Aspect-Based Sentiment Analysis for Arabic Food Delivery Reviews

IBRAHIM AL-JARRAH, AHMAD M. MUSTAFA, and HASSAN NAJADAT, Jordan University of Science & Technology, Jordan

Business customers and consumers share their reviews online on social platforms such as Twitter. So, Twitter data sentiment analysis is extremely useful for both research and commercial purposes. Manually analyzing reviews takes a long time and effort, hence, automatic sentiment analysis is required. In this paper, we address aspect-based sentiment analysis for Arabic food delivery reviews using several deep learning approaches. In particular, we propose to use Transformer-based models (GigaBERT and AraBERT), Bi-LSTM-CRF, and LSTM, as well as a classical machine learning algorithm (SVM). We also present our dataset of food delivery service reviews, which we collected from Twitter. We annotated them and used them for training and evaluating our approaches.

The experiments show that both GigaBERT and AraBERT outperformed the other models in all the tasks. The Transformer-based models received F1-scores of 77% in the aspect terms detection task, 82% in the Aspect category detection task, and 81% in the aspect polarity detection task, gaining 2%, 4%, and 4% over Bi-LSTM-CRF and LSTM in the first, second, and third tasks respectively.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Information systems** → **Sentiment analysis**.

Additional Key Words and Phrases: Aspect-based Sentiment classification, Deep Learning, BERT

1 INTRODUCTION

The internet has revolutionized the way people communicate and interact with each other. The emergence of social media platforms such as Twitter and Facebook has provided a new avenue for individuals to connect and share their ideas, information, and feelings online. Social media has become an essential tool for communication and business services, making it possible for companies and organizations to interact with their customers in various fields [5]. Among these social media platforms, Twitter is one of the most popular and widely used platforms worldwide, with millions of members. Twitter was founded in February 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams [7]. The platform has become accessible to everybody, allowing for interactive communication between individuals and the expression of opinions and ideas. With the growth and development of Twitter in recent years and the increase in the number of users, Twitter has become a virtual community for people to express their opinions. The platform offers features such as groups and pages for businesses and restaurants to allow people to express their opinions freely.

Twitter data has become a valuable resource for researchers to investigate the opinions of people on a particular topic or product [4]. However, the volume of data or tweets can be enormous, making it difficult to process manually. Reading all these reviews manually takes a long time, and therefore automation is inevitable. Researchers have tackled these challenges through various fields, making the analysis of Twitter data more accessible and efficient.

Authors' address: IBRAHIM AL-JARRAH, imaljarrah18@cit.just.edu.jo; AHMAD M. MUSTAFA, ammustafa@just.edu.jo; HASSAN NAJADAT, najadat@just.edu.jo, Jordan University of Science & Technology, Irbid, Jordan, 22110.

2375-4699/2023/6-ART $15.00
https://doi.org/10.1145/3605146

Aspect-Based Sentiment Analysis (ABSA) is a task within the field of Natural Language Processing (NLP). Its primary objective is to extract aspects from a given text, along with their corresponding category and polarity [15]. In contrast to traditional Sentiment Analysis (SA) [10] which categorizes texts into neutral, negative, or positive, ABSA provides more detailed insights into a given product or service.

ABSA is of significant research value as it enables consumers to evaluate a product or service from a more holistic standpoint, allowing for a more complete and transparent understanding [28]. By identifying and analyzing the different aspects of a product or service, ABSA provides a more comprehensive picture of its strengths and weaknesses, enabling businesses to better understand and address the needs of their customers.

In food delivery reviews, a sentence may contain several aspects and sentiments, for instance, the following example shows a tweet, complaining about a food delivery system. This example illustrates the necessity for ABSA in this domain.

لا يمكننا تعديل العنوان ، ولم يتم الاتصال بالمطعم للأسف خدمة سيئة

English Translation: **We can not correct the address, and the restaurant has not been contacted, unfortunately, poor service**

The complaint mentioned two aspects:

(1) تعديل العنوان **(correct the address)** indicates a *location* and the sentiment polarity is *negative.*

(2) خدمة سيئة **(poor service)** refers to a *service* with *negative* polarity.

Numerous studies have investigated the use of ABSA (Aspect-Based Sentiment Analysis) for English reviews [55, 57]. However, Arabic is considered a low-resource language [16], which has resulted in limited research on this topic [8, 11]. Our objective is to aid businesses and organizations in comprehending their customers' expectations and desires regarding food delivery services. In addition, these models can provide performance indicators for client satisfaction and business services. The following section outlines the novelty of our approach.

- We aim to develop a gold standard dataset for three aspects of aspect-based sentiment analysis (ABSA) - aspect terms detection (T1), aspect categorization (T2), and aspect sentiment polarity estimation (T3). To accomplish this, we collected Arabic tweets about food delivery services and collaborated with a team of Arabic native speakers to annotate the data.
- Our dataset includes a variety of Arabic dialects, including the most frequent ones - Egyptian (EGP) and Gulf (including Iraqi (GLF) and Jordanian (JOR)) - as well as Modern Standard Arabic (MSA).
- To address the ABSA tasks, we use five different approaches: Bi-LSTM-CRF, LSTM, GigaBERT, AraBERT (all deep learning models), and a classic machine learning algorithm (SVM). Our evaluation results show that the BERT-based models outperform the other approaches.

The remainder of this paper is organized into four sections. Section 2 provides a review of previous relevant work in sentiment analysis (SA) and aspect-based sentiment analysis (ABSA) for both English and Arabic. Section 3 outlines the materials and methods we recommend, including details about the dataset, data pre-processing, feature extraction, and machine learning models we used. In Section 4, we present the experimental outcomes of our used methods, including the evaluation metrics and results we obtained. Finally, in Section 5, we conclude the paper by summarizing our findings and suggesting future directions for research in this area.

2 RELATED WORK

Deep learning has gained popularity in natural language processing (NLP) tasks such as Semantic Search [37], Chatbots [39], Social media analysis [24], and Customer satisfaction analysis [22]. The most common approaches for NLP tasks are Transformers [18] and Recurrent Neural Networks (RNN) [19].

While there has been little research on Aspect-Based Sentiment Analysis (ABSA) in food delivery services, several studies have applied Sentiment Analysis (SA) to food delivery data. These studies have utilized deep learning models based on Transformers (BERT) [46, 58], Convolutional Neural Networks (CNN) [36, 38, 43, 56], RNN [38, 56], and LSTM [43]. Other studies have proposed Regression [25], Decision tree [30], Random Forest [45], Support vector machine (SVM) [20, 27, 31], Naïve Bayes [14, 30], and KNN for SA. In this study, we have performed ABSA using both deep learning and SVM and found that deep learning approaches are more effective at capturing the features of review aspects and, as a result, produce more accurate results. However, these deep learning approaches were primarily used to detect the general sentiment of text, classifying the entire sentence as either positive or negative.

Several studies have performed ABSA on datasets from the SemEval-2014 competition, particularly the Restaurants, Laptops, and Twitter datasets. Recursive Conditional Random Field (CRF), LSTM, and LSTM with hand-crafted features are among the models proposed for ABSA in English [13]. Song et al. [49] proposed a novel strategy for using BERT's intermediate layers to enhance the performance of fine-tuning BERT for ABSA tasks.

Given the lack of ABSA data in the Arabic language [51], we present studies conducted on Arabic reviews (mostly tweets) of several other domains, such as Arabic book reviews, news reviews, hotel reviews, and applications and phone reviews.

For Arabic tweets, Ashi et al.[26] evaluated different ABSA embedding models for Arabic tweets using AraVec-Web and FastText Arabic Wikipedia templates for aspect recognition and an SVM classifier to determine polarity. The dataset consisted of approximately 5,000 tweets, and the results showed that fastText Arabic in Wikipedia is better than other models. Aspect detection achieved 70% accuracy, and sentiment polarity detection achieved 89%. Alassaf Manar [35] examined tweets regarding Qassim University in Saudi Arabia to identify the institution's strengths and drawbacks, using a feature selection method and an SVM classifier. The suggested model received an F-1 measure of 70% for aspect detection and 87% for aspect polarity. Masadeh [44] developed a strategy to identify aspect polarity that combines a corpus-based approach with a lexical-based approach, using a dataset of 1,000 reviews. The results revealed an accuracy of 78% for polarity classification and 30% for aspect polarity.

For Arabic book reviews, Al-Smadi et al. [8] studied ABSA in Arabic to review Arabic books using approximately 1,513 reviews. They employed the Dice coefficient similarity measure to determine the distance between sentence phrases for training and testing. The F-1 score for aspect term extraction yielded a 23% success rate, while the classification accuracy for aspect term extraction indicated a 42.57% success accuracy. Behdenna et al. [48] suggested an aspect-based sentiment analysis approach to Arabic book reviews using the semantics of description logics and linguistic rules [6] to identify entities, aspects, and opinions. The authors applied the proposed approach to review Arabic books (HAAD Dataset). Salima et al. [59] suggested using description logic to identify explicit aspects and their polarities in a HAAD dataset for ABSA tasks.

For Arabic News reviews, Al-Smadi et al. [11] also assessed the impact of news publications on readers, using ABSA to analyze Arabic-language news posts about the Israeli raid on Gaza in 2014. They used classifiers including CRF, J48, NB, and K-NN to extract properties such as POS, NER, and N-Grams, and found that J48 performed best in extracting aspect terms and that CRF and NB performed best in detecting the polarity of the aspect term. Rajae and Taher [60] recommended using transfer learning with pre-trained Arabic language models such as AraBERT for ABSA tasks, including aspect term extraction and identification of aspect categories.

For Arabic Hotel reviews, Al-Smadi et al.[12] improved the model provided in [8] using the SVM with n-Unigrams derived from Arabic hotel reviews, using a threshold value of 0.2. The findings indicated that the F-score was 40%. The authors then compared machine learning methods for ABSA of Arabic Hotels reviews in [23]. Two approaches were employed: RNN and SVM, the two approaches were trained on the dataset. The results showed that SVM classifier outperformed RNN, achieving an accuracy of 95% in the sentiment polarity, and a 93% F-1 score in the aspect category. Al-Smadi et al. also developed two models in [32], to obtain aspect terms at the word level for neural networks, utilizing the first Bi-LSTM-CRF model based on word2vec word embeddings. The other is the AB-LSTM-PC ad hoc sentiment polarity classification model. The results showed an outperformance over the baseline. The Bi-LSTM-CRF model received F-1 scores of 66% for extracting aspect terms. The AB-LSTM-PC model had an accuracy of (82.7%) which outperformed the baseline sentiment polarity classification compared to the conventional approach with the lexicon (76.4%). Trigui et al. [29] proposed a system for detecting ABSA aspects. Features were extracted from the SemEval 2016 collection of hotel reviews. These characteristics included POS, NER, Syntactic, and Numeric, with a total of 4802 reviews. They used the Adaboost, DT, NB, and RepTree classifiers. The results showed the superiority of the Adaboost classifier, as it produced outcomes in terms of accuracy (97.9%) and recall (96.9%). Abdelgwad et al. [52] employed two GRU-based models for ABSA to extract both the primary opinionated aspects (T1) and aspects sentiment polarity (T2). For (T1), the first model includes a bidirectional GRU, CNN, and CRF, whereas the second model employs an interactive attention network based on a bidirectional GRU (T2). He utilized a dataset comprised of Arabic-language hotel reviews. The results showed that the proposed approaches outperformed prior research's, with an F1 score of 70.67% in (T1) and an accuracy of 83.98% in (T2). Al-Dabet et al. [54] proposed two deep learning models for handling ABSA challenge, establishing the aspect category and categorization of aspects-sentiments. The first challenge was to use a model to determine the aspect class through a CNN with the use of (Indy-LSTM). A stacked (Bi-Indy-LSTM) network and several attention layers were used in the second challenge. They used the ArabicSemEval-2016 dataset. The findings demonstrated superiority over other models, The model achieved an F-1 measure of 58% in the first challenge and accuracy of 87.31% in the second.

For Arabic Telecom, Apps, and phone reviews, Alshammari and Almansour [33] compared between classical machine learning and deep learning approaches on tweets in the Arabic language, as well as the impact of POS tagging and word embedding on deep learning. They used tweets on Saudi telecom business, with a total of 1098 tweets. The findings indicated that deep learning with word embedding reached an F-1 score of 81% and that the Unigram language model outperformed the Bigram language model. Agreed et al. [34] presented an approach that worked on lexicon- and rule-based systems for classifying terms and extracting the aspect. They used a dataset of 2071 reviews of Arab government applications. The top aspect of extraction accuracy was 96% and F-1 of 92%, while the sentiment classification job obtained an accuracy of 95.8% and F-1 measures of 90%. Alhamad and Kurdy [42] presented a report on the evaluation of products to work on improving the services related to the phone. Three dictionaries have been extracted: the first is related to the telephone, the second is related to the entities, and the last is to the words of feelings with their poles. They gathered about 1,594 comments from Facebook ad posts. The results showed that the model employed F-1 measures of 88%. Al-Ayyoub et al. [17] built a dataset for mobile device reviews in Arabic, to work with two tasks: aspect category prediction and sentiment polarity. Use to extract features n-grams, and for training data Use an SVM classifier, results are shown in the task aspect category F1 measure is 0.315%, and in the second task, the sentiment polarity assignment is F1 measure is 73%.

Our approach is different than the presented studies in two ways. Unlike previous work which focus on measuring the sentiment of a review, we aim to discover all the aspects within a review, because there might be more than one aspect, moreover, we detect the category and polarity of each aspect. The second difference is that we target Arabic language.

3 MATERIALS AND METHODS

In order to perform ABSA tasks, we collected our dataset and applied the preprocessing pipeline. Then we applied our suggested classification methods. Figure 1 shows our methodology and the pre-processing steps applied to the data. We describe these parts in the next sections.

3.1 Dataset Collection and Annotation

We collected the data using Twitter API. We use keywords related to food delivery services. We observed that choosing a short form of the keywords maximizes the number of retrieved Tweets. For example, the keyword

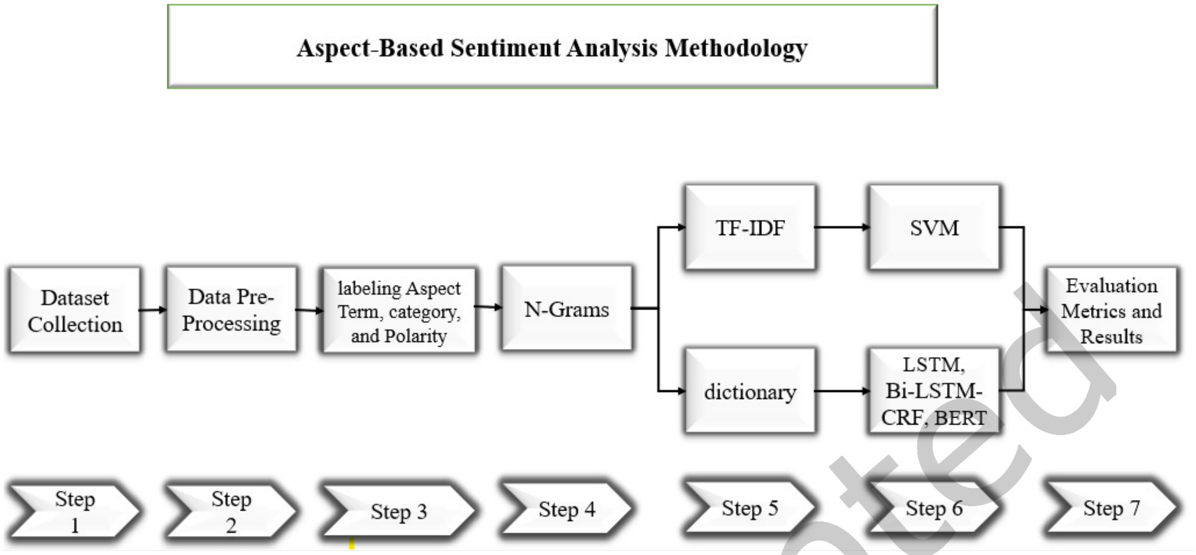


Fig. 1. The proposed methodology.

Table 1. All Keywords used to collection dataset

Keywords	English Translation
سعر	Price
اخر	Late
شكر	Thank
خدمه	Service
لوكيشن	Location
مكان	Location
عوض	Compensation

"شكر" (which means "Thank") returns tweets with several word variations ("شكرا", "شكراً", etc..) in addition to the exact keyword. Table 1 shows our keywords.

After manually removing unrelated Tweets, our dataset contained a total of 4,879 tweets. We have manually detected 23,733 aspects, each with one category and sentiment polarity. Non-aspect terms have neither category nor polarity. We observed that our aspects can be categorized into four categories; *Price*, *Time*, *Service*, and *Location*. Table 2 describes these categories. We also label each aspect with its sentiment polarity *Positive* or *Negative*.

Table 2. Description of each aspect category

Category	Description
Service	Tweets related to Service issues such as (rude customer service, etc...)
Price	Tweets related to price issues such as (High prices, method of payment, etc...)
Time	Tweets related to time issues such as (around late deliveries, cold food being delivered, etc...)
Location	Tweets about a location-related concern, such as (incorrect orders being delivered, the driver requiring a lot of guidance to find the delivery location., etc...)

Table 3. Abbreviations Table

Abbreviations	Description
T1	Task 1: Aspect term identification
T2	Task 2: Aspect category identification
T3	Task 3: Aspect sentiment polarity
B	Beginning of a chunk
I	Inside a chunk
O	Outside any chunk

Manual annotation is more accurate and produces less mislabeled terms as compared to the automatic approaches of labeling. To assure the quality of our labels, three Native speakers have labeled each term in the dataset. Each annotator determined whether a word is an aspect or not then annotates the aspect category and polarity. Majority voting is applied to select the final labels.

For instance, in a review:

أسعار مقبولة ولكنها تصل متأخرة وخدمة العملاء لا تستجيب لنا

"Prices are acceptable, but they arrive late and customer service does not respond to us"

We first label aspects for T1 (see Table 3 for the abbreviations). Here we found "**prices**" and "**late**". Aspect terms are labeled "B", and non-aspect terms are labeled "O". If an aspect is a phrase of more than one word, we label the first word "B" and the remaining words "I". So, a word has one of three labels: "B", "I", and "O". Then we assign a category (T2) for each term, for example, **prices** is labeled *Price* and **late** is *Time*. Finally, we annotate the sentiment polarity of each term. For example, **prices** is *Positive* and **late** is *Negative*. Figure 2 illustrates the example.

To our utmost knowledge, our data is the only publicly available dataset on Arabic reviews of food delivery services. The dataset is available for noncommercial use[1].

We counted the number of aspect terms, category, and sentiment polarity, and have shown the statistics in Tables 4, 5, and 6.

Figure 3 depicts the word count distribution in the dataset. This is necessary for defining the data padding sequence length. We observed that the average sentence in the dataset contains about 21 words, and the longest sentence in the corpus contains 58 words.

[1] https://github.com/IBRAHIMALJRRAH/DataSet-Aspect-Based-Sentiment-Analysis.git

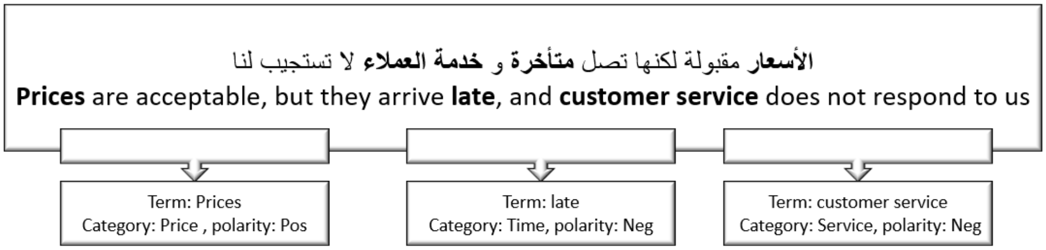

الأسعار مقبولة لكنها تصل متأخرة و خدمة العملاء لا تستجيب لنا
Prices are acceptable, but they arrive late, and customer service does not respond to us
Term: Prices
Category: Price , polarity: Pos
Term: late
Category: Time, polarity: Neg
Term: customer service
Category: Service, polarity: Neg

Fig. 2. An example of using models to classify aspects of Term, category, and sentiment polarity in a review sentence

Table 4. Number of terms of each aspect label

No	Labels of Aspect Terms	The number of terms
1	Beginning (B)	13,650
2	Inside (I)	10,083
3	Outside (O)	80,741
4	Total	104,374

Table 5. Number of terms per aspect category

No	Aspect Category	The number of terms
1	Service	12,071
2	Price	5,103
3	Time	5,091
4	Location	1,468
5	Total	23,733

Table 6. Number of terms per sentiment polarity

No	Aspect Polarity	The number of terms
1	Negative	17,965
2	Positive	5,768
3	Total	23,733

3.2 Text Pre-Processing and Cleaning

Text pre-processing is used as a method to clean the text in order to make it ready to feed to models. It is the initial stage in every NLP project. Table 7 shows an example from the dataset before and after applying each pre-processing step. Pre-processing stages include the following:

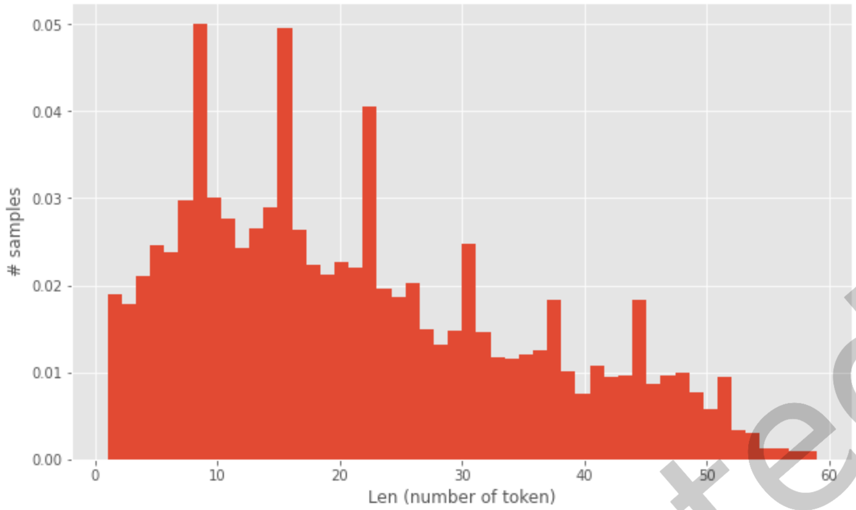


Fig. 3. The distribution of words count in the dataset. The x-axis represents the length of sentences. The y-axis shows the number of sentences falling in each length category

- Delete non-Arabic words.
- Delete punctuation marks and symbols including("? ' ! @ $ # |")
- Delete Arabic diacritics (اُ اِ اٌّ اَّ).
- Delete hyperlinks and hashtags.
- Delete Stop words.

3.3 Datasets Format and Convert To XML

We have chosen the format of the SemEval-2015 Task 12 [9] dataset to be the format for our dataset. The data is in XML format. Each aspect has three attributes (aspect terms, category, and sentiment polarity) See an example of the dataset format in Figure 4.

3.4 Features Extraction

A Feature Extraction phase was done to prepare the data for analysis. This step includes NGrams [12], and Term Frequency–Inverse Document Frequency (TF-IDF) [33]. We use these features because they improve the performance of ABSA in Arabic. Extracted features are as follows:

N-Grams. We use the IOB2 scheme chunking, proposed by Sang et al. [2], to annotate review sentences, where the "B-Aspect" denotes the beginning of a chunk, the "I-Aspect" shows that the tag is included inside a chunk, and the "O" tag indicates that the token does not belong to any entity or chunk. N-Grams for T1 (aspect terms identification).

Word Features. This set of features was utilized in ABSA-related work. TF-IDF was produced for the SVM model. Also, we build a dictionary (word2index) that assigns a unique integer value to every word. word2index corpus is used for LSTM and Bi-LSTM-CRF, which reduces memory usage. BERT is a pre-trained model. Since it demands input data in a specified format, GigaBERT and AraBERT special tokens will be required to denote

Table 7. Text Pre-Processing and Cleaning

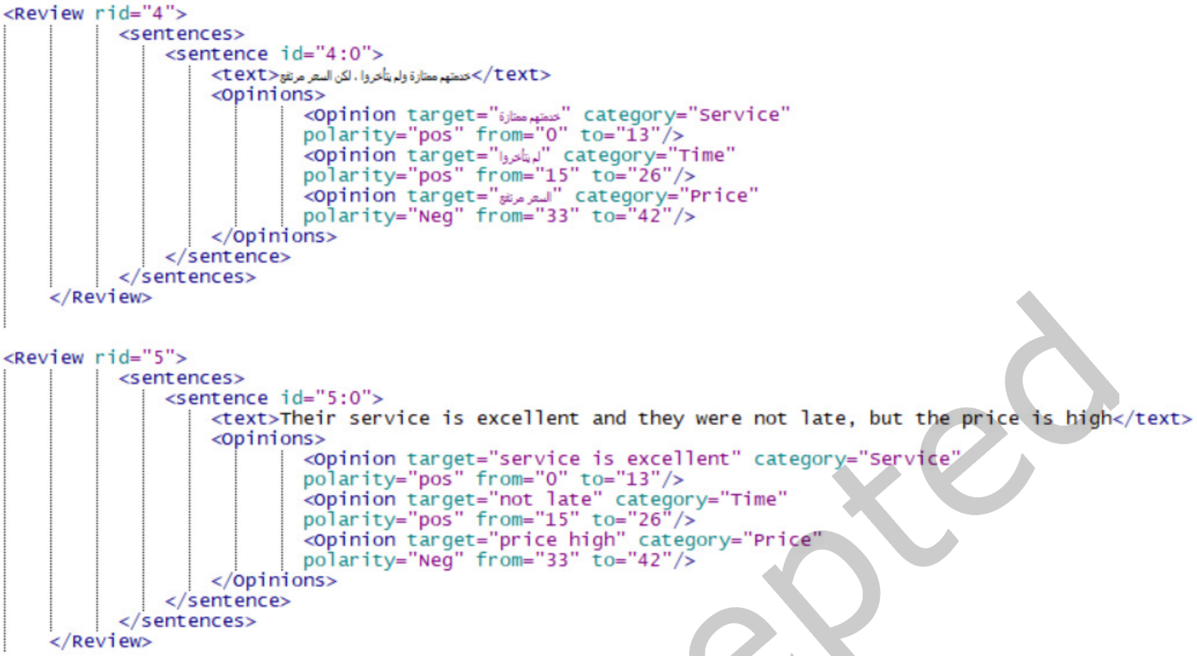
Stages Text Pre-Processing and Cleaning	Output Pre-Processing
Text English language	’@Talabat The worst application and the worst service. I ordered at half past 6, and until now it hasn’t arrived yet, and the problem I paid!!!:)!!’
Text Arabic language	’@Talabat اسوء تطبيق واسوء خدمه طالبه من ٦ ونص وللحين ماوصل والمشكلة دافعه فلوسي!!!:)!!”
Delete non-Arabic words.	” اسوء تطبيق واسوء خدمه طالبه من @ ونص وللحين ماوصل والمشكلة دافعه فلوسي!!!:)!!”
Delete punctuation marks and symbols	اسوء تطبيق واسوء خدمه طالبه من ونص وللحين ماوصل والمشكلة دافعه فلوسي
Delete Arabic diacritics	اسو تطبيق واسو خدمه طالبه من ونص وللحين ماوصل والمشكلة دافعه فلوسي
Delete hyperlinks	اسو تطبيق واسو خدمه طالبه من ونص وللحين ماوصل والمشكلة دافعه فلوسي
Delete Stop words	اسو تطبيق واسو خدمه طالبه ونص ماوصل والمشكلة دافعه فلوسي
Buckwalter Transliteration	Asw tTbyq wAsw xdmh TAlbh wnS mAwSl wAlm$klp dAfEh flwsy

the beginning ([CLS]) and separation/end of sentences ([SEP]) tokens that conform to the fixed vocabulary. We convert the phrase from a collection of strings to a list of vocabulary indices after dividing the text into tokens.

3.5 Suggested Approaches

We suggest several approaches to analyzing ABSA for the three tasks. We use Transformer-based models (GigaBERT and AraBERT), Bi-LSTM-CRF, LSTM, and a classical machine learning model (SVM). We next go through these recommended approaches.

GigaBERT. GigaBERT [40] is a language model that is specifically designed for cross-lingual transfer from English to Arabic. It is trained on a combination of newswire text from the Gigaword corpus, as well as data from Wikipedia and web crawls. GigaBERT is based on Bidirectional Encoder Representations from Transformers (BERT) language model. BERT is built on transformers that understand contextual links between words in a text; when this is evident from its common value, attachment, notation, and position. In contrast to conventional single-pass converters, which scan the input text sequentially from left to right or vice versa. BERT can interpret the text in both renderings simultaneously. Figure 5 shows the architecture of GigaBERT. In this paper, we employ GigaBERT-v3 [41], which has been pre-trained with around 10 billion tokens, and we use it in three tasks (T1, T2, T3).



```
<Review rid="4">
        <sentences>
            <sentence id="4:0">
                <text>خدمتهم ممتازة ولم يتأخروا ، لكن السعر مرتفع</text>
                <Opinions>
                        <Opinion target="خدمتهم ممتازة" category="Service"
                        polarity="pos" from="0" to="13"/>
                        <Opinion target="لم يتأخروا" category="Time"
                        polarity="pos" from="15" to="26"/>
                        <Opinion target="السعر مرتفع" category="Price"
                        polarity="Neg" from="33" to="42"/>
                </Opinions>
            </sentence>
        </sentences>
    </Review>

<Review rid="5">
        <sentences>
            <sentence id="5:0">
                <text>Their service is excellent and they were not late, but the price is high</text>
                <Opinions>
                        <Opinion target="service is excellent" category="Service"
                        polarity="pos" from="0" to="13"/>
                        <Opinion target="not late" category="Time"
                        polarity="pos" from="15" to="26"/>
                        <Opinion target="price high" category="Price"
                        polarity="Neg" from="33" to="42"/>
                </Opinions>
            </sentence>
        </sentences>
    </Review>
```

Fig. 4. Shows the XML file

AraBERT. AraBERT [1] is a pre-trained Arabic language model based on Google's BERT architecture. AraBERT employs the same BERT-base configuration. Their final pre-training dataset is 24 GB in size and contained around 3 billion words. Two versions of AraBERT were released, AraBERTv2 and AraBERTv1, we have chosen to use AraBERTv2.

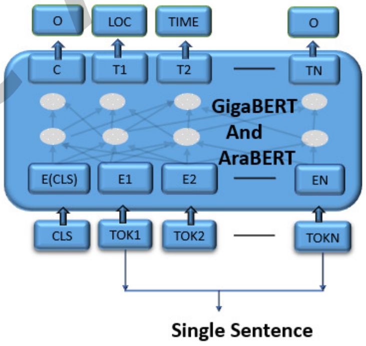


Fig. 5. GigaBERT and AraBERT Architecture

LSTM. LSTM is a form of Recurrent neural network (RNN) that can learn Long short-term memory information [3] . In many problems, LSTM has been very successful and widely used. Among the practical applications that use LSTM are speech recognition, handwriting recognition, text translation from one language to another, sign language translation, traffic prediction, and many more. LSTM is designed for handling the vanishing gradient problem faced by RNN. All RNNs contain a sequential form of recurrent neural network modules. A standard LSTM consists of three gates: input gate determines fresh information in the input, output gate identifies which information is no longer required, and forget gate determines what to output based on the input and the memory cell's content. The architecture of the LSTM model is shown in Figure 6. The example shown explains how LSTM classifies the polarity of an aspect in T3.

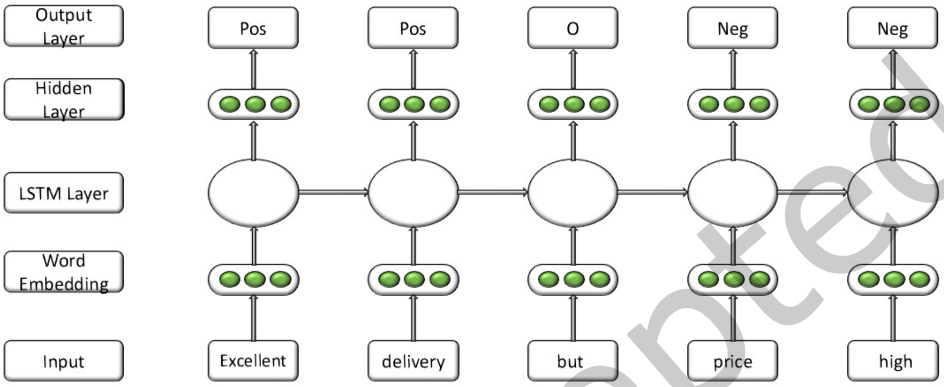


Fig. 6. The aspect-based LSTM for sentiment polarity model

Bi-LSTM-CRF:. To do tasks T1 and T2 (aspect term, and category), we create an ABSA model using Bi-LSTM-CRF [21]. The CRF layer receives all predicted scores from the Bi-LSTM block. The tag sequence with the greatest prediction score will be chosen as the response in the CRF layer. Figure 7 depicts the structure of Bi-LSTM-CRF model.

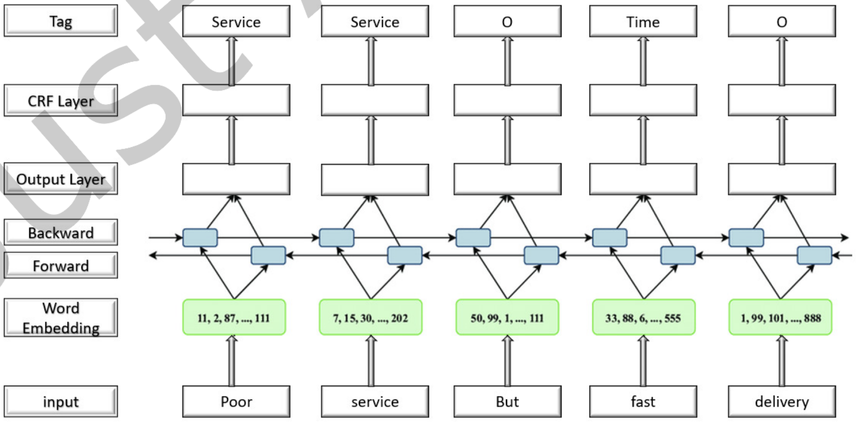


Fig. 7. The bidirectional LSTM with conditional random fields model

***SVM*.** SVM is one of the most important classical algorithms in the world of machine learning [47]. It is a supervised machine learning algorithm that can perform regression and classification operations and is more used in classification. SVM is used in Image Classification and Segmentation, and Category Assignment, and in NLP tasks. It is an algorithm based on the idea of finding a hyperplane that divides a dataset into two classes in the best way. The types of SVM are Linear and Non-Linear. We use SVM in the three tasks (T1, T2, T3). We build TF-IDF matrices from our dataset. Then we train classifiers for training and testing sets in SVM. TF-IDF is a statistical tool used for determining the relevance of specific textual elements.

4 EVALUATION METRICS AND RESULTS

4.1 Evaluation Metrics

To evaluate the methods, we use the following metrics: precision, recall, and F1-score. These metrics are adopted for several tasks in text mining, machine learning, and deep learning. Classified items can be True Positives-TP (i.e., the number of terms which the classifier correctly predicted as the positive), False Positives-FP (the number of cases where the model incorrectly predicted a negative term as positive), False Negatives-FN (the incorrectly predicted terms that are actually positive), and True Negatives-TN (the correctly classified negative terms). The metrics are given as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

We have computed Precision, Recall, and F1-Score for each class. As described in Section 3.3, our tasks have different class labels. For example the classes in task T1 are ‘I’, ‘O’, and ‘B’. For each class the Precision, Recall, and F1-Score are computed assuming the class is positive and the remaining classes are negative. Then macro average precision, recall, and f1-score are reported. The metrics for all tasks are reported in the section that follows.

4.2 Result

We have tested the used methods on a dataset of food delivery services reviews. Our suggested approaches are Bi-LSTM-CRF for tasks T1 and T2, LSTM for task T3, and GigaBERT, AraBERT and SVM for T1, T2, and T3. We randomly assigned 80% of the reviews in the dataset to training, 5% to validation, and 15% to testing. While training the models, we set the Max length in all the models to 58, which equals the maximum sentence length. We use a batch size of 12 for GigaBERT and AraBERT, and 32 for Bi-LSTM-CRF and LSTM. We perform 5 epochs of training for all the models. For both Bi-LSTM-CRF and LSTM, the word embedding size is 300, dropout is 0.1, and the learning rate equals 0.001. The other parameters are shown in Table 8.

Task 1: Detecting Aspect Terms (T1). Our first task is the aspect term detection (or T1). It is regarded as one of the most critical tasks in Aspect Based Sentiment Analysis. Each sentence is labeled with an IOB format (short for Inside, Out, and Begin). For instance, in the sentence.

Table 8. Parameters used to post-train all models

Model	Parameter	Values
GigaBERT, AraBERT	optimizer	AdamW
	learning rate	3.00E-05
	epsilon	1.00E-08
	max_grad_norm	1
	weight_decay_rate	0.1
Bi-LSTM-CRF	optimizer	Adam
	Loss_Function	CRF
LSTM	optimizer	RMSprop
	Loss_Function	categorical_crossentropy
SVM	All Parameters	Default sklearn values

تطبيق ممتاز ، لكن فريق التوصيل سيء.

"Excellent application, but the delivery staff is bad."

The phrase **Excellent application** is one aspect, with two terms. Note that, as mentioned in previous sections, if an aspect consists of only one term, the term is labeled as "B". On the other hand, if the aspect is a phrase of more than one term, the first term is labeled "B", and the remaining terms are labeled "I". Lastly, if a word is not part of an aspect, it is tagged "O".

Table 9. Results of applying the models to task T1 (Aspect term identification)

Model	Precision	Recall	F1-score
GigaBERT	79%	76%	77%
AraBERT	78%	77%	77%
BI_LSTM_CRF	76%	73%	75%
SVM	72%	60%	65%

Table 9 shows the results of applying GigaBERT, AraBERT, Bi-LSTM-CRF, and SVM to T1. GigaBERT and AraBERT outperformed the other models in terms of F1-score, achieving a precision greater than 78%, recall greater than 76%, and F1 of 77%, while Bi-LSTM-CRF model received a precision of 76%, recall of 73%, and F1 score of 75%, and SVM model received a precision of 72%, recall of 60%, and F1 of 65%.

Table 10 illustrates the performance of models for aspect terms detection. As seen in the table, GigaBERT and AraBERT show better performance as compared to Bi-LSTM-CRF and SVM. We observed that GigaBERT has missed many Inside (I) aspect terms more than AraBERT. This is evident by the Recall value. We also noticed that Bi-LSTM-CRF correctly detects Inside (I) terms, outperforming both GigaBERT and AraBERT by 1% in F1.

Table 10. T1 results per class

model	classes	precision	Recall	F1-score
GigaBERT	Beginning(B)	66%	67%	65%
	Inside(I)	73%	63%	68%
	O	97%	98%	98%
AraBERT	Beginning(B)	65%	66%	65%
	Inside(I)	72%	66%	68%
	O	97%	98%	98%
BI_LSTM_CRF	Beginning(B)	61%	54%	64%
	Inside(I)	70%	67%	69%
	O	98%	99%	98%
SVM	Beginning(B)	58%	34%	43%
	Inside(I)	69%	49%	59%
	O	89%	97%	93%

Figure 8 shows curves (F1-score and loss) of GigaBERT for the training and testing set in the first task. We stop training at the 5^{th} epoch when the validation loss starts to increase (while the training loss is less) indicating the beginning of overfitting.

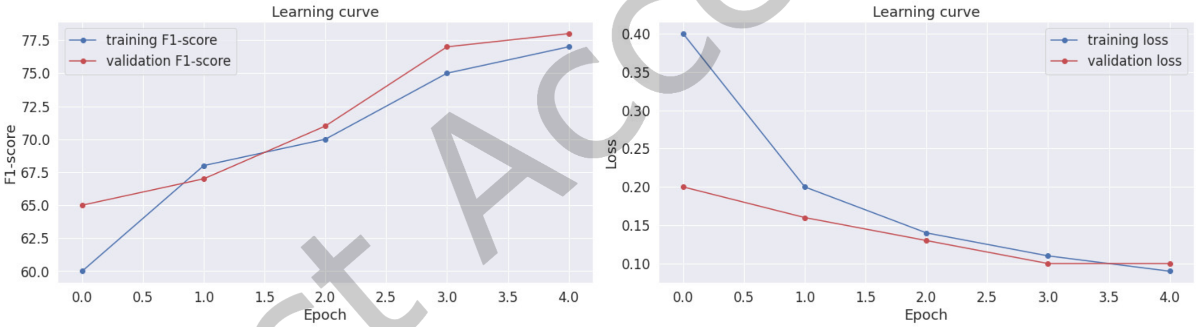


Fig. 8. The training and validation F1-score and loss at each epoch of GigaBERT training for task T1

Task 2: Detecting Aspect Category (T2). The second task (T2) is called the aspect category. The purpose of this task is to assign one aspect category to each aspect term in the sentence. As mentioned before, we specify four categories; service, price, time, and location. In our last example, the aspect category of **Excellent application** is *service*. Table 11 reports our results on T2.

Similar to the outcomes of T1, GigaBERT and AraBERT outperforms Bi-LSTM-CRF and SVM in terms of F1 with scores higher than 81%, while Bi-LSTM-CRF and SVM both scored 78% and 65% respectively. Table 12 reports the performance of each individual category on T2. The results are similar for each category. We noticed that Bi-LSTM-CRF is performing well when predicting Time category with F1-score of 85%, increasing by 3% over AraBERT and GigaBERT. See the curves (F1-score and loss) for the training and the testing set in the second task in Figure 9.

Table 11. The results of the models on task T2 (Aspect category detection)

Model	Precision	Recall	F1-score
AraBERT	83%	81%	82%
GigaBERT	83%	79%	81%
BI_LSTM_CRF	84%	74%	78%
SVM	78%	60%	65%

Table 12. T2 results per class

model	Category	precision	Recall	F1-score
AraBERT	Price	84%	82%	84%
	Service	80%	76%	79%
	Time	84%	81%	82%
	Location	84%	76%	80%
GigaBERT	Price	84%	80%	83%
	Service	80%	76%	77%
	Time	84%	80%	82%
	Location	84%	76%	80%
Bi-LSTM-CRF	Price	85%	79%	81%
	Service	69%	75%	74%
	Time	90%	80%	85%
	Location	92%	62%	71%
SVM	Price	80%	60%	66%
	Service	74%	57%	59%
	Time	82%	64%	75%
	Location	75%	60%	60%

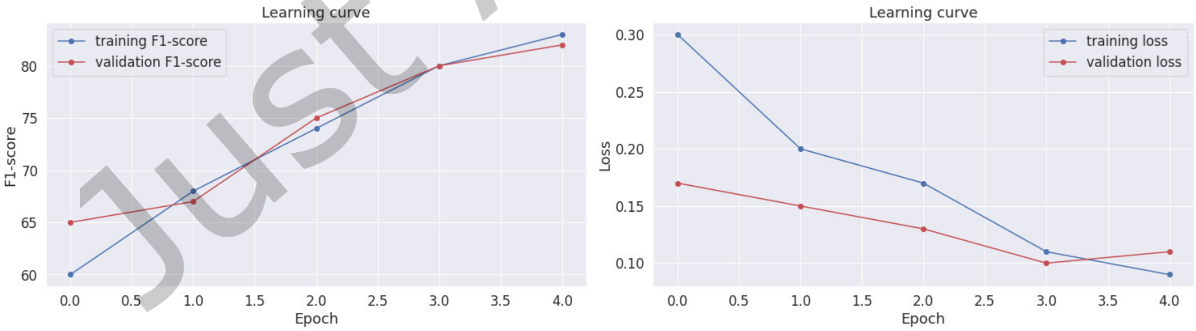


Fig. 9. The training and validation F1-score and loss at each epoch of GigaBERT training for task T2

Task 3: Aspect Sentiment Polarity (T3). Each aspect of sentiment polarity is detected. The aspect sentiment polarity is either positive or negative. In our previous example, the review **Excellent application** holds a *positive* sentiment.

Table 13. The results of the models on task T3 (Aspect polarity classification)

Model	Precision	Recall	F1-score
GigaBERT	83%	81%	81%
AraBERT	82%	77%	80%
LSTM	84%	72%	77%
SVM	79%	70%	71%

Table 14. T3 results of each Classes

model	classes	precision	Recall	F1-score
GigaBERT	negative	84%	83%	82%
	positive	81%	79%	81%
AraBERT	negative	83%	80%	81%
	positive	81%	74%	79%
LSTM	negative	85%	74%	79%
	positive	82%	70%	75%
SVM	negative	82%	70%	74%
	positive	77%	70%	68%

Table 13 shows the results of LSTM, GigaBERT, AraBERT, and SVM on T3. We observed that GigaBERT and AraBERT outperformed the other models with F1-score over 80%, whereas, LSTM and SVM scored 77% and 71% respectively. The performance on each polarity class is shown in Table 14. Note that the dataset is imbalanced, containing less number of positive aspects as compared to the negative. Hence, the performance of all the models is better on the negative class. We also noticed that GigaBERT outperforms AraBERT in T3 in terms of the recall. This was not obvious in T1 and T2. Figure 10 shows curves (F1-score and loss) for the training and the testing set in the third task.

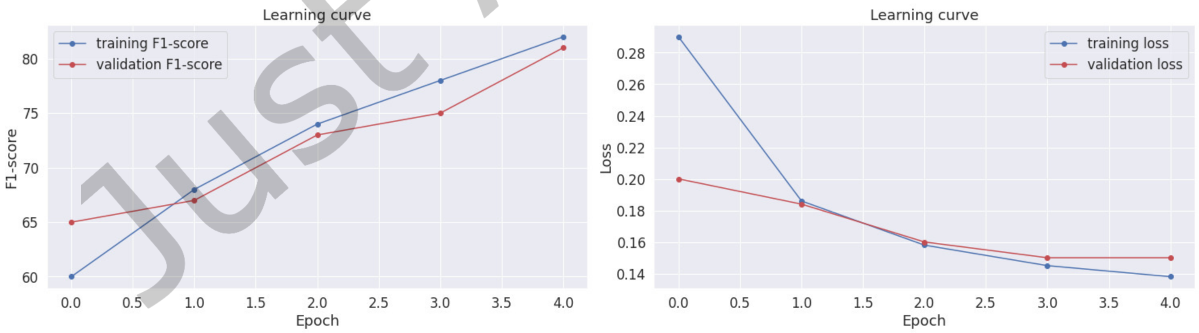


Fig. 10. The training and validation F1-score and loss at each epoch of GigaBERT training for task T3

GigaBERT and AraBERT are performing better than Bi-LSTM-CRF and SVM because they use Transformers. Transformers learns quicker than LSTM models because all input is ingested once [50]. Training LSTMs is more difficult than training transformers because LSTM networks have a much larger number of parameters. Moreover, both AraBERT and GigaBERT are pre-trained on larger amount of data, which results in a better learned language representation [53]. All these factors boost up the performance of BERT-based models, and they also come with less complexity and computational cost, as compared to LSTM and Bi-LSTM-CRF.

BERT-based models has revolutionized many fields in machine learning such as question answering, text grading, sentiment analysis, and aspect-based sentiment analysis. Our findings on Arabic food delivery reviews conform with previous work.

5 CONCLUSION AND FUTURE WORK

This paper proposed a dataset for performing aspect-based sentiment analysis (ABSA) for Arabic food delivery reviews. The dataset contains about 4,880 reviews, with 23,733 aspects, each labeled with a category and polarity. Our experimental results showed that Bert-based models (particularly GigaBERT and AraBERT) have outperformed LSTM, Bi-LSTM-CRF, and SVM.

We intend to continually expand our dataset with more reviews. Also we plan to address the problem of Out-of-domain ABSA, where novel domains of aspects not seen by the model during training emerge during the testing/deployment phase.

REFERENCES

[1] Wissam Antoun, Fady Baly, and Hazem Hajj. [n. d.]. AraBERT: Transformer-based Model for Arabic Language Understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*. 9.

[2] Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006* (1999).

[3] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*. Springer, 799–804.

[4] A Phillip Owens 3rd and Nigel Mackman. 2011. MP's and VTE's: Fact or fiction. *Thrombosis research* 128, 6 (2011), 505–506.

[5] Dedi Rianto Rahadi and Leon Andretti Abdillah. 2013. The utilization of social networking as promotion media (Case study: Handicraft business in Palembang). *arXiv preprint arXiv:1312.3532* (2013).

[6] Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*. 165–173.

[7] Jack Dorsey, Noah Glass, Biz Stone, Evan Williams. 2014. Twitter. https://en.wikipedia.org/wiki/Twitter

[8] Mohammad Al-Smadi, Omar Qawasmeh, Bashar Talafha, and Muhannad Quwaider. 2015. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *2015 3rd International Conference on Future Internet of Things and Cloud*. IEEE, 726–730.

[9] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 486–495.

[10] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems* 89 (2015), 14–46.

[11] Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Huda Al-Sarhan, and Yaser Jararweh. 2016. An Aspect-Based Sentiment Analysis Approach to Evaluating Arabic News Affect on Readers. *J. Univers. Comput. Sci.* 22, 5 (2016), 630–649.

[12] AL-Smadi Mohammad, Omar Qwasmeh, Bashar Talafha, Mahmoud Al-Ayyoub, Yaser Jararweh, and Elhadj Benkhelifa. 2016. An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study. In *2016 11th International conference for internet technology and secured transactions (ICITST)*. IEEE, 98–103.

[13] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679* (2016).

[14] MS Mubarok. 2017. Adiwijaya, and MD Aldhi,“. In *Aspect-based sentiment analysis to review products using Naïve Bayes," in AIP Conference Proceedings*. 020060.

[15] Rajesh Piryani, D Madhavi, and Vivek Kumar Singh. 2017. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management* 53, 1 (2017), 122–150.

[16] Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16, 4 (2017), 1–20.

[17] Mahmoud Al-Ayyoub, Amal Gigieh, Areej Al-Qwaqenah, Mohammed N Al-Kabi, Bashar Talafhah, and Izzat Alsmadi. 2017. Aspect-Based Sentiment Analysis of Arabic Laptop. In *ACIT'2017, The International Arab Conference on Information Technology.*

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[19] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923* (2017).

[20] Ike Pertiwi Windasari, Fajar Nurul Uzzi, and Kodrat Iman Satoto. 2017. Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek. In *2017 4th international conference on information technology, computer, and electrical engineering (ICITACEE)*. IEEE, 266–269.

[21] Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

[22] Sridhar Ramaswamy and Natalie DeClerck. 2018. Customer perception analysis using deep learning and NLP. *Procedia Computer Science* 140 (2018), 170–178.

[23] Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2018. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of computational science* 27 (2018), 386–393.

[24] Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making* 18, 2 (2018), 77–87.

[25] Natalia Klyueva, Yunfei Long, Chu-Ren Huang, and Qin Lu. 2018. Food-Related Sentiment Analysis for Cantonese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 25th Joint Workshop on Linguistics and Language Processing*.

[26] Mohammed Matuq Ashi, Muazzam Ahmed Siddiqui, and Farrukh Nadeem. 2018. Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets. In *International Conference on Advanced Intelligent Systems and Informatics*. Springer, 241–251.

[27] Zulkarnian Zulkarnain, Isti Surjandari, and Reggia Aldiana Wayasti. 2018. Sentiment Analysis for Mining Customer Opinion on Twitter: A Case Study of Ride-Hailing Service Provider. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 512–516.

[28] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2VLDA: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications* 91 (2018), 127–137.

[29] Sana Trigui, Ines Boujelben, Salma Jamoussi, and Yassine Ben Ayed. 2019. ADAL System: Aspect Detection for Arabic Language. In *International Conference on Hybrid Intelligent Systems*. Springer, 31–40.

[30] Nilesh Korde, Gaurav Kawade, Sunita Rawat, Kavita Kalambe, and Abhijeet Thakare. 2019. Exploration of Opinion from Twitter Data. *International Journal of Recent Technology and Engineering (IJRTE)* 8 (2019).

[31] Andia Enggar Mayasari and Anggit Dwi Hartanto. 2019. User Satisfaction Levels Sentiment Analysis Toward Goods Delivery Service On Twitter Using Support Vector Machine Algorithm (SVM). In *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 239–243.

[32] Mohammad Al-Smadi, Bashar Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2019. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics* 10, 8 (2019), 2163–2175.

[33] Norah Fahad Alshammari and Amal Abdullah AlMansour. 2020. Aspect-based Sentiment Analysis for Arabic Content in Social Media. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 1–6.

[34] Sufyan Areed, Omar Alqaryouti, Bilal Siyam, and Khaled Shaalan. 2020. Aspect-based sentiment analysis for Arabic government reviews. In *Recent advances in NLP: the case of Arabic language*. Springer, 143–162.

[35] Manar Alassaf and Ali Mustafa Qamar. 2020. Aspect-Based Sentiment Analysis of Arabic Tweets in the Education Sector Using a Hybrid Feature Selection Method. In *2020 14th International Conference on Innovations in Information Technology (IIT)*. IEEE, 178–185.

[36] Manjubala Bisi, Ankam Divija, Sowmya Namala, and Rohan Sarap. 2020. CNN-BPSO Model for Multi Classification of Tweets. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. IEEE, 1–5.

[37] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv preprint arXiv:2006.09595* (2020).

[38] Irfan Nasrullah and Rila Mandala. 2020. COMPARISON INTENT RECOGNITION ON FOOD DELIVERY SERVICE COMPLAINT IN TWITTER WITH RECURRENT AND CONVOLUTIONAL NEURAL NETWORK. *IT for Society* 5, 1 (2020).

[39] Satyendra Praneel Reddy Karri and B Santhosh Kumar. 2020. Deep learning techniques for implementation of chatbots. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–5.

[40] Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

[41] Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. *arXiv preprint arXiv:2004.14519* (2020).

[42] Ghady Alhamad and Mohamad-Bassam Kurdy. 2020. Feature-Based Sentiment Analysis for Arabic Language. *International Journal of Advanced Computer Science and Applications* 11, 11 (2020).

[43] Abdullah M Bdeir and Farid Ibrahim. 2020. A framework for arabic tweets multi-label classification using word embedding and neural networks algorithms. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering*. 105–112.

[44] Raja Masadeh and Bassam Hammo Sa'ad Al-Azzam. 2020. A Hybrid Approach of Lexicon-based and Corpus-based Techniques for Arabic Book Aspect and Review Polarity Detection. *International Journal of Advanced Trends in Computer Science and Engineering* 9 (2020).

[45] Ravindra Kumar Singh and Harsh Kumar Verma. 2020. Influence of social media analytics on online food delivery systems. *International Journal of Information System Modeling and Design (IJISMD)* 11, 3 (2020), 1–21.

[46] Phivos Mylonas. 2020. SMAP 2020 15th International Workshop on Semantic and Social Media Adaptation & Personalization. In *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*. IEEE, 1–4.

[47] Derek A Pisner and David M Schnyer. 2020. Support vector machine. In *Machine learning*. Elsevier, 101–121.

[48] Salima Behdenna, Fatiha Barigou, and Ghalem Belalem. 2020. Towards semantic aspect-based sentiment analysis for arabic reviews. *International Journal of Information Systems in the Service Sector (IJISSS)* 12, 4 (2020), 1–13.

[49] Youwei Song, Jiahai Wang, Zhiwei Liang, Zhiyue Liu, and Tao Jiang. 2020. Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference. *arXiv preprint arXiv:2002.04815* (2020).

[50] Harshith Nadendla. 2020. Why are LSTMs struggling to matchup with Transformers? https://medium.com/analytics-vidhya/why-are-lstms-struggling-to-matchup-with-transformers-a1cc5b2557e3

[51] Ruba Obiedat, Duha Al-Darras, Esra Alzaghoul, and Osama Harfoushi. 2021. Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review. *IEEE Access* (2021).

[52] Mohammed M Abdelgwad, Taysir Hassan A Soliman, Ahmed I Taloba, and Mohamed Fawzy Farghaly. 2021. Arabic aspect based sentiment analysis using bidirectional GRU based models. *Journal of King Saud University-Computer and Information Sciences* (2021).

[53] Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*. 21–31.

[54] Saja Al-Dabet, Sara Tedmori, and AL-Smadi Mohammad. 2021. Enhancing Arabic aspect-based sentiment analysis using deep learning models. *Computer Speech & Language* 69 (2021), 101224.

[55] Azadeh Mohammadi and Anis Shaverizade. 2021. Ensemble deep learning for aspect-based sentiment analysis. *International Journal of Nonlinear Analysis and Applications* 12, Special Issue (2021), 29–38.

[56] Krutuja S Lasne, Sejal S Nandrekar, Ashraf A Khan, and Tushar Ghorpade. 2021. Food Reviews Classification using multi-label convolutional neural network text classifier. In *ITM Web of Conferences*, Vol. 40. EDP Sciences, 01009.

[57] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2021. Multilingual Review-aware Deep Recommender System via Aspect-based Sentiment Analysis. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–33.

[58] Jahanur Biswas, Md Mahbubur Rahman, Al Amin Biswas, Md Akib Zabed Khan, Aditya Rajbongshi, and Hasnaine Amin Niloy. 2021. Sentiment Analysis on User Reaction for Online Food Delivery Services using BERT Model. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1. IEEE, 1019–1023.

[59] Salima Behdenna, Barigou Fatiha, and Ghalem Belalem. 2022. Ontology-Based Approach to Enhance Explicit Aspect Extraction in Standard Arabic Reviews. *International Journal of Computing and Digital Systems* 11, 1 (2022), 277–287.

[60] Rajae Bensoltane and Taher Zaki. 2022. Towards Arabic aspect-based sentiment analysis: a transfer learning-based approach. *Social Network Analysis and Mining* 12, 1 (2022), 1–16.