SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا
Sana'a University Journal of Applied Sciences and Technology
https://journals.su.edu.ye/index.php/jast/

Lexicon-Based Approach in Sentiment Analysis of Yemeni Dialect for Social Media Network

Alaa Abdulkareem Hameed Brihi[1] **and Mossa Ghurab** [1]

[1]**Department of Computer Science, Faculty of Computer and Information Technology, University of Sana'a, Sana'a, Yemen**

*Corresponding author:alaa.brihi.@su.edu.y

Abstract

Recently, the number of Yemeni users has been expanding quickly on social media platforms. Most research in Arabic sentiment analysis has gained on Modern Standard Arabic (MSA) and some specific dialects, such as Egyptian, Levantine, and Gulf. However, there is a noticeable gap in Yemeni dialect sentiment analysis research. The reason for that is the lack of reliable Yemeni lexical and corpus and a real dataset for social media sentiment analysis. This research addresses this lack by presenting the Yemeni Dialect sentiment lexicon and corpus. This lexicon and corpus provide valuable resources for researchers and practitioners seeking to analyze sentiment in Yemeni dialect social media content, contributing to a better understanding of Yemeni public opinion, social media monitoring, marketing, cultural understanding, and assisting in efforts to respond to crises in Yemen. The Yemeni Dialect sentiment lexicon is enriched with a reasonable number of words and phrases categorized according to their positive and negative sentiment tendencies. Moreover, we constructed a corpus dataset of more than 54,000 comments built from the Facebook platform. A large dataset of unlabeled comments from the main Yemeni telecommunications companies in Yemen (Yemen Telecom, Yemen Mobile, YOU, and Sabafon), are people commenting on a public issue related to the services provided by those companies. The lexicon-based approach is used to extract the sentiment's polarity and label each of the provided comments to formulate a corpus dataset as being either positive, negative, or neutral. The evaluation metrics of experiments are accuracy, recall, precision, f-measure, and the confusion matrix. The accuracy result of the lexicon-based labeling approach was calculated through a comparison between the achieved results and the ones achieved through manually labeled comments by three Yemeni experts. Evaluation results using a lexicon-based approach achieved an accuracy of 90.05%. .

Article Info

Keywords:
Arabic Sentiment Analysis, Yemeni Dialect, Lexicon-Based, Yemeni Lexicon and courps

Article History:
Received: 10-June-2024,
Revised: 17-August-2024,
Accepted: 13-Septermber-2024,
Available online: 31 October 2024.

1. INTRODUCTION

With the increase of Arab users on social media networks expressing their thoughts, socializing, and sharing their comments, opinions, and sentiments, data has grown explosively. Such growth has created a huge amount of unstructured data. The demand for sentiment analysis in Arabic has experienced a significant surge. According to UNESCO in 2023 [1], the Arabic language is one of the most widely spoken languages in the world, used daily by more than 400 million people, making it the fifth most widely spoken language worldwide, following Mandarin, Spanish, English, and Hindi. Sentiment analysis, also known as opinion mining, can simplify reforming the unstructured data and place it within a structured form. This area of study analyzes people's sentiments, opinions, evaluations, emotions, and attitudes in their language to understand their perspective concerning a precise topic or goal as being either positive, negative, or neutral. In recent years, the number of Yemeni users has been expanding quickly on social media platforms. Especially Facebook, YouTube, and Twitter, as they are the most famous platforms on which the standard Arabic language and Arabic dialect are used. According to [2] the Social Media Stats in Yemen in December 2023, 64.22% of Yemeni users use Facebook, 16.43% use YouTube, 12.54% use Twitter, and 5.91% use Instagram. The sentiment analysis of the Yemeni Arabic Dialect domain needs a comprehensive survey to understand Yemeni public opinion, cultural trends, and societal issues. However, research on this domain was rarely found. The development of the SA system for the Yemeni dialect faces many challenges due to the limitations of freely available resources, such as reliable lexicon and corpus and a real dataset for sentiment analysis of social media. Moreover, the absence of standard orthographies and tools dedicated to this dialect.This research aims to fill these gaps by creating a reliable Yemeni corpus and lexicon for sentiment analysis. This will allow us to understand numerous social media users and obtain their opinions and attitudes to provide better services. We constructed a corpus from Facebook, a rich source and the most popular social platform among Yemenis. The resulting Yemeni lexical sentiment corpus provides a valuable resource for training and evaluating future sentiment analysis models, opinion mining models, and natural language processing models specifically tailored to the Yemeni dialect. The accompanying sentiment lexicon serves as a foundational reference map, categorizing words and phrases according to their positive or negative sentiment tendencies. The primary contributions to this paper are as follows: It focuses on Arabic Sentiment Analysis and provides solutions to one of the challenges that face Arabic SA by creating the largest Yemeni dialect sentiment resource. This resource is based on data extracted from Facebook public pages for Yemeni telecom companies. The remainder of the article is structured as follows: Section 2 overviews the related work. In Section 3, the methodology used in this research, we describe our approach and observations while creating the lexicons and corpus. In Section 4, we report on the results of lexicon and corpus validation and discuss them. In Section 5, conclusions are drawn.

2. RELATED WORK

Research in sentiment analysis and building resources for Arabic dialects is increasing. Social media platforms are the most commonly used for building resources and constructing datasets of Arabic dialects. However, the detection of sentiment polarity is a challenging task due to the lack of sentiment resources in Arabic dialects. While a substantial body of research exists for English and other languages [3], it remains largely. Recently, research in Arabic Sentiment Analysis (ASA) has been interested in the various dialects. That is because the majority of Arabic interactions in social media are produced in local dialects. The literature of (ASA) focuses on sentiment analysis on social media platforms such as Facebook and Twitter, which use standard Arabic language and colloquial Arabic. Most of the literature on sentiment analysis is for specific dialects like Egypt, Levantin, and Gulf dialects, while sentiment analysis research in the Yemeni Dialect is limited. Figure1 displays the number of studies for Arab country dialects. Senti-

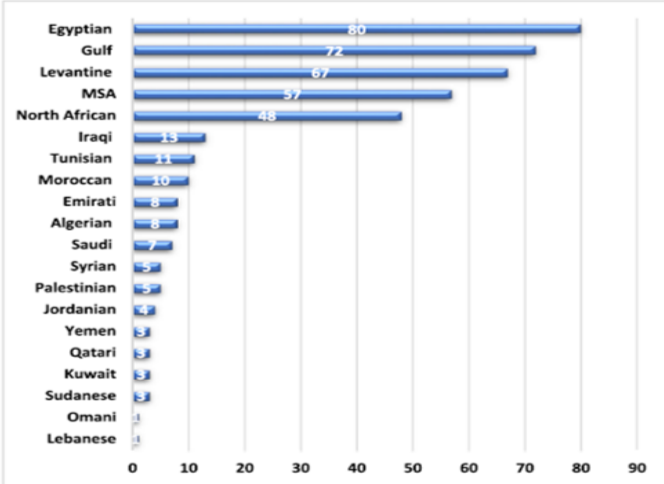

Figure 1. Research per country/regional dialect[4]

ment analysis approaches are divided into lexicon-based, Machine learning (ML), and Hybrid approaches. Much of the lexicon-based research has focused on using adjectives or adjective phrases as the primary source of subjective content in a document. For instance, “good” and “beautiful” were positive sentiments, whereas “bad" and "terrible” were negative-feeling words [5]. Lexicon-based techniques are an unsupervised method that does not need training and high-speed classification but requires large-scale external lexical resources. The accuracy of a result is dependent upon the size and quality of the lexicon [6]. This approach is divided into two techniques: Corpus-Based and Dictionary-based. Lexicon-based techniques fundamentally concentrate on analyzing the sentiment lexicon (i.e., the collation of words where each one contains a mark that indicates the negative, neutral, or positive tone of the text to be explored). For the chosen text information, marks for the subjective words are assessed and inputted separately, and the maximum score will decide the overall polarity. The text is analyzed via this sentiment lexicon. One of the major advantages of the Lexicon-Based Approach is its domain independence and ability to be easily extended and improved. However, it is prohibitively costly in terms of annotator time and effort. In the study of AlTwairesh et al. [7], the deployed AraSenti-Tweet corpus contains 17,573 tweets manually labeled with four sentiment labels: positive, negative, neutral, and mixed. The accuracy of their work registered at 76.31%. Nahar et al. [8] concentrated on the SA of Facebook Arabic comments for Jordanian telecommunications companies. The lexicon-based approach was used to determine the polarity of each of the provided Facebook comments. Data samples come from Jordanians commenting on a public issue related to the services provided by Jordan’s main telecommunications companies. The results of the evaluation of the Arabic sentiment lexicon were promising. They created a large dataset of unlabeled comments, the lexicon was used to label a set of Facebook comments. Then, the resulting labeled dataset frequently used ML algorithms to classify comments in the absence of lexicons. This model is still restricted to the availability of the words or phrases in the lexicons. It is considered unsupervised learning that depends on a mathematical counting formula. Aloqaily et al.[9] developed lexicon-based sentiment analysis for Arabic tweet datasets concerning the Syrian civil war and crisis. Arabic Tweets, expressed as bag-of-words (BOW), are classified as positive and negative by looking up the mentioned sentiments in an Arabic sentiment lexicon. Their work was accurate 68% of the time. Another sentiment lexicon is NileULex [10, 11, 12], which includes compound phrases and single words from dialectal Arabic and MSA. Terms and compound phrases were derived from social media automatically, even though they were manually annotated. Abdul-Mageed and Diab [13] constructed SANA, which is a combination of many lexicons, such as SIFAAT (3,325 Arabic adjectives), HUDA (4,905 entries extracted from chat records in the Egyptian dialect) and an automatically collected corpus (with both statistical method and machine translation). In [14], they proposed a new lexicon-based model for Arabic sentiment analysis with the support of the Vader Module. The model's accuracy was 86.6%. In [6], they developed the lexicon-based analysis of the Saudi dialect. Specifically, a morphologically annotated corpus of the Saudi dialects, consisting of 7000 tweets, was collected. This lexicon is domain-specific and corresponds to the issue of unemployment in Saudi Arabia. Then, we applied multi-factor lexicon-based sentiment analysis. The results indicate that the proposed combined lexicon approach (light stemming, emojis, intensifiers, negations, and special phrases, such as supplications, proverbs, and interjections) obtained an 89.80% accuracy score. In [5], the authors prepared a sentiment analysis dataset gathered from Arabic tweets, called Arabic Sentiment Tweets Dataset (ASTD). ASTD is an Arabic sentiment corpus that contains 10,000 tweets that were manually annotated and classified as positive, negative, mixed, and objective. They annotated the tweets dataset and constructed a seed sentiment lexicon from the dataset. AraSenTi by [3] is about the Saudi dialect in multiple domains, such as education, sports, news, etc. However, their lexicon was based on extracting the lexicon from a set of tweets automatically and then reviewing it manually. These lexicons were extracted from the datasets of tweets using the MADAMIRA tool and contain 131,342 terms. The accuracy of their work registered at 76.31%, In this study [15], a mixed lexicon was used. The lexicon was a combination of "AraSentiLexicon" made by [3] and an Arabic translation of Bing Lius Lexicon. In [16] This work introduces an approach to analyzing the sentiment effects of emoji as textual features. Using an Arabic dataset as a benchmark. The results confirm the borrowed argument that each emoji has three different norms of sentiment role (negative, neutral, or positive) and an emoji can play different sentiment roles depending upon context. NurMaulidiahElfajr et al. [17] concentrated on the emoticon dictionary and the weighting of emoticons. They determined the emotions conveyed in a sentence through the use of emoticons. They assumed that emoticons express emotions more effectively than words. The findings of their investigation revealed that the inclusion of an emoticon-based model significantly improved the results compared to the SentiWordNet process without such a model. In [18] introduces a distant supervision algorithm that automates the collection and labeling of 'TEAD', a dataset for Arabic Sentiment Analysis (SA), by utilizing emojis and sentiment lexicons. The researchers employed an emoji lexicon as search keywords to gather data and addressed the challenge of using dialect instead of MSA. Furthermore, they used an algorithm to replace dialect words with their respective synonyms in the MSA. The lexicons used for translation of the Arabic dialect from Egypt, Levantine, Maghrebi, and Gulf lexicons. Several benchmark experiments were conducted to compare TEAD with ASTD. As our focus is on the Yemeni Arabic dialect, research has been carried out on the Yemeni dialect by [19, 20, 21, 22, 23]. The initial attempt to create an annotated corpus for the Sana'ani dialect was made by [19]. They present annotated morphological corpora resources for each dialect of Moroccan and Sanaani Yemeni Arabic. DIWAN tool [24] was used to morphologically annotate the corpus for each dialect. The YEMS Corpus size is 32.5K word tokens from various sources such as a Sanaani Radio Station program social texts, poems, and political texts. Similarly, [20] conducted by the same authors and using the same corpus size and tool in [19]. However, this study encompasses two Yemeni dialects, Sana'ani and Taizi, along with five other Arabic dialects. Each word in the corpus was annotated with CODA, lemma, morphological information, prefix, stem, and suffix to establish a common ground with Modern Standard Arabic (MSA) and other Arabic dialects. In this study [21], they developed Normalizer courps for San'ani Arabic Social media texts that are extracted from Facebook and Telegram apps, representing daily fictional conversations written throughout the year. Their corpus consists of 447,401 tokens and 51,073 types. The normalizer is limited to dealing with San'ani Arabic spoken in Yemen. Another study by the same author with others [22] presents a grammatically annotated corpus for Sana'ani Arabic adopted from an earlier research presented in [21]. The corpus consists of 7,295 tokenized sentences. The annotation performed is rather a grammatical annotation ignoring morphological inflections. A more recent study on Yemeni corpus was conducted by [23]. Supervised machine learning was applied to a developed and classified MSA and Yemeni dialects dataset using RapidMiner. A constructed dataset was collected from Twitter and Facebook involved in the political domain, consisting of 2000 MSA and Yemeni dialects records used for training and 300 MSA and Yemeni dialects records for testing purposes. The current literature review did not provide a suitable Yemeni Dialect Arabic lexicon that could fulfill the aims of the study.

Due to the lack of freely and publicly available Yemeni dialectal Arabic sentiment lexicons, a new lexicon construction approach is proposed to fill these gaps. A reliable Yemeni corpus and lexicon contains many terms for various domains and topics and can be used for Arabic sentiment analysis to understand numerous users on social media and get their opinions and attitudes to provide better services.

3. METHODOLOGY

To enrich the study's sentiment lexicon and corpora with the Yemeni dialect, we did the following steps :

3.1. Data Collection

Below are the methodologies employed for the collection of data for the Yemeni dialect sentiment lexicon and corpora:

3.1.1. Sentiment Lexicons Collection

Lexicon-based approaches are widely utilized in sentiment analysis research. These approaches consider the semantic orientation of words in a given text and compute sentiment scores accordingly. In this methodology, a lexicon or dictionary consisting of positive and negative terms is constructed, with each word assigned a sentiment value. These sentiment values are then incorporated into the text being analyzed, which is transformed into a collection of words, and subsequently matched with the lexicon. In the realm of lexicon-based approaches, significant attention has been devoted to English sentiment lexicons [25], while Arabic sentiment lexicons have received relatively little focus. Conversely, the majority of these endeavors have concentrated on addressing specific problem statements. There are various factors of Arabic language sentences that pose a challenge. Furthermore, there is a degree of difference between dialects in the same country. This experiment considered different Yemeni dialects, such as Sana'ani, Taizi, Adni, Hadrami, and Tahami. Therefore, the lexicon contains words and phrases from different regions of Yemen, which are manually added to the word list.

The Yemeni Dialect sentiment lexicon was developed to assess the polarity, emoji, and special phrases related to the Yemen dialect. In special phrases, we take into account cases such as (supplications, negations, proverbs, and interjections).

We manually extracted all the sentiment Yemeni dialect words from our collected Facebook datasets. Furthermore, the authors asked some of their friends from different regions in Yemen to give us more Sentiment words and phrases that are used in their area and added to the lexicon. Next, we divided all these words and phrases based on their polarities. Three native-speaking annotators manually classified the words and phrases into positive and negative polarity levels. If there is any disagreement among the three annotators, we solve it by voting. The lexicon comprises 2902 words and phrases that are linked to different emotions (183 negative phrases, 87 positive phrases, 1855 negative words, and 777 positive words). As we see in Table 1, there are examples of lexicon words/ phrases. Moreover, we used 140 negative emojis and 362 positive emojis. Figure 2 displays the percentage of lexicon words/ phrases. The diversity in Yemeni dialects is such that there is a difference in dialect between one village and another and one city and another. Therefore, the same word is added to the lexicon in different slang. Also, some Arabic vocabulary in the Yemeni dialect has connotations that are different from MSA and other dialects. It has a semantic specificity, not a syntactic specificity. Furthermore, there are a lot of loanwords from Turkish or Indian like ستاره perde 'curtain' and ميز maiz 'table'. Furthermore, we added English words written using the Arabic alphabet which is frequently used on social media [26] like ثانكس which means Thanks.

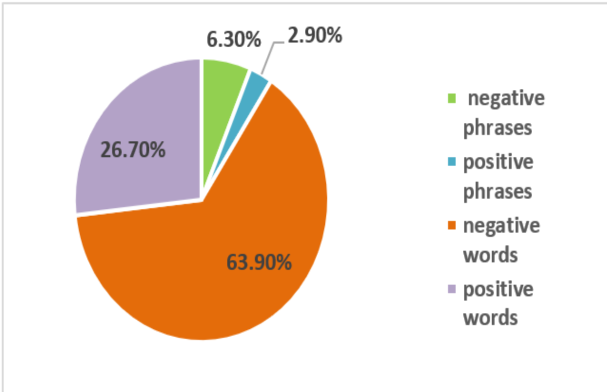

Figure 2. The percentage of lexicon word /phrase

3.1.2. Corpus Dataset Collection

According to information from [2] Facebook ,Twitter, and Instagram are considered popular social networking platforms in Yemen. Facebook is the most popular social media platform, where 64.22% of social media users in Yemen. Therefore, the dataset in this research work was collected from the Facebook platform. We focus here on gathering comments on Facebook written in

Table 1. Examples of lexicon word /phrase

Positive Word	Negative Word	Positive Phrase	Negative Phrase
اوكي	اتجعجع	احر من الجمر	اتقوا الله
جبر	اتفه	اطيب الامنيات	الله يهديكم
حالي	اجحاف	اجحاف	هدره على الفاضي
قرعه	اخترط	الف مبروك	سوق سوداء
يابلاشاه	ادوع	الله معاكم	خافوا الله
توفيق	ازفت	الله يديمكم	الله المستعان
توكل	استحو	اعانكم الله	الله يشغلكم
تيسير	اقطبو	الى الامام	الله يشلكم
ثانكش	التطم	الله يوقفكم	قرحنا جو
جود	ودافة	ان شاء الله	مابش خراج
جياد	انقلع	انت الاصل	الله يقلعكم
مليح	بايخ	باذن الله	عصدتونا عصيد
حصانة	بجاحه	بارك الله	اللي استحوا ماتوا
حياك	بعسسه	يسعد صباحكم	تحت الصفر
خبرات	بغران	يرفع الراس	خارج التغطية
خرافي	بلطجه	تقبل الله	جعلكم السم
خنفشاري	بقبقه	قوه القوه	شلوك الجن
محسنه	تحفه	جزاكم الله خير	ياخزا البلا
فشعة	طنش	جمعه مباركه	حدث ولاحرج

Yemeni dialects and MSA. Python scripts retrieve comments from Facebook, utilizing the Python code library to extract publicly accessible data on the platform. Then, the data is stored in JSON file format. The comments gathered are from the public and formal Facebook pages of the Yemeni telecommunication companies. There are four main Yemeni telecommunication companies: Yemen Telecom, Yemen Mobile, YOU, and Sabafon. The collected comments were to obtain users' opinions regarding services provided by these companies. Table 2 shows examples of collected comments.

3.2. Data Filtering

Initially, we were able to collect around 80,000 comments based on the Yemeni dialect or MSA; Spam comments presented the main challenge, such as advertisements and comments related to events at the time of the data collection. The spam comments constituted the majority of the corpus because at the time the data was collected, it coincided with two sporting events: The World Cup 2022 and the Arabian Gulf Cup 25. Resulting in a lot of comments related to predictions of match results and discussions about match events. Therefore, eliminating all these comments was unrelated to our goal and their effects on accuracy. Subsequently, the dataset size was reduced to 54,163 comments, divided between the companies as we see in Table 3 From our corpus dataset, we found that most users' opinions regarding services were written in non-standard dialectal Arabic. Furthermore, most comments were written without concentration, with improvised words, orthographic mistakes, and slang vocabulary like Sana'ani and Taizi.

Table 2. Examples of collected comments

الباقة حلوه شكرا النت سريع😍
التغطية ضعيفة في رداع
حلوه بس غاليه
كيف نعرف كم المتبقي من رصيد الانترنت
مابش النت خالص وشبكة االتصال تتقطع من وقت لآخر عندنا مابش إهتمام مديرية التحيتاء
مشكورررررين إبداع متواصل
تمام قوه القوه
خلو سعر الباقات والرصيد موحد في جميع أنحاء اليمن وسعرها موحد عشان نكون نعبي رصيد
بطلوا تضحكو على الناس شبكتكم االسوء في الجمهوريه مافي شبكه لالتصال استحواعلى انفسكم
حلوووووو
تحياتي لكم .... وهللا في عونكم
ان شاء الله ..
واحنا متئ كل المناطق المجاوره لنا موجود اال احنا
مميزين 😊
انترنت ضعيف وفورجي فاشل
متميزون كاالعاده وسباقون في المشاركات

Table 3. Yemeni telecommunication companies and the number of comments for each company

Company	# of Unique comments
Yemen Mobile	16319
YOU	14477
Sabafon	11902
Yemen telecom	11465
Total	54163

3.3. Data Preprocessing

In this study, we developed a Python script using the NLTK (Natural Language Toolkit) library to implement preprocessing. Preprocessing includes many subtasks as follows:

- Tokenization :
  It is a first step in preprocessing which involves breaking up the text into a set of words (tokens) separated by white spaces and stored in a vector that can be

dealt with in the next steps of the processing phase [27, 28].

- Cleaning:

  The collected comments contain a lot of noise. Therefore, we cleaned the data from irrelevant content, such as User information, URLs, and mentions. We also removed content that did not affect the meaning, such as diacritics (Tashdid, Fatha, Tanwin Fath, Damma, Tanwin Damm, Kasra, Tanwin Kasr, Sukun) and Arabic and English punctuation marks. In addition, we removed the non-Arabic words and numbers. Elongated words were processed by deleting the repeated letters many times from the words that are usually used for emphasis, such (for example, “رووووعه“ was returned to its original form “روعه“).

  Moreover, stop words were removed, which involved eliminating words that are used to structure language but do not add to its content, such as (هذا ، هو ،هؤلاء ، التي ،الذي).

- Normalization:

  We normalized different Arabic letter forms which were implemented in most of the explored studies and involved replacing the Arabic letters (إ آ أ) with (ا), replacing (ة) with (ه), replacing (ى ئ) with (ي), and replacing (ؤ) with (و). This process increases the accuracy of the analysis because the words are unified in the way they are written.

- Stemming:

  It is a preprocessing technique that reduces inflected words to their stems or root forms. The stemming step in our work is difficult because we are treated with a dialect that is not modern standard Arabic, and there is no standard pattern in the Yemeni dialect. Moreover, the data from social media is written differently. However, by reducing the word to the root form, it misses out on some important morphological information [28]. So we try to use a stemmer algorithm which keeps the meaning of information. [28] Researchers investigated the impact of stemming on sentiment classification and reported that light-stemming methods outperformed root extraction methods. In this study, an Arabic light stemmer is used to enable the removal of prefixes, waw, and suffixes based on the Information Science Research Institute (ISRI) stemming algorithm [28]. ISRIStemmer is a valuable tool in the NLTK (Natural Language Toolkit) library a rule-based stemmer specifically designed for Arabic Language. Arabic Light stemming maintains the meaning of information by deleting just the suffix and prefix terms. For example, the sentence "تضحكو على الناس شبكتكم الاسوء في اليمن والتغطية ضعيفه استحوا على انفسكم" which means" You are laughing at people. Your network is the worst in Yemen and the coverage is weak, shame on you". ISRIStemmer is designed to deal with Arabic roots. The sentence uses a Yemenidialect that may contain non-standard words or abbreviations and contains many suffixes and compound words, such as "تضحكو laugh", which may not be handled perfectly by ISRIStemmer. We might get the following results: The word "تضحكو " is reduced to "ضحك " which means “laugh”, which is a positive word. In our example "تضحكو على الناس” sentence, the meaning for this word in Yemeni slang is the negative phrase “cheating/lying to people”. When dealing with such situations, it’s important first to look up our lexicon (words and phrases). If the words or phrases exist in the lexicon or not. If not, then apply to ISRIStemmer. which means" You are laughing at people. Your network is the worst in Yemen and the coverage is weak, shame on you". ISRIStemmer is designed to deal with Arabic roots. The sentence uses a Yemeni dialect that may contain non-standard words or abbreviations and contains many suffixes and compound words, such as " تضحكو laugh", which may not be handled perfectly by ISRIStemmer. We might get the following results: The word "تضحكو" is reduced to " ضحك" which means “laugh”, which is a positive word. In our example "تضحكو على الناس” sentence, the meaning for this word in Yemeni slang is the negative phrase “cheating/lying to people”. When dealing with such situations, it’s important first to look up our lexicon (words and phrases). If the words or phrases exist in the lexicon or not. If not, then apply to ISRIStemmer.

The pseudo-code of function: The NLP for Corpus pre-processing

---

Input: collected comments
Output: The pre-processed data Begin
For each comment in the dataset
Removal of irrelevant information
Remove URL, remove hashtag
Remove punctuations
Delete the non-Arabic words and numbers
For each sentence in a comment in
pos_phrases OR neg_phrases:
add a sentence to the list of sentences
delete the sentence in the comment
End For
Split the words in the comment
For each word in words:
if word not exists(pos_words OR neg_words):
Remove repeated letters
if word not in Stop Words:
Apply normalisation
Apply Arabic light stemmer
End if

Add the word to the list of words
End For
Display the result list of words + list of sentences
End

3.4. LEXICON-BASED APPROACH

A lexicon-based approach that matches the separated words and phrases with positive and negative words and phrases vocabulary. Furthermore, If a comment contains emojis, then it matches the positive and negative emojis list. Last, calculate the label sentiment of the comment by counting the number of positive and negative words, phrases, and emojis. The pseudocode of the function comment Score Calculation is as follows:

Pseudocode: comment Score Calculation

Input: The Pre-processed data, lexicons
Output: Sentiment labels like positive, negative, neutral Begin
Set Score ← 0
FOR EACH sentence in the list of sentences DO
IF sentence is PositivePhraseLexicon THEN
Score ← Score +1.0
ELSEIF sentence is NegativePhraseLexicon THEN
Score ← Score - 1.0
ENDIF
END FOR
FOR EACH word in the list of words DO
IF word is PositiveWordLexicon THEN
Score ← Score + 1.0
ELSEIF word is NegativeWordLexicon THEN
Score ← Score - 1.0
ELSEIF word is PositiveEmojiLexicon THEN
Score ← Score + 1.0
ELSEIF word is NegativeEmojiLexicon THEN
Score ← Score - 1.0
ENDIF
IF Score > 0 THEN
Label ← Positive comment
ELSEIF Score < 0 THEN
Label ← Negative comment
ELSE Label ← Neutral comment
ENDIF
RETURN Label

4. RESULTS AND DISCUSSION

This paper has addressed Yemeni dialect sentiment analysis in Facebook comments. To identify the polarity of the provided text, we implemented a lexicon-based approach. We developed a Yemeni dialect lexicon that depended on sentiment words and phrases and was divided into two types: Positive and negative Yemeni words and phrases. The dataset contains Yemeni comments on a public issue related to the services provided by the main telecommunication companies in Yemen. The procedure involves applying a Yemeni dialect lexicon based on the collected dataset, which resulted in a (49%) positive, (16.5%) negative, and (34.5%) neutral. As we show in Figure 3. Table 4 contains examples of comments with their labels.

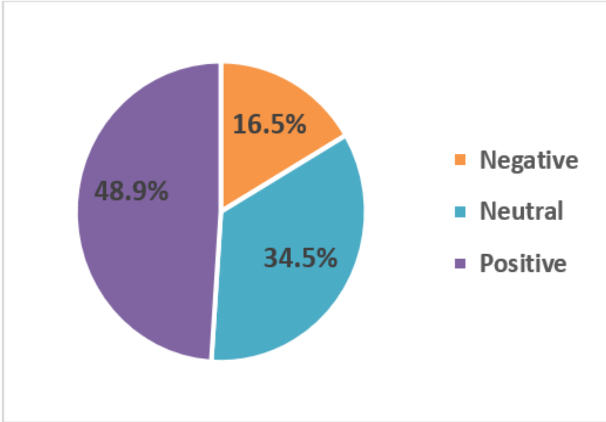

Figure 3. Distribution of polarity scores for the lexicon-based approach

Table 4. Examples of comments with their labels

Label	Comment
Positive	الباقة حلوه شكرا النت سريع 😊 The package is nice, thanks. The internet is fast
Negative	انترنت ضعيف وفورجي فاشل Weak internet and failed 4G
Neutral	هل موجود هذا العرض بصنعاء فقط او بكل المحافظات Is this offer available in Sanaa only or in all governorates?

To evaluate the lexicon's accuracy after getting polarity for each comment in the dataset, we manually annotated around 25,184 comments and categorized them as positive, negative, or neutral for the same courps dataset. Each comment was annotated by the three Yemeni native speakers. A basic Python annotation bot on Telegram was developed using the python-telegram-bot library. The annotator is prompted to choose between three labels - "positive," "negative," or "neutral" - for each comment. The sentiment of the comment was validated by establishing the sentiment that most annotators concurred on. If there was a conflict between the annotators, another expert annotator was assigned. The accuracy of the lexicon-based labeling approach was calculated through a comparison between the achieved results and the ones achieved through manually labeled comments by experts. Four evaluation metrics were utilized in this paper to evaluate the lexicon-based labeling approach. They are precision (P), recall (R), F measure (F), and accuracy (Acc), and their mathematical equations are as follows:

Precision (P) = TP / (TP+ FP)
Recall (R) = TP / (TP + FN)
Accuracy (Acc) = (TP + TN) / (TP + FP +TN +FN)
F-Score (F) = 2TP / (2TP + FP + FN)

Where TP, or True Positive indicates to number of comments that are correctly predicted as a positive, TN, or True Negative are number of comments that are correctly predicted as a negative, FP, or False Positive indicates to number of comments that are incorrectly predicted as a positive, FN, or False Negative is the number of comments that are incorrectly predicted as negative. The results for these measurements are accuracy(Acc) (90.05%), F-Score(F) (90.04%), Precision (P) (90.08%), and Recall (R) (90.05%). Table 5 and Figure 4 compare the performance of the study-proposed sentiment analysis approach against that of Aloqaily et al. [9], Al-Twairesh et al. [7], and Alwakaid[6], who used their experiments' corpora.

Table 5. Demonstrates the comparison of the results of existing work with the proposed work

Author	Dialect	Accuracy (%)
Aloqaily et al. (2020)	MSA& Leventain	68%
AlTwairesh et al.(2017)	MSA & Saudi	76.31%
Alwakaid (2020)	MSA & Saudi	89.80%
Current work	MSA & Yemeni	90.05%

The confusion matrix provides a visual representation of how well the model performs in categorizing instances into various classes. It shows the number of instances that were classified correctly, or incorrectly, and where the model may face challenges. Figure 5 shows the utilization of the confusion matrix in analyzing the sentiment of Yemeni Dialect comments on social media revealing its effectiveness in categorizing sentiments as positive, neutral, or negative. The model successfully identified 14064 positive sentiments, demonstrating its capability to recognize favorable comments regarding telecommunication services. Nonetheless, it incorrectly labeled 449 neutral comments as positive and 259 negative remarks as positive, suggesting challenges in distinguishing between neutral and negative sentiments from positive ones in user comments. The matrix further reveals specific inconsistencies between the model's predictions and the actual sentiments. It indicates that 102 negative sentiments were misclassified as positive and 454 as neutral. Moreover, 697 neutral instances were erroneously categorized as positive, while 541 negative instances were incorrectly classified as neutral. This suggests a need for enhancement in the identification of neutral comments. Naturally, a lexicon-based approach should produce results that are comparable to a qualitative evaluation by a human. Nevertheless, there were a lot of comments that were conflicting about polarity sentiment between the annotators. Humans seldom reach a consensus when it comes to determining the emotional content of a word, sentence, or paragraph. It is hardly surprising that the output mirrors the confusion in the data. Only when there are distinct categories and accurate annotations without any interference can we achieve high levels of accuracy in classification tasks [29]. However, the accuracy of the Yemeni dialect sentiment analysis can be influenced by different aspects, such as the complexity of the Yemeni dialect, the diversity of sentiments expressed, and the

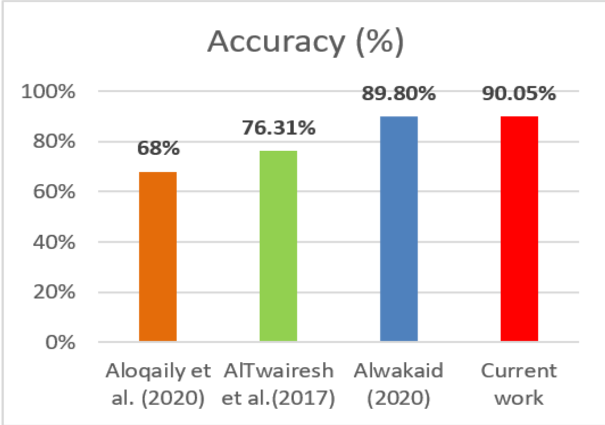

Figure 4. Graphical representation of the accuracy result between the previous lexicon and our lexicon

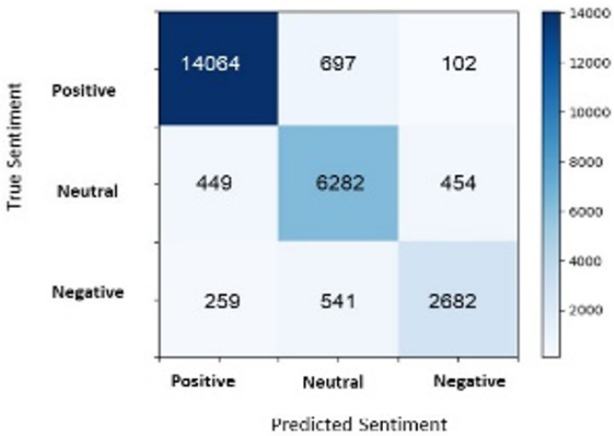

Figure 5. Confusion Matrix for Sentiment analysis of Yemeni Dialect

context in which the analysis is being performed. The lexicon-based approach frequently faces challenges in discerning sentiment within context, potentially resulting in inaccurate categorizations. It is regarded as unsupervised learning, relying on a mathematical counting formula. The lexicon may not encompass all Yemeni dialect vocabulary, which could result in misclassifications. Additionally, the Yemeni dialect exhibits notable regional differences that may not be adequately represented in the lexicon. Hence, it is imperative to develop a more comprehensive Yemeni dialect sentiment lexicon, encompassing regional variations, incorporating nuanced expressions of sentiment, and addressing potential biases to achieve higher accuracy values.

5. CONCLUSION

This paper has introduced Yemeni lexical sentiment and the corpus offers a valuable resource for training and evaluating future sentiment analysis models tailored, opinion mining, and natural language processing specifically for the Yemeni dialect. This resource enables the comprehension of Yemeni public opinion, cultural trends, and societal issues. The study applied a novel lexicon-based sentiment analysis of social media content in Yemeni dialect social media. This approach integrates the processing of several factors, such as special phrases and emojis, to improve classification accuracy. Yemeni telecom companies' services were used as the target problem domain. Also, this methodology applies effective preprocessing steps with a light stemming approach. The accuracy of the lexicon-based labeling approach was calculated by comparing the results achieved and 25,184 manually annotated comments by experts. The study-proposed result was compared to [9], [7], and [6] studies that utilized the same approach with their Arabic dialect experiments' corpora. According to Table 3, the results indicate that the Yemeni lexicon combined lexicon approach (light stemming, emojis, and special phrases, such as (supplications, negation, proverbs, and interjections) achieved an accuracy of 90.05%, surpassing the accuracy of the other studies. While the lexicon shows promise, further refinement is needed to address challenges related to dialectal variations and context-dependent sentiment expressions. Future work will focus on expanding the Yemeni dialect sentiment lexicon, encompassing regional variations. Moreover, The highest consideration for future work is to explore the Yemeni dialect courps with machine learning approaches and hybrid approaches that combine lexicon-based methods with machine learning techniques to leverage the strengths of both approaches.

REFERENCES

[1] Abubakar Yusuf Abdullahi, Isah Zubairu Achara, and Mohammad Usman. "Arabic as an International Language for Renaissance: Impact on the Muslim Ummah". In: (). URL: https://doi.org/10.11648/j.allc.20230801.1%201.

[2] Salah Naji Taher Sanad, NA AL-Shameri, and Shaima Saleh Mohammed Al-Radai. "MOVING TOWARDS DIGITALIZATION: UNVEILING CHALLENGES AND PROSPECTS OF EMARKETING IN LEAST DEVELOPED ECONOMIES–THE CASE OF YEMEN". In: (2023). URL: https://doi.org/10.29121/granthaalayah.v1%201.i8.2023%20.5273.

[3] Nora Al-Twairesh, Hend Al-Khalifa, and AbdulMalik Al-Salman. "AraSenTi: Large-scale Twitter-specific Arabic sentiment lexicons". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 697–705. URL: https://doi.org/10.18653/v1/P16-1066.

[4] Ashraf Elnagar et al. "Systematic literature review of dialectal Arabic: identification and detection". In: *IEEE Access* 9 (2021), pp. 31010–31042. URL: https://doi.org/10.1109/ACCESS.2021.3059504.

[5] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. "Astd: Arabic sentiment tweets dataset". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 2515–2519. URL: https://doi.org/10.18653/v1/d15-1299.

[6] Ghadah Alwakid. *Sentiment analysis of dialectical Arabic social media content using a hybrid linguistic-machine learning approach*. Nottingham Trent University (United Kingdom), 2020. URL: https://irep.ntu.ac.uk/id/eprint/42474.

[7] Nora Al-Twairesh et al. "Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets". In: *Procedia Comput. Sci.* 117 (2017), pp. 63–72. URL: https://doi.org/10.1016/j.procs.2017.10.094.

[8] Khalid MO Nahar et al. "Sentiment analysis and classification of arab jordanian facebook comments for jordanian telecom companies using lexicon-based approach and machine learning". In: *Jordanian J. Comput. Inf. Technol.* 6.3 (2020). URL: https://doi.org/10.5455/jjcit.71-.

[9] Ahmad Aloqaily et al. "Sentiment analysis for arabic tweets datasets: Lexicon-based and machine learning approaches". In: *J. Theor. Appl. Inf. Technol* 98.4 (2020), pp. 612–623. URL: https://api.semanticscholar.org/CorpusID:218979351.

[10] Samhaa R El-Beltagy. "Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2900–2905. URL: https://aclanthology.org/L16-1463.

[11] Samhaa R El-Beltagy and Ahmed Ali. "Open issues in the sentiment analysis of Arabic social media: A case study". In: *2013 9th International Conference on Innovations in information technology (IIT)*. IEEE. 2013, pp. 215–220. URL: https://doi.org/10.1109/Innovations.2013.6544421.

[12] Samhaa R El-Beltagy et al. "Combining lexical features and a supervised learning approach for Arabic sentiment analysis". In: *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*. Springer. 2018, pp. 307–319. URL: https://doi.org/10.48550/arXiv.1710.08451.

[13] Muhammad Abdul-Mageed and Mona T Diab. "Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis." In: *LREC*. 2014, pp. 1162–1169. URL: http://www.lrec%20conf.org/proceedings/lrec2014/pdf/919_Paper.pdf.

[14] Mohamed A Rahim et al. "Sentiment Analysis for colloquial Arabic Language". In: 1.3 (2019), pp. 7–11. URL: https://fcihib.journals.ekb.eg/article_107519_142f35973da506eb32a8191138823e2d.pdf.

[15] Abdulrahman Alruban et al. "Improving sentiment analysis of arabic tweets". In: *International Symposium on Security in Computing and Communication*. Springer. 2019, pp. 146–158. URL: https://doi.org/10.1007/978-981-15-4825-3_12..

[16] Shatha Ali A Hakami, Robert J Hendley, and Phillip Smith. "Emoji sentiment roles for sentiment analysis: A case study in arabic texts". In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. 2022, pp. 346–355. URL: https://doi.org/10.18653/v1/2022.wanlp-1.32.

[17] Nur Maulidiah Elfajr and Riyananto Sarno. "Sentiment analysis using weighted emoticons and SentiWordNet for Indonesian language". In: *2018 International Seminar on Application for Technology of Information and Communication*. IEEE. 2018, pp. 234–238. URL: https://doi.org/10.1109/ISEMANTIC.2018.8549703.

[18] Houssem Abdellaoui and Mounir Zrigui. "Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis". In: *Comput. y Sistemas* 22.3 (2018), pp. 777–786. URL: https://doi.org/10.13053/cys-22-3-3031.

[19] Faisal Al-Shargi et al. "Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic". In: *10th Language Resources and Evaluation Conference (LREC 2016)*. 2016. URL: https://aclanthology.org/L16-1207.

[20] Faisal Alshargi et al. "Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, iraqi and Moroccan". In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. 2019, pp. 137–147. URL: http://doi.org/10.18653/v1/W19-4615.

[21] Mohammed Sharaf Addin. "Developing a Normalizer for San'ani Arabic Social Media Texts". In: (). URL: http://www.researchpublish.com/journal/IJIRI/%20Issue-2-April-2019-June-2019/15.

[22] Sabah Al-Shehabi and Mohammed Sharaf Addin. "A Grammatically Annotated Corpus for Sana'ani Arabic Dialect". In: *Test Eng. Manag.* 83 (Mar. 2020), pp. 4953–4961. URL: https://www.researchgate.net/publication/340298682_A_Grammatically_Annotated_Corpus_for_Sana'ani_Arabic_Dialect.

[23] Emran Al-Buraihy et al. "An Ml-based classification scheme for analyzing the social network reviews of yemeni people." In: *Int. Arab. J. Inf. Technol.* 19.6 (2022), pp. 904–914. URL: https://iajit.org/portal/images/Year2022/No.6/19506.pdf.

[24] Faisal Al-Shargi and Owen Rambow. "Diwan: A dialectal word annotation tool for Arabic". In: *Proceedings of the Second Workshop on Arabic Natural Language Processing*. 2015, pp. 49–58. URL: http://doi.org/10.18653/v1/W15-3206.

[25] Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. "Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases". In: *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*. 2016, pp. 42–51. URL: https://doi.org/10.18653/v1/S16-1004.

[26] J Lana, M Renad, et al. "The Frequency of the English Language Used in Social Media by Undergraduate English Majors in Jordan". In: *World J. Engl. Lang.* 12.6 (2022), pp. 352–352. URL: https://doi.org/10.5430/wjel.v12n6p352.

[27] Zainab A Khalaf and Zainab M Jawad. "Measuring the Impact of Using Different Tools on Classification System Results". In: *Journal of Physics: Conference Series*. Vol. 1591. 1. IOP Publishing. 2020, p. 012025. URL: https://doi.org/%2010.1088/1742-6596/1591/1/012025..

[28] Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. "Arabic stemming without a root dictionary". In: *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*. Vol. 1. IEEE. 2005, pp. 152–157. URL: http://doi.org/10.1109/ITCC.2005.90.

[29] Emily Öhman. "The validity of lexicon-based sentiment analysis in interdisciplinary research". In: *Proceedings of the workshop on natural language processing for digital humanities*. 2021, pp. 7–12. URL: https://aclanthology.org/2021.nlp4dh-1.2.