater improvement in terms of accuracy and F1 score and they outperform the state-of-the-art approaches in Arabic sentiment analysis.

Keywords

Arabic Sentiment Analysis; Natural language processing; Deep Learning; Dependency-based rules.

1. Introduction

Sentiment Analysis (SA) is currently the focus of considerable interest from industry and academia. With the ubiquity of social media platforms on the Internet, many people can express their opinions or feelings about a product, brand or service via text. Consequently, a huge amount of unstructured data is available by way of the Internet. Understanding whether the sentiment found in a text is positive or negative can be a challenging task.

According to [1], Sentiment Analysis is defined as "the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes and emotions concerning entities such as products, services, organizations, individuals, issues, events, topics, along with their attributes". SA is classified into three distinct levels, specifically document, sentence and aspect levels [2,3].

There are three approaches applied in the literature to classify the sentiment of a given review: supervised, unsupervised or hybrid. Supervised approaches (also called corpus-based method), typically require a labelled dataset to build the classification model. Unsupervised approaches (also called lexicon-based method) rely on lexicons such as dictionaries, while hybrid approaches combine supervised and unsupervised approaches [2].

In lexicon-based methods, sentiment analysis counts the sentiment terms in the review to determine the general sentiment for a given review. However, this approach does not take into account word order or dependency relations between words, which have an important function in identifying the general sentiment of "الكتاب قديم لكن القصة ممتعة", " The book is old but the story is interesting". In this example, although the reviewer has expressed a negative sentiment in the first part of the review, the general sentiment of the review is positive. Therefore, the general sentiment of the review depends on the sentiment of the term and the relative terms, as well as on the dependency relations between these terms. Interestingly, the general sentiment of the review above is neutral when the review is analyzed using Mazajak [4], an Arabic sentiment analysis system based on deep learning models.

Furthermore, classifying sentiments based on dependency relations, i.e., linguistic rules, provides a logical explanation of why the review has that sentiment and maintains decision transparency with respect to that prediction. This is in contrast to black-box models, for instance deep learning models, which yield high accuracy without any justification or in which features are used in the prediction since the feature engineering is performed implicitly by the model.

To overcome these limitations, a hybrid framework for Arabic sentiment analysis is proposed that combines dependency-based grammatical rules with deep learning models. Arabic dependency-based rules are based on linguistic structures that allow mapping sentiment terms to concepts based on the dependency structure of a sentence. This proposed framework evolved from an insight obtained from the framework developed by [5] which is applied in the Persian language.

Several experiments are conducted on corresponding Arabic sentiment analysis datasets, and the hybrid Arabic sentiment analysis framework is compared to Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Networks (DNN) including Convolutional Neural Networks(CNN), Long-Short Term Memory (LSTM) and Bidirectional Long-Short Term Memory (BiLSTM). The comparative results demonstrate that the hybrid approach outperforms other methods.

Arabic dependency-based rules are not able to classify the entire dataset because term sentiment was not available in the lexicon or the rules were not triggered for some reviews. Therefore, a hybrid Arabic sentiment analysis is proposed.

The paper makes the following main contributions:

1. An innovative approach to Arabic sentiment analysis based on dependency rules. These rules are fully explainable and explore the terms and dependencies more comprehensively to provide a justification for each production. Thus, understanding the model predictions in an interpretable way can provide trust and transparency.

2. A comparative analysis of the proposed hybrid Arabic Sentiment Analysis framework with Logistic Regression, Support Vector Machine, Convolutional Neural Network, Long-Short Term Memory and Bidirectional LSTM.

3. An ablation study of the proposed Arabic dependency rule-based approach on different datasets illustrating the significance of each rule.

4. A solution for the limitation of unclassified reviews with the Arabic dependency rule-based approach by combining the rules with DNN models.

The remainder of the article is organized as follows: Section 2 reviews the literature on sentiment analysis. Section 3 provides a detailed methodology and Arabic dependency-based rules. Section 4 describes a hybrid framework that combines Arabic dependency-based rules with deep learning models. Section 5 describes the performance evaluation and results. Finally, Section 6 is the conclusion and provides recommendations for future research.

2. Review of Literature

This section surveys the literature on English sentiment analysis, Arabic sentiment analysis and dependency-based rules.

2.1. English Sentiment Analysis

Most research in sentiment analysis focuses on the English language because it is the most widely spoken language in the world. One of the earliest studies in the field of English sentiment analysis [6] analyzed a dataset of IMDB movie reviews for positive and negative sentiment analysis. In this study, the hand-engineered features were examined using machine learning techniques, namely Naive Bayes (NB), SVM and Maximum Entropy (ME). The evaluation results of the machine learning methods outperformed the human-generated baselines. In one of the initial works on sentiment classification in microblogging services such as Twitter, [7] used the same machine learning classifiers as [6] and added hand-engineered features that matched the unique features of Twitter. The paper also described the preprocessing steps required in the implementation of machine learning classifiers. The classifier that had the highest accuracy in this study was SVM with over 80%. [8] integrated linguistic, lexicon and micro-blogging features to identify the sentiment of Twitter sentences. Several experiments were conducted with several selected features.

In 2013, The International Workshop on Semantic Evaluation (SemEval) held a competition on sentiment analysis in Twitter (SemEval-2013) to address the lack of suitable datasets which could be used for comparison purposes [9]. SemEval-2013 was followed by further contests namely, SemEval-2014 [10], SemEval-2015 [11], SemEval-2016 [12] and SemEval-2017 [13].

Deep learning has been explored for sentiment analysis and has shown excellent performance [14]. The SwissCheese model, implemented by [15], performed best on SemEval-2016 by training two layers of convolutional neural networks and combining their results in a random forest classifier. The two layers had similar architectures but differed in hyperparameters and word embedding, Word2Vec and GloVe, respectively. [16] used two layers of BiLSTM architecture with an attention mechanism in the last layer to capture important terms. This model achieved the best performance results on SemEval-2017 [13]. The study by [17] developed a convolutional stacked BiLSTM with a multiplicative attention mechanism for the purpose of detecting aspect category and sentiment polarity. The evaluation of the model was carried out as a multiclass classification. Both the SemEval-2015 and the SemEval-2016 datasets were utilized in the evaluation of the model. In aspect-based sentiment analysis, the model performed better than the state-of-the-art results.

Ensemble models were also explored by [18] to enhance the evaluation metrics of sentiment analysis utilizing DNN models. In this study, the authors predicted the sentiment of IMDB movie reviews and the SST2 dataset by averaging the probability scores of CNN and BiLSTM. The BiLSTM captured the forward and backward context and the CNN extracted the local information. The evaluation results showed that the ensemble model performed better in terms of accuracy than the two models when used independently. Another ensemble model proposed by [19] combined both bidirectional LSTM and GRU with CNN. It also applied the attention mechanism to the outputs of the bidirectional layers to focus on important words in the text. Unlike the model of [18], the features extracted from the bidirectional layers were concatenated and then used in the CNN to capture the local structure. The results showed that the model performed well in classifying both short tweets and long reviews.

Other research has combined different DNN models to improve sentiment analysis. For example, [20] proposed a novel LSTM-CNN model based on hyperparameter optimization to predict sentiment analysis for two datasets, IMDB movie reviews and Amazon. In their work, the model was implemented using a grid-search approach and compared with CNN, KNN, LSTM, CNN– LSTM and LSTM–CNN. The results showed that the model outperformed the baselines.

A single layer of BiLSTM with a global pooling mechanism was employed by [21] on three datasets. The results were competitive with state-of-the-art models. The authors concluded that using one layer of BiLSTM was computationally efficient and beneficial for a real-time application such as sentiment analysis.

2.2. Arabic Sentiment Analysis

There are three main approaches used in Arabic sentiment analysis: lexicon-based, corpus-based, and hybrid methods [22] [23]. The lexicon-based method generally determines the polarity of sentences by measuring the sentiment words in the sentence. It usually uses predefined lexicons with annotated sentiment words [24].

In [25], a lexicon-based method for Arabic sentiment analysis was developed. The authors created 120,000 of their own terms. The system was tested on their collected tweets and achieved 87% accuracy.

The corpus-based method, also termed machine learning, is primarily based on a corpus and machine learning algorithms. In this method, the classifier is trained to predict the sentiment of the sentences [24]. Although several machine learning algorithms have been used to classify Arabic sentiment analysis, only three algorithms have shown good performance: SVM, K-Nearest Neighbor (KNN) and NB [26–28].

Deep learning non-contextual embeddings have also been explored in Arabic sentiment analysis and have shown excellent performance. The most important aspect of deep learning is that it works efficiently without feature engineering [24]. The study by [29] investigated four different architectures based on deep learning algorithms. The evaluation of the proposed models confirmed that the Recursive Auto Encoder (RAE), outperformed all other models. A study by [30] demonstrated a health services dataset which was collected from Twitter. Several deep learning and machine learning algorithms were investigated for sentiment analysis classification. The results showed that the deep learning approaches exhibited a promising performance and the results outperformed the SVM classifier.

Furthermore, [31] highlighted the integrating of CNN and LSTM deep learning models and tested the hybrid architecture on three benchmark datasets, namely Arabic Health Services Dataset (AHS) [30], ArTwitter [32] and Arabic Sentiment Tweets Dataset (ASTD) [27]. [33] evaluated the CNN and LSTM model along with different preprocessing techniques on a Moroccan dialect dataset manually collected from different social media sources. The performance results indicated that the deep learning approaches outperformed the classical approaches. Additionally, a Bidirectional LSTM network was investigated by [34] with the aim of enhancing Arabic sentiment analysis. The results of several benchmark datasets demonstrated the usefulness of this model on sequential data and in extracting contextual information in both forward and backward sequences.

Ensemble models were explored by [35] using deep learning models. In this framework, the authors predicted the sentiment of an ASTD dataset using soft voting, where the CNN and LSTM outputs were averaged to obtain the final results. The evaluation results revealed that the ensemble model achieved better scores in terms of accuracy and F1-score than the two models when they were used independently. Another ensemble model proposed by [36] integrated word embeddings with hand-engineered features. The architecture was evaluated on several datasets, specifically the SemEval 2017 [13], AraSenTi [37] and ASTD [27] for Arabic tweets, and outperformed previous results on all these datasets.

Much previous research into determining the sentiment of an Arabic sentence relied on word frequencies, lexicons or black-box methodologies. Clearly, semantic relations and dependencies between words are not taken into account in these approaches. However, some studies did use lexical rules in sentiment analysis [38] and [39]. The main objective of these studies was to analyze the effect of inverters such as negation to classify sentiment using machine learning approaches.

2.3. Dependency Grammar Based Rules Sentiment Analysis

To the best of our knowledge, the study by [5] was the first model combining dependency-based rules and DNN algorithms in the Persian language. It involved the dependency grammar based rules in the first stage of classifying the sentiment. Then, if the rules failed to classify the sentence, the algorithm used DNNs. In this study, two types of DNNs were used: CNN and LSTM. The results showed that their framework outperformed the state-of-the-art approaches in benchmark datasets for Persian product and hotel reviews.

3. Methodology

In this section, the innovated Arabic dependency-based rules for Arabic sentiment analysis are explained.

To explain the framework’s methodology, the following example will illustrate the concept. If the frequency of the positive and negative words is employed to determine the sentiment of the example review, “The book is very old but the story is not bad” “ الكتاب جدا قديم لكن القصة ليست سيئة”, the review will be categorized as negative due to the negative words “old” and “bad” “قديم” and “سيء”, even though the general sentiment of the review is positive given that the dependent tokens “but” and “not” appeared in the review. However, in the dependency-based rules approach, the dependent tokens are considered in the review and determine the general sentiment. Typically, the token “جدا” “very “ following “قديم” “old”, does not switch the general sentiment, but the negation particle “ليست” “not” and the negative token “سيئة” “bad” switches the general sentiment to positive. Furthermore, the existence of the token “لكن” “but” results in considering the sentiment of the second part after “لكن” “but.” Figure 1 explains the logical flow of the sentiment in order to determine the general sentiment for this review. For more details, see Section 3.1.

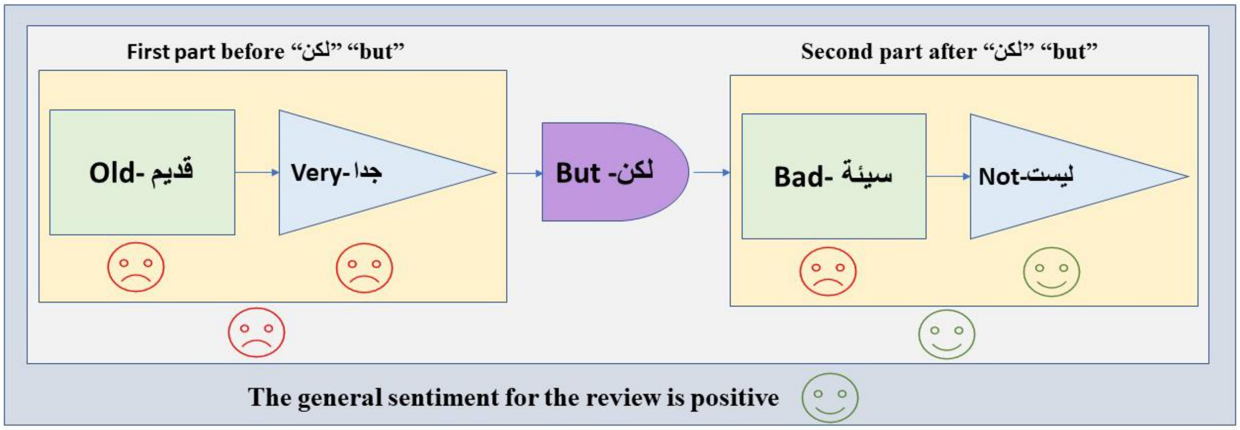
First part before “لكن” “but”
Second part after “لكن” “but”
Old- قديم
Very- جدا
But -لكن
Bad- سيئة
Not-ليست
The general sentiment for the review is positive

Fig 1. The logical flow of sentiment.

3.1. Arabic Dependency-Based Rules

In this section, we describe innovative Arabic dependency-based rules for sentiment analysis.

1.Sentiment Inversion

Trigger: when a review contains one of the Arabic negation particles.

There are many forms of negation in Arabic. In Standard Arabic, there are five known elements for negation, as shown in Table 1.

Negation in a review can be considered as a switch operator. For instance, when a negation particle is employed with a positive verb, the review’s general sentiment is negative, and when a negation particle is employed with a negative verb, the review’s general sentiment is positive. For instance, in “The customer does not like the service” “ لا تعجب العميل الخدمة ”, the general review’s sentiment is negative.

Since the dialectal words (مش and مو) have the same meaning as " ليس/ليست " "laysa", they have been added in this rule.

Negation Particle	Pronunciation	Usage	Example in Arabic	Translation
لا	lA	For present form	لا تعجب الخدمة العميل	The customer does not like the service.
لم	lam	For past form	لم تعجب الخدمة العميل	The customer did not like the service.
ما	mA	For past form	ما اعجبت الخدمة العميل	The customer did not like the service.
لن	lan	For future form	لن تعجب الخدمة العميل	The customer will not like the service
ليس ليست	Laysa laysat	For noun-verb form	ليست الخدمة جيدة	The service is not good

Table 1. Sentiment inversion particles.

2. Adversative Clause

Trigger: the adversative words such as "but" "لكن " , "although" "بالرغم من "or "however" "بالرغم ان" are used to join two opposite sentences.

The reviews are divided into two segments depending on the appearance of a word such as "but", and the sentiment of the second segment is taken into account. For instance, in the review, “The car is really good but it‘s very expensive” “ السيارة جيده لكنها جدا غالية”, the first segment in advance of the word “but” is positive, whereas the second segment is negative. Therefore, the review’s general sentiment is negative. Likewise, in “The car is really costly but it‘s very luxurious” “السيارة جدا مكلفة ولكنها جدا فخمة”, the first segment of the review is negative, whilst the second segment is positive. Hence, the review’s general sentiment is positive.

3. Adverbial Clause

Trigger: if a review includes an adverbial clause like “whereas” “بينما”.

The action of “whereas” “بينما” in a review is like the word “but”. If a review includes “whereas”, the review is divided into two segments. The sentiment of the second segment is regarded as the general review’s sentiment. For example, “In the user guide, they said the camera has a zoom lens, whereas the lens is without the zoom” "في دليل المستخدم قالوا إن الكاميرا بها عدسة تكبير بينما العدسة بدون تكبير". The sentiment of the first segment is positive and the sentiment of the second segment is negative. For this reason, the review’s general sentiment is negative.

4. Exclamation Clause

Trigger: when a review starts with “ما” “mA” to express exclamation such as surprise or strong emotion in the Arabic language.

In this case, “ما” is not regarded as a negation particle. For example, if the “ما” particle is followed by a positive word, the general review’s sentiment is positive and if the “ما” particle is followed by a negative word, its general sentiment is negative. For example, with respect to “what a beautiful book “ ما أجمل الكتاب ”, the general review’s sentiment is positive because of the word “beautiful” “ أجمل”.

5. Superlative Clause

Trigger: when a review has a noun on the pattern “أفعل” to represent the best or worst action.

A superlative clause is used to represent concepts, such as best design, worst doctor, etc. For example, “This book is one of the best I have read” “هذا الكتاب من اجمل ما قرات”. The general review’s sentiment depends on the sentiment of the noun on the pattern “أفعل”. In this instance, the sentiment of “best” “ اجمل” is positive. Hence, the general review’s sentiment is positive.

6. Joint noun and adjective

Trigger: when a review includes joint noun and adjective.

If there is a connection between the noun and the adjective, both words are considered in the review. The sentiment of the adjective is taken into account. For example, "The bag is new" or "The bag is old" " الحقيبة جديدة" or " الحقيبة قديمة". There is a subject relationship between "bag" and "new or old".

7. Preposition rule

Trigger: when a review has the preposition “against” “ضد”.

Although the preposition “against” “ضد” is normally employed in negative reviews, it can also be employed in positive reviews. Generally, if an action comprises a positive sentiment “A group of scientists presented a report on their discovery of drugs” “قدم مجموعة من العلماء تقرير عن اكتشافهم ادوية “ and is followed by a negative preposition modifier “ against infectious diseases” ” ضد امراض معدية”, the general review’s sentiment is changed to positive. Conversely, if an action comprises a negative statement “Israel launched a war” “ شنت اسرائيل حرب ” and is followed by a negative preposition modifier “against Palestine” “ضد فلسطين”, the general review’s sentiment is changed to negative. Table 2 summarizes the preposition “ضد” rule.

Sentence before “ضد” sentiment	Sentence after “ضد” sentiment	General sentiment	Example in Arabic	Translation
Positive	Negative	Positive	قدم مجموعة من العلماء تقرير عن اكتشافهم ادوية ضد امراض معدية	A group of scientists presented a report on their discovery of drugs against infectious diseases.
Negative	Positive	Negative	مؤامرة ضد الشرفاء	A conspiracy against honest people.
Negative	Neutral	Negative	شنت اسرائيل حرب ضد فلسطين	Israel launched a war against Palestine.
Neutral	Negative	Negative	الحكومات ضد المفاعل النووي	Governments against nuclear reactors.

Table 2. Preposition " against " rule summary.

8. Adverbial/Adjective Sub-rule

Trigger: The noun “غير” “other than” can be followed by adverbs or adjectives to express contradiction or negation. It can represent “not, non-, un-, dis-,in-” in the English language.

If positive adverbs or adjectives are followed by “غير”, the sentiment of the concept can be changed to negative and vice versa. For example, in “I am not happy today” “ انا غير سعيد اليوم”, the sentiment of the review is negative because the noun “غير” appears in the review.

9. Other Rules

From exploring the reviews on social media, we found three types of rules: Supplications, Aggressive Words and Apologetic Feelings.

9.1 Supplications

Trigger: when the review has one of the words generally used during prayer in Arabic supplications, such as “اللهم” “Oh God”, “يا الهي” “My God”, “يا رب” “Oh Lord” and “امين” “Amen”. In this case, the general review’s sentiment is positive.

9.2 Aggressive Words

Trigger: when the review has an aggressive word typically employed in Arabic reviews, for instance “يلعن” “Damned” and “يلعنكم” “Damn you”. In this case, the general review’s sentiment is negative.

9.3 Apologetic Feelings

Trigger: when the review has apologetic words, such as “للأسف” “Unfortunately”, “يؤسفني” “I'm sorry" ”and “اسف” “sorry”. If these words exist at the beginning or near the end of the review, the general review’s sentiment is negative.

Table 3 provides a summary of the Arabic dependency-based rules.

Rule	Behaviour
Sentiment Inversion	When negation is used with a positive word, the general review’s sentiment is negative and when negation is used with a negative word, the general review’s sentiment is positive.
Adversative Clause	A word such as “but” “لكن ”“although” “بالرغم من ”or “however” “ بالرغم ان” is used in the review. The reviews are divided into two segments and the sentiment of the second part is considered.
Adverbial Clause	The review is divided into two segments by “بينما” and the sentiment of the second part is employed as the general review’s sentiment.
Exclamation Clause	When the “ما” particle is followed by a positive word, the general review’s sentiment is positive and vice versa.
Superlative Clause	The general review’s sentiment depends on the sentiment of the noun on the pattern “أفعل”. If the pattern is positive, the general review’s sentiment is positive and vice versa.
Joint noun and adjective	If there is a relationship between the noun and adjective, the sentiment of the adjective will be considered.
Preposition	When the word “against” “ضد” appears in the review. The general review’s sentiment is presented in Table 2.
Adverbial/Adjective Sub-rule	If positive adverbs or adjectives are followed by “غير”, the sentiment of the concept can be changed to negative and vice versa.
Other rules	The sentiment of a review depends on the appearance of particular terms.

Table 3. Summary of the Arabic dependency-based rules.

4. Hybrid Framework

In this section, we discuss the details of our hybrid framework that combines Arabic dependency-based rules and Deep Learning models.

4.1. Framework Overview

The framework integrates Arabic dependency-based rules and a deep learning based model. Unclassified reviews from Arabic dependency-based rules will feed into the deep learning based model. **Algorithm 1** demonstrates the proposed hybrid approach.

The first step regarding the framework is preprocessing each review then tokenizing and recognizing the Part-of-Speech tagging (PoS tags) for each token based on each rule as a bag of concepts. The tokens’ sentiments of the dependency concepts will feed into the triggered rule. If the rule classifies the review, the framework will return the review sentiment. For the reviews that were not classified by the dependency rules because the token sentiment was not available in the lexicon or no rules were triggered, the reviews are input to the selected DNN algorithm to obtain the reviews' sentiments. Figure 2 illustrates the proposed framework.

```
Results: The review sentiment
for each review acquire the tokens and PoS tags do
for each token in review do
      if token in lexicon then
          Assign sentiment for each obtained token
      else
          Assign zero sentiment
      end
Utilize dependency-based rules
if sentiment assigned by dependency-based rules then
   return sentiment
else
    Utilize DNN algorithm
    return  sentiment
end
```

Algorithm 1 hybrid approach.

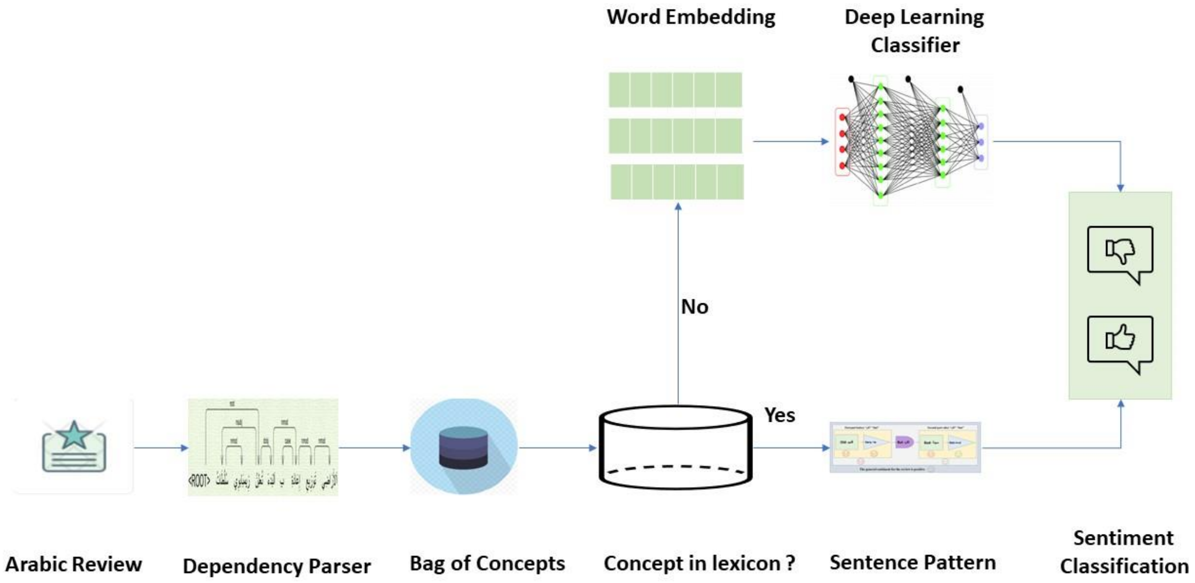

Fig 2.The proposed framework.

4.2. Deep Neural Networks (DNNs)

This section presents the DNN models used in the proposed hybrid framework and the same models used independently for a comparative study.

4.2.1 Long Short-Term Memory (LSTM)

LSTM is the most popular Recurrent Neural Network (RNN) model that has been successfully used in sentiment analysis with recognizable results. This model can handle long-term dependencies due to its internal memory. Therefore, it is often used with sequential data.

The LSTM consists of an input gate, a memory, an output gate and a forget gate. The goal of the memory is to remember previous data. Both the current and previous input are considered when making a prediction [40].

Our LSTM implantation is similar to the architecture used in [41] for sentiment analysis. The LSTM network takes the sentence as a sequence of words and returns the sentiment value as positive or negative. LSTM can be represented mathematically as follows:

$$h_t = f(W_h \; . \; x_t \; + \; U_h \; . \; h_{t-1} + b_h), \qquad (1)$$

where x_t is the word embedding, W_h and U_h are weight matrices, b_h is a bias, $f(x)$ is a nonlinear function normally selected as tanh and, h_t is the hidden state.

$$f_t = \sigma(W_f \, . \, x_t + U_f \, . \, h_{t-1} + b_f) \qquad (2)$$

$$i_t = \sigma(W_i \, . \, x_t + U_i \, . \, h_{t-1} + b_i) \qquad (3)$$

$$o_t = \sigma(W_o \, . \, x_t + U_o \, . \, h_{t-1} + b_o) \qquad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_c \, . \, x_t + U_c \, . \, h_{t-1} + b_c) \qquad (5)$$

$$h_t = o_t \circ \tanh(c_t), \qquad (6)$$

where i_t is called the input gate, f_t is the forget gate, c_t is the memory cell, σ is the sigmoid function, and $\circ$ is the Hadamard product [34].

4.2.2 Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory [42] is another type of RNN model. It consists of two stacked LSTMs; the first processes the sequence forward and the second backward. The output is computed based on the final hidden state h_t^{bilstm} of both LSTMs. BiLSTM can be represented as follows:

$$h_t^{bilstm} = h_t^{forward} \oplus h_t^{backward}, \qquad (7)$$

where $\oplus$ is concatenation operator [34].

4.2.3 Convolutional Neural Networks (CNN)

The CNN model is primarily used in image and video classifications as well as in sequential data like text processing [43]. It consists of convolutional and pooling layers, followed by a series of fully connected layers. The convolutional layer has several kernels in order to apply the convolution operation on input. It then sends the result to the subsequent layer. In Natural Language Processing (NLP) tasks, the text is usually represented as word embeddings rather than pixels of the image [44]. The convolution operation can be defined as follows:

$$c_i = f(\textstyle\sum_{j,k} w_{j,k} \, (X_{[i:i+h-1]_{j,k}}) + b), \qquad (8)$$

where X is a matrix dimension, w is a filtering matrix, h is the size of the convolution, b is a bias term , $f(x)$ is a non-linear function usually chosen to be the ReLU function and the output c is a concatenation of the convolution operator [45].

5. Performance Evaluation

This section discusses the datasets, lexicons, data preprocessing, evaluation metrics, environment setup and parameters, experimental setup, experimental results, and ablation study.

5.1. Datasets

To evaluate the performance of the proposed framework, several Arabic benchmark datasets are used.

Arabic Sentiment Tweets Dataset (ASTD) [27]**:** ASTD consists of approximately 10000 Egyptian tweets annotated as positive, negative, neutral, and objective. The tweets were collected in September 2013. As our interest is in positive and negative classes, the objective and neutral classes were removed creating a resultant set of 2482 unbalanced tweets, which we refer to as **ASTD-U**. In addition, the balanced shape of the dataset **ASTD-B** which was sampled by [46], is used in our experiments. **ASTD-B** consists of 777 tweets belonging to positive class and 812 tweets belonging to negative class.

Arabic Jordanian General Tweets (AJGT) dataset [47]**:** The AJGT includes two balanced classes, 900 positive and 900 negative tweets. Hence, there are 1800 in total in Modern Standard Arabic (MSA) and the Jordanian dialect. The tweets were collected in May 2016 by way of Twitter API. The dataset was manually annotated by two experts along with one extra expert for consultation.

ArTwitter dataset [32]: It was collected using a tweet crawler in several topics such as arts and politics written in MSA and Jordanian dialect. In our experiments, we used the same dataset evaluated by [46] , which consists of about 2000 balanced classes for positive and negative tweets.

5.2. Arabic Sentiment Lexicons

Several Arabic sentiment lexicons are used collectively in our experiments. The first one is NileULex [48], which is Egyptian dialect and MSA terms. This lexicon has 5953 distinctive sentiment terms for both negative and positive sentiments, where 45% of the terms are Egyptian and 55% are modern standard Arabic. The second lexicon is Ar-SenticNet created by [49] which comprises a total of 48k terms. Some terms in Ar-SenticNet are built by translating the English SenticNet_v4 into Arabic using wordnet mapping, whereas other terms are translated using the Google translate method. The third lexicon belongs to the Large-scale Arabic Book Review (LABR) dataset [50]. LABR is a huge sentiment analysis Arabic book review dataset which was collected from the Goodreads website in the Egyptian dialect.

5.3. Data Preprocessing

Typically, data preprocessing is the first step applied to the raw data to prepare the text data for sentiment analysis. Table 4 provides an example of the preprocessing steps.

We applied the following steps in the dependency grammar-based rules part:

- **Cleaning:** this step carries out general cleaning to remove unwanted text parts, including English letters and numbers, URLs, mentions, retweets, hashtags, punctuation marks, extra spaces and diacritics “Tashkeel”.
- **Letter normalization:** this step unifies the letters in Arabic which are normally written in several forms. For example , ("أإآ" replaced with 'ا'), ("ة" replaced with 'ه') and ("ي" replaced with 'ى').
- **Elongation removal:** as social media reviews have a certain writing style like repeating various characters which are used to express strong negative or positive emotions, this step returns the actual form of the word.

For the DNN part, we added the following steps along with the previous preprocessing steps:

- **Stop words removal:** in this step, most words that carry no or very little meaningful information are removed.
- **Stemming:** The Information Science Research Institute’s (ISRI) stemmer is employed in the stemming step.

Original review	المطعم الذي بجانب البحيرة قديم جدااااااااااا ، أغلب الطلبات التي في قائمة الطعام رقم 2 غير موجودة .
Cleaning	المطعم الذي بجانب البحيرة قديم جدااااااااااا أغلب الطلبات التي في قائمة الطعام رقم غير موجودة
Letter normalisation	المطعم الذى بجانب البحيره قديم جدااااااااااا اغلب الطلبات التى في قائمه الطعام رقم غير موجوده
Elongation removal	المطعم الذى بجانب البحيره قديم جدا اغلب الطلبات التى في قائمه الطعام رقم غير موجوده
Stop words removal	المطعم بجانب البحيره قديم جدا اغلب الطلبات قائمه الطعام رقم غير موجوده
Stemming	طعم جنب بحر قدم جدا غلب طلب قئم طعم رقم وجد

Table 4. Data preprocessing example.

5.4. Evaluation Metrics

Evaluation metrics commonly used in relation to sentiment analysis are accuracy and F1 score.

- **Accuracy:** Accuracy describes how frequently the sentiment rating predicted by the model is correct. Accuracy is calculated as:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, \quad (9)$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively.

- **F1 score:** Both precision and recall of test data are used to calculate the F1 score. It is calculated as follows.

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$\boldsymbol{F1} = \frac{\mathbf{2}(\boldsymbol{Precision} \times \boldsymbol{Recall})}{\boldsymbol{Precision} + \boldsymbol{Recall}} \quad (12)$$

5.5. Environment Setup and Parameters

As the Python language has a number of APIs, we chose this language to perform the performance evaluation through Google Colaboratory Environment. For reading and writing files, the Pandas API was chosen. For the deep learning algorithm, we used Keras on the TensorFlow back_end deep learning platform. For tokenization and PoS tagging, CAMeL Tools, an open source Python toolkit developed by [51] was used.

Several hyperparameters and settings were explored to determine the optimal parameters. For all DNNs, the number of epochs for all experiments is 5. For neural network regularization and to avoid the problem of overfitting, we use a dropout rate of 0.2 for all experiments. For the CNN, the filter size for all experiments is 3. This is followed by a fully connected layer with ReLU activation function. Finally, a softmax layer with two output units is used to predict the positive or negative sentiment of the tweet. In all experiments, the Adam Optimizer [52] with binary cross entropy was used. The selected hyperparameters for the DNN experiments are listed in Table 5.

DNN Models	LSTM cell	Recurrent Dropout	Output Dropout	#Filters	Filter Size	Hidden Units
LSTM	64	0.2	0.2	-	-	-
BiLSTM	32	0.2	0.2	-	-	-
CNN	-	-	-	250	3	250

Table 5. Hyperparameters employed for DNN models.

5.6. Experimental Setup

For the ASTD-U, ASTD-B, AJGT and ArTwitter datasets, an 80% training set and 20% test set splitting was applied. The dependency-based rules were not necessary for the training step. However, they were used to evaluate the test set. The experiment begins to train the selected deep learning models, i.e. LSTM, BiLSTM and CNN using a training set. With respect to word embedding, two specific methods are used: word embedding and pre-trained word embedding. First, in word embedding, the training set is vectorized into a list of integers (vectors). Each vector maps to a specific value in a dictionary by using Keras's text pre-processing library. The second embedding, i.e., pre-trained word embedding, utilizes fastText, which is available in several languages including Arabic [53]. Next, the dependency rules are evaluated using a test set. Unclassified reviews from the dependency rules are switched to the selected DNN models.

The framework performance is compared with the selected deep learning models which are investigated individually using the same splitting and hyperparameters setup for all datasets. Additionally, the hybrid framework is compared with two machine learning baselines, namely logistic regression and support vector machine. The baselines experiments are run using Term Frequency (TF) and the Term Frequency * Inverse Document Frequency (TF*IDF) weighting scheme by means of unigram. Furthermore, the performance of the proposed framework is compared with the state-of-the-art Arabic Sentiment Analysis models.

5.7. Experimental Results

The performance results of the proposed framework, deep learning models and the baseline of the machine learning classifiers for the ASTD-U, ASTD-B, AJGT and ArTwitter test sets, are presented in Tables 6 and 7.

In terms of **evaluation metrics**, the experimental results show that the proposed hybrid framework which combines the Arabic dependency-based rules with deep learning models outperformed deep learning models when they are evaluated individually. It also demonstrated improvement over baseline machine learning methods for all datasets.

In the ASTD-U dataset, the hybrid framework achieves a significant improvement in terms of accuracy and F1 score (86.36 % and 0.87, respectively) over deep learning models when they and the baseline methods are evaluated individually. The best performance framework for ASTD-U is Hybrid 3 which combines the CNN and dependency-based rules using word embedding and fastText word embedding. For ASTD-B and ArTwitter datasets, the best performance framework is Hybrid 2, which combines the BiLSTM and dependency-based rules using word embedding. The accuracy and F1 score for ASTD-B are 89.44% and 0.89, respectively, and for ArTwitter are 93.85% and 0.95, respectively.

It is also significant that the performance of Hybrid 1, which combines LSTM with the dependency-based rules using word embedding, is the best among the other hybrid frameworks on the AJGT dataset in terms of accuracy and F1 score (89.72% and 0.90, respectively). The comparison of the hybrid frameworks in terms of accuracy on the ASTD-U, ASTD-B, AJGT, and ArTwitter test sets is shown in Figure 3, which indicates that ArTwitter achieves the best accuracy results.

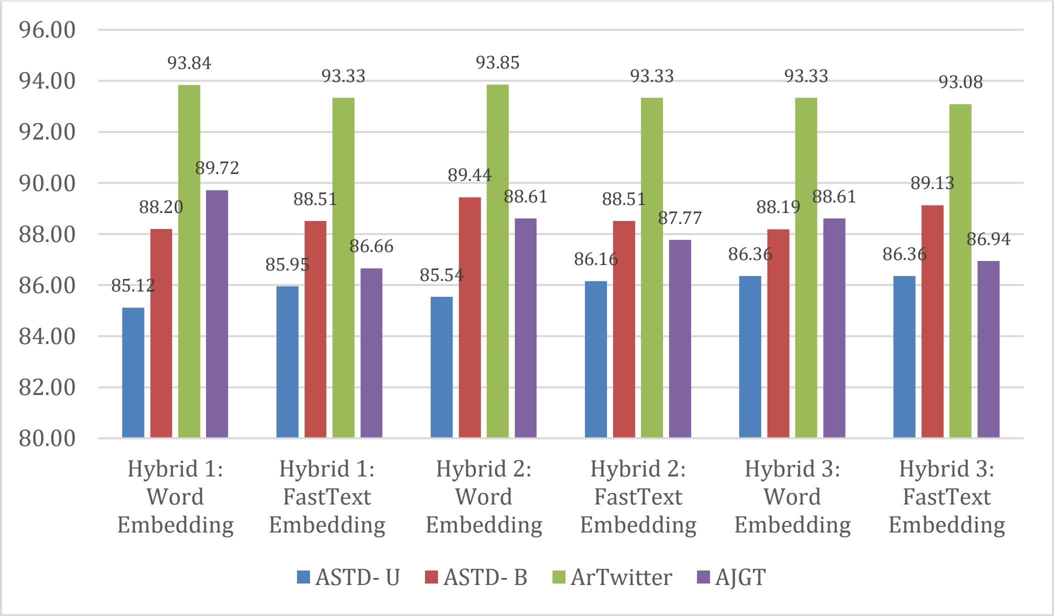

Fig 3. The accuracy of the hybrid models of ASTD-U, ASTD-B, AJGT and ArTwitter.

To compare the DNN models used in the hybrid framework, we used the Wilcoxon signed-ranks test [54]. This is a nonparametric test that compares two classifiers with a significance level of 0.05. Thus, we compared Hybrid 1 versus Hybrid 2, Hybrid 1 versus Hybrid 3, and Hybrid 2 versus Hybrid 3 . Table 8 shows the p-values of the hybrid classifiers that used fastText word embedding for all datasets. As can be seen in Table 8, there is no significant difference between the DNNs in the hybrid framework in terms of the evaluation metrics as the 0.05 < p-values.

In terms of **explainability**, while the performance results pertaining to the deep learning models speak for themselves, a well-known shortcoming of these models relates to why a model achieves a particular prediction. However, the dependency-based rules part of the proposed framework provides a full explanation of a specific prediction, and this makes our framework semi explainable.

Table 9 represents the performance results of the state-of-art in Arabic sentiment analysis approaches. It is obvious that our framework has improved the performance of Arabic sentiment analysis as it achieved 86.36% accuracy in ASTD-U dataset, 89.44% in ASTD-B dataset, 93.85% in ArTwitter dataset and 89.72% in AJGT dataset. This outperforms the state-of-the-art approaches.

The error analysis of our proposed Arabic dependency-based rules is as follows:

- A long review can trigger many rules. Consequently, an incorrect prediction will occur. However, our framework works perfectly in a short review.
- Some reviews have inappropriate labelling and require reannotation in order to achieve the correct sentiments.
- Parsing tools have a significant impact on implementing dependency-based rules. Thus, a reliable tool for MSA and dialects is essential to improve the proposed framework.
- Dependency-based rules depend on the availability of the term sentiment; however the term sentiment of the Arabic lexicon cannot always be found to fit these rules.
- Spelling mistakes in online reviews are also a challenge because they can affect the performance of the dependency-based rules.

	Approach	ASTD- U				ASTD- B			
		Recall	Precision	F1	Accuracy (%)	Recall	Precision	F1	Accuracy (%)
	Dependency-Based Rules	0.81	0.50	0.61	66.79	0.80	0.67	0.73	71.68
Baselines	**LR (TF)**	0.38	0.96	0.54	66.94	0.84	0.72	0.78	76.71
	LR (TF*IDF)	0.58	0.87	0.69	73.76	0.81	0.79	0.80	80.43
	SVM (TF)	0.16	1.00	0.27	56.40	0.82	0.62	0.70	66.15
	SVM (TF*IDF)	1.00	0.52	0.68	51.65	0.80	0.80	0.80	80.43
Proposed Framework	**LSTM Word Embedding**	0.52	0.78	0.63	67.77	0.71	0.81	0.76	77.95
	Hybrid 1: LSTM + Dependency-Based Rules Word Embedding	0.88	0.84	0.86	85.12	0.89	0.87	0.88	88.20
	LSTM fastText Embedding	0.48	0.81	0.61	67.36	0.82	0.87	0.84	84.67
	Hybrid 1: LSTM + Dependency-Based Rules fastText Embedding	0.88	0.86	0.87	85.95	0.90	0.87	0.88	88.51
	BiLSTM Word Embedding	0.48	0.80	0.60	67.15	0.76	0.82	0.79	80.12
	Hybrid 2: BiLSTM + Dependency-Based Rules Word Embedding	0.87	0.85	0.86	85.54	0.90	0.88	**0.89**	**89.44**
	BiLSTM fastText Embedding	0.55	0.80	0.65	69.63	0.77	0.79	0.78	78.57
	Hybrid 2: BiLSTM +Dependency-Based Rules fastText Embedding	0.89	0.85	0.87	86.16	0.91	0.86	0.89	88.51
	CNN Word Embedding	0.50	0.79	0.62	67.56	0.69	0.77	0.73	74.84
	Hybrid 3: CNN + Dependency-Based Rules Word Embedding	0.88	0.86	**0.87**	**86.36**	0.89	0.87	0.88	88.19
	CNN fastText Embedding	0.37	0.85	0.52	64.26	0.81	0.79	0.80	80.43
	Hybrid 3: CNN + Dependency-Based Rules fastText Embedding	0.86	0.87	**0.87**	**86.36**	0.91	0.87	0.89	89.13

Table 6. Summary of the evaluation results of the ASTD-U and ASTD-B test sets.

	Approach	ArTwitter				AJGT			
		Recall	Precision	F1	Accuracy (%)	Recall	Precision	F1	Accuracy (%)
	Dependency-Based Rules	0.92	0.84	0.88	83.51	0.91	0.84	0.87	82.84
Baselines	LR (TF)	0.84	0.91	0.87	86.41	0.80	0.82	0.81	81.67
	LR (TF*IDF)	0.88	0.90	0.89	87.69	0.85	0.81	0.83	83.06
	SVM (TF)	0.76	0.91	0.83	82.31	0.74	0.83	0.78	79.72
	SVM (TF*IDF)	0.89	0.90	0.90	88.72	0.85	0.82	0.83	83.33
Proposed Framework	LSTM Word Embedding	0.91	0.92	0.91	90.51	0.89	0.88	0.89	88.61
	Hybrid 1: LSTM + Dependency-Based Rules Word Embedding	0.94	0.95	0.95	93.84	0.92	0.88	**0.90**	**89.72**
	LSTM fastText Embedding	0.93	0.91	0.92	90.51	0.89	0.81	0.85	84.16
	Hybrid 1: LSTM + Dependency-Based Rules fastText Embedding	0.95	0.92	0.94	93.33	0.92	0.83	0.87	86.67
	BiLSTM Word Embedding	0.92	0.91	0.92	90.51	0.82	0.89	0.86	86.38
	Hybrid 2: BiLSTM + Dependency-Based Rules Word Embedding	0.95	0.95	**0.95**	**93.85**	0.88	0.89	0.88	88.61
	BiLSTM fastText Embedding	0.95	0.89	0.92	90.51	0.85	0.85	0.85	85.55
	Hybrid 2: BiLSTM +Dependency-Based Rules fastText Embedding	0.95	0.93	0.94	93.33	0.90	0.86	0.88	87.77
	CNN Word Embedding	0.88	0.94	0.91	90.25	0.86	0.85	0.86	86.11
	Hybrid 3: CNN + Dependency-Based Rules Word Embedding	0.94	0.93	0.94	93.33	0.91	0.87	0.89	88.61
	CNN fastText Embedding	0.95	0.87	0.91	88.97	0.90	0.80	0.85	84.17
	Hybrid 3: CNN + Dependency-Based Rules fastText Embedding	0.95	0.92	0.94	93.08	0.93	0.83	0.87	86.94

Table 7. Summary of the evaluation results of the ArTwitter and AJGT test sets.

Metrics	Hybrid 1 vs. Hybrid 2	Hybrid 1 vs. Hybrid 3	Hybrid 2 vs. Hybrid 3
Recall	0.9	0.9	0.9
Precision	0.71	0.32	0.32
F1	0.16	0.32	0.32

Table 8. The P-values of the Wilcoxon signed-ranks test of the hybrid classifiers.

Dataset	Model	Technique	Accuracy (%)	F1 Score (%)
ASTD-U	[31]	Combined CNN and LSTM	77.62	-
	Our best hybrid	Hybrid 3	86.36	86.69
ASTD-B	[34]	BiLSTM	79.25	76.83
	[55]	CNN	82.48	82.57
	Our best hybrid	Hybrid 2	89.44	89.31
ArTwitter	[34]	BiLSTM	91.82	92.39
	[31]	Combined CNN and LSTM	88.10	-
	[56]	Combined LSTMs	87.27	87.28
	Our best hybrid	Hybrid 2	93.85	94.52
AJGT	[57]	KNN, LR	82	-
	Our best hybrid	Hybrid 1	89.72	89.81

Table 9. The performance results of the state-of-art methods for all datasets.

5.8. Ablation Study

The ablation study is performed separately for the Arabic dependency-based rules, using all the lexicons described in Section 5.2 together. The results for the ASTD-U, ASTD-B, AJGT, and ArTwitter datasets are shown in Table 10. The table clearly shows that the accuracy of all rules in the ArTwitter dataset exceeds that of the other datasets. This is possibly because the sentiment of the terms in the selected lexicons contains more related terms or because the logic behind the ArTwitter annotation is closer to our dependency-based rules.

Exclamation clauses, superlative clauses and preposition rules yielded high accuracy in all datasets. Furthermore, the adversative clause rule achieved the lowest performance in the ArTwitter and AJGT datasets, while the sentiment inversion rule and adverbial/adjective sub rule achieved the lowest performance in the ASTD-U and ASTD-B datasets. There is no example of an adverbial clause rule in the ArTwitter and AJGT datasets. Therefore, no evaluation result is reported.

Table 11 illustrates some examples of classified reviews using the Arabic dependency-based rules.

Rule Type	ASTD-U Accuracy (%)	ASTD-B Accuracy (%)	ArTwitter Accuracy (%)	AJGT Accuracy (%)
Sentiment Inversion	51.92	46.00	72.5	65.45
Adversative Clause	68.75	70.00	33.33	45.45
Adverbial Clause	75.00	100	-	-
Exclamation Clause	100	100	80.00	83.33
Superlative Clause	100	100	100	100
Joint Noun Adjective	67.75	74.10	76.18	75.36
Preposition	96.88	93.75	100	100
Adverbial/Adjective Sub-rule	55.56	27.27	100	85.71
Other Rules	66.67	85.11	94.81	96.67
All Rules	66.79	71.68	83.51	82.84

Table 10. Ablation study of all datasets.

Arabic Review	Translation	Sentiment	Rule Type
اكيد الله لا يحرمنا من هالنعمه العظيمة.	Certainly, May God not deprive us of this great blessing.	Positive	Sentiment Inversion
اخبار مسخره	Ridiculous News.	Negative	Joint Noun Adjective
مش فاهم عليك لانك صاير غير مفهوم.	I do not understand you because you are becoming incomprehensible.	Negative	Adverbial/Adjective Sub-rule
رفع الدعم عن الكهرباء والغاز والسولار والبنزين اجرام سياسي ضد الفقراء.	Lifting subsidies on electricity, gas, solar and petrol is a political crime against the poor.	Negative	Preposition

Table 11. Examples of classified reviews using dependency-based rules.

6. Conclusion

With the growth of social platforms and the development of communication media through them, there is a considerable amount of textual data rich in opinions and attitudes. Sentiment analysis is a task that helps organizations and businesses to analyze this textual data to understand consumers' thoughts regarding services or products. Although many studies on Arabic sentiment analysis are described in the literature, few of them pay attention to word order or dependency relationships between words, although these have an important function in detecting the general sentiment in reviews. The hybrid framework proposed in this paper combines Arabic dependency-based rules with deep learning models. The results of our proposed framework show a visible improvement in accuracy and F1 score and also outperform the state-of-the-art in Arabic sentiment analysis.

Future work will undoubtedly help to improve this model. The deep learning part of the hybrid framework requires the development of a new model in conjunction with explainability to provide trust and transparency in each prediction so that the framework can achieve full explainability [58,59]. There is ample room for improvement when we use the hierarchical hybrid ensemble–based AI model and compare it to our framework. This model employs two lexicon-based methods integrated with a pre-trained deep learning-based model (BERT) [60]. Further work needs to be done to extend our Arabic unimodal framework, i.e., text modality, to include Arabic multimodal sentiment analysis, which includes multiple modalities (text, audio and visual) [61]. Another potential area of future research would be to improve Arabic aspect-based sentiment analysis through transfer learning [62]. Moreover, addressing the challenges posed by the limited number of the Arabic language lexicon would have a significant impact on improving the proposed framework. A useful approach to address this issue could be to extend SenticNet to include the Arabic language [63,64]. This version of SenticNet 7 [65] generalizes semantically related concepts of terms and expressions with multiple terms into a set of primitives that are later identified as superprimitives. The advantage of this method is that we only need the polarity of the superprimitives instead of building a huge lexicon with polarity. Finally, these dependency-based rules should be modified and applied across several Arabic language dialects.

7. Acknowledgments:

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions. Hussain would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

8. REFERENCES

[1] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining Text Data, 2012: pp. 415–463. https://doi.org/10.1007/978-1-4614-3223-4_13.

[2] I. Guellil, F. Azouaou, M. Mendoza, Arabic sentiment analysis: studies, resources, and tools, Social Network Analysis and Mining. 9 (2019) 1–17. https://doi.org/10.1007/s13278-019-0602-x.

[3] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: Approaches, challenges and trends, Knowledge-Based Systems. 226 (2021) 1–26. https://doi.org/10.1016/j.knosys.2021.107134.

[4] I. Abu Farha, W. Magdy, Mazajak: An Online Arabic Sentiment Analyser, in: Proceedings of the Fourth Arabic Natural Language Processing Workshop, 2019: pp. 192–198. https://doi.org/10.18653/v1/w19-4621.

[5] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, A. Hussain, A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks, Neurocomputing. 380 (2020) 1–10. https://doi.org/10.1016/j.neucom.2019.10.009.

[6] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002: pp. 79–86. https://doi.org/10.3115/1118693.1118704.

[7] A. Go, R. Bhayani, L. Huang, Twitter Sentiment Classification using Distant Supervision, CS224N Project Report, Stanford. (2009) 1–12.

[8] E. Kouloumpis, T. Wilson, J. Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011: pp. 538–541. https://doi.org/10.1016/0378-1097(92)90668-E.

[9] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, T. Wilson, SemEval-2013 task 2: Sentiment analysis in Twitter, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), 2013: pp. 312–320.

[10] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 Task 4: Aspect Based Sentiment Analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: pp. 27–35. https://doi.org/10.3115/v1/s14-2004.

[11] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, V. Stoyanov, SemEval-2015 Task 10: Sentiment Analysis in Twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation, 2015: pp. 451–463. https://doi.org/10.18653/v1/s15-2078.

[12] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, V. Stoyanov, SemEval-2016 Task 4: Sentiment Analysis in Twitter Preslav, in: Proceedings of SemEval-2016, 2016: pp. 1–18. https://doi.org/10.18653/v1/s16-1032.

[13] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 Task 4: Sentiment Analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), 2017: pp. 502–518. https://doi.org/10.18653/v1/s16-1001.

[14] M.M. Agüero-Torales, J.I. Abreu Salas, A.G. López-Herrera, Deep learning and multilingual sentiment analysis on social media data: An overview, Applied Soft Computing. 107 (2021) 107373. https://doi.org/10.1016/j.asoc.2021.107373.

[15] J. Deriu, A. Lucchi, M. Gonzenbach, V. de Luca, F. Uzdilli, M. Jaggi, SwissCheese at SemEval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision, in: Proceedings of SemEval-2016, 2016: pp. 1124–1128. https://doi.org/10.18653/v1/s16-1173.

[16] C. Baziotis, N. Pelekis, C. Doulkeridis, DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), 2017: pp. 747–754. https://doi.org/10.1109/iembs.1997.757075.

[17] A.K. J, T.E. Trueman, E. Cambria, A Convolutional Stacked Bidirectional LSTM with a Multiplicative Attention Mechanism for Aspect Category and Sentiment Detection, Cognitive Computation. 13 (2021) 1423–1432. https://doi.org/10.1007/s12559-021-09948-0.

[18] S. Minaee, E. Azimi, A. Abdolrashidi, Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models, (2019). http://arxiv.org/abs/1904.04206.

[19] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis, Future Generation Computer Systems. 115 (2021) 279–294. https://doi.org/10.1016/j.future.2020.08.005.

[20] I. Priyadarshini, C. Cotton, A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis, Journal of Supercomputing. 77 (2021) 13911–13932. https://doi.org/10.1007/s11227-021-03838-w.

[21] Z. Hameed, B. Garcia-Zapirain, Sentiment Classification Using a Single-Layered BiLSTM Model, IEEE Access. 8 (2020) 73992–74001. https://doi.org/10.1109/ACCESS.2020.2988550.

[22] M. Al-Ayyoub, A.A. Khamaiseh, Y. Jararweh, M.N. Al-Kabi, A comprehensive survey of arabic sentiment analysis, Information Processing and Management. 56 (2019) 320–342. https://doi.org/10.1016/j.ipm.2018.07.006.

[23] I. Abu Farha, W. Magdy, A comparative study of effective approaches for Arabic sentiment analysis, Information Processing and Management. 58 (2021) 102438. https://doi.org/10.1016/j.ipm.2020.102438.

[24] O. Oueslati, E. Cambria, M. ben HajHmida, H. Ounelli, A review of sentiment analysis research in Arabic language, Future Generation Computer Systems. 112 (2020) 408–430. https://doi.org/10.1016/j.future.2020.05.034.

[25] M. Al-Ayyoub, S. Bani Essa, I. Alsmadi, Lexicon-based sentiment analysis of Arabic tweets, International Journal of Social Network Mining. 2 (2015) 101–114. https://doi.org/10.1504/IJSNM.2015.072280.

[26] R. Duwairi, M. El-Orfali, A study of the effects of preprocessing strategies on sentiment analysis for Arabic text, Journal of Information Science. 40 (2014) 501–513. https://doi.org/10.1177/0165551514534143.

[27] M. Nabil, M. Aly, A.F. Atiya, ASTD: Arabic Sentiment Tweets Dataset, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015: pp. 2515–2519. https://doi.org/10.18653/v1/D15-1299.

[28] H. Ibrahim, S. Abdou, M. Gheith, SENTIMENT ANALYSIS FOR MODERN STANDARD ARABIC AND COLLOQUIAL, International Journal on Natural Language Computing (IJNLC). 4 (2015) 95–109. https://doi.org/10.5121/ijnlc.2015.4207.

[29] A.A. al Sallab, R. Baly, G. Badaro, H. Hajj, W. el Hajj, K.B. Shaban, Deep Learning Models for Sentiment Analysis in Arabic, 2015.

[30] A. Alayba, V. Palade, M. England, R. Iqbal, Arabic language sentiment analysis on health services, Institute of Electrical and Electronics Engineers (IEEE). (2017) 114–118. https://doi.org/10.1109/asar.2017.8067771.

[31] A. Alayba, V. Palade, M. England, R. Iqbal, A combined CNN and LSTM model for Arabic sentiment analysis, Proc. International Cross-Domain Conference for Machine Learning and Knowledge Extraction. 11015 (2018) 179–191. https://doi.org/10.1007/978-3-319-99740-7_12.

[32] N. Abdulla, N. Ahmed, M. Shehab, M. Al-Ayyoub, Arabic sentiment analysis: Lexicon-based and corpus-based, in: 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT, IEEE, 2013: pp. 1–6. https://doi.org/10.1109/AEECT.2013.6716448.

[33] A. Oussous, F.-Z. Benjelloun, A. Lahcen, S. Belfkih, ASA: A framework for Arabic sentiment analysis, Journal of Information Science. 46 (2020) 544–559. https://doi.org/10.1177/0165551519849516.

[34] H. Elfaik, E.H. Nfaoui, Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text, Journal of Intelligent Systems. 30 (2021) 395–412. https://doi.org/10.1515/jisys-2020-0021.

[35] M. Heikal, M. Torki, N. El-Makky, Sentiment Analysis of Arabic Tweets using Deep Learning, Procedia Computer Science. 142 (2018) 114–122. https://doi.org/10.1016/j.procs.2018.10.466.

[36] N. Al-Twairesh, H. Al-Negheimish, Surface and deep features ensemble for sentiment analysis of Arabic tweets, IEEE Access. 7 (2019) 84122–84131. https://doi.org/10.1109/ACCESS.2019.2924314.

[37] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, Y. Al-Ohali, AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets, Procedia Computer Science. 117 (2017) 63–72. https://doi.org/10.1016/j.procs.2017.10.094.

[38] O. Alharbi, Negation Handling in Machine Learning-Based Sentiment Classification for Colloquial Arabic, International Journal of Operations Research and Information Systems. 11 (2020) 33–45. https://doi.org/10.4018/ijoris.2020100102.

[39] S. Kaddoura, M. Itani, C. Roast, Analyzing the effect of negation in sentiment polarity of facebook dialectal arabic text, Applied Sciences. 11 (2021) 1–13. https://doi.org/https://doi.org/10.3390/app11114768.

[40] A.B. Nassif, A. Elnagar, I. Shahin, S. Henno, Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities, Applied Soft Computing. 98 (2020) 1–27. https://doi.org/https://doi.org/10.1016/j.asoc.2020.106836.

[41] A. Gulli, S. Pal, Deep Learning with Keras, Packt, Birmingham, 2017.

[42] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing. 45 (1997) 2673–2681. https://doi.org/10.1109/78.650093.

[43] D. Otter, J. Medina, J. Kalita, A Survey of the Usages of Deep Learning for Natural Language Processing, IEEE Transactions on Neural Networks and Learning Systems. 32 (2019) 604–624. https://doi.org/10.1109/TNNLS.2020.2979670.

[44] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: pp. 1746–1751. https://doi.org/10.3115/v1/d14-1181.

[45] M. Cliche, BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs, (2018) 573–580. https://doi.org/10.18653/v1/s17-2094.

[46] A.A. Altowayan, L. Tao, Word embeddings for Arabic sentiment analysis, Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. (2016) 3820–3825. https://doi.org/10.1109/BigData.2016.7841054.

[47] K.M. Alomari, H.M. Elsherif, K. Shaalan, Arabic tweets sentimental analysis using machine learning, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 10350 LNCS (2017) 602–610. https://doi.org/10.1007/978-3-319-60042-0_66.

[48] S.R. El-Beltagy, NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016. (2016) 2900–2905.

[49] A. Nasser, H. Sever, A concept-based sentiment analysis approach for Arabic, International Arab Journal of Information Technology. 17 (2020) 778–788. https://doi.org/10.34028/iajit/17/5/11.

[50] M. Aly, A. Atiya, LABR: A large scale arabic book reviews dataset, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2013: pp. 494–498. https://doi.org/10.13140/2.1.3960.5761.

[51] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, N. Habash, CAMeL tools: An open source python toolkit for arabic natural language processing, in: Proceedings of the

12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association, 2020: pp. 7022–7032.

[52] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: The 3rd International Conference for Learning Representations, 2015: pp. 1–15.

[53] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Trans Assoc Comput Linguist. 5 (2017) 135–146. https://doi.org/https://doi.org/10.1162/tacl_a_00051.

[54] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research. 7 (2006) 1–30.

[55] A. Dahou, M.A. Elaziz, J. Zhou, S. Xiong, Arabic Sentiment Classification Using Convolutional Neural Network and Differential Evolution Algorithm, Computational Intelligence and Neuroscience. 2019 (2019) 1–16. https://doi.org/10.1155/2019/2537689.

[56] S. Al-Azani, E.S.M. El-Alfy, Hybrid Deep Learning for Sentiment Polarity Determination of Arabic Microblogs, Neural Information Processing. 10635 (2017) 491–500. https://doi.org/10.1007/978-3-319-70096-0_51.

[57] N. Bolbol, A. Maghari, Sentiment Analysis of Arabic Tweets Using Supervised Machine Learning, in: 2020 International Conference on Promising Electronic Technologies (ICPET), IEEE, 2020: pp. 89–93. https://doi.org/10.1109/ICPET51420.2020.00025.

[58] S. Islam, W. Eberle, S.K. Ghafoor, M. Ahmed, Explainable Artificial Intelligence Approaches: A Survey, (2021) 1–14. http://arxiv.org/abs/2101.09429.

[59] L. Bacco, A. Cimino, F. Dell'orletta, M. Merone, Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach, Electronics (Switzerland). 10 (2021) 1–19. https://doi.org/10.3390/electronics10182195.

[60] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United, J Med Internet Res. 23 (2021). https://doi.org/10.2196/26627.

[61] K. Dashtipour, M. Gogate, E. Cambria, A. Hussain, A novel context-aware multimodal framework for persian sentiment analysis, Neurocomputing. (2021) 377–388. https://doi.org/https://doi.org/10.1016/j.neucom.2021.02.020.

[62] N. Majumder, R. Bhardwaj, S. Poria, A. Gelbukh, A. Hussain, Improving aspect-level sentiment analysis with aspect extraction, Neural Computing and Applications. (2020). https://doi.org/https://doi.org/10.1007/s00521-020-05287-7.

[63] E. Cambria, A. Hussain, Guest Editorial: A Decade of Sentic Computing, Cognitive Computation. 14 (2022) 1–4. https://doi.org/10.1007/s12559-021-09972-0.

[64] Y. Susanto, E. Cambria, B.C. Ng, A. Hussain, Ten Years of Sentic Computing, Cognitive Computation. 14 (2022) 5–23. https://doi.org/10.1007/s12559-021-09824-x.

[65] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis, Proceedings of LREC (2022). (2022).