Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Evaluating the Effectiveness of Implementing the SauDiSenti Lexicon in Saudi Dialect Sentiment Analysis

Haya Albader[1], Nora Almezeini[1*]
[1]Department of Information Management Systems, King Saud University, Riyadh, Saudi Arabia
Corresponding Author: Nora Almezeini, e-mail: nalmezeini@ksu.edu.sa

ARTICLE INFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Introduction: One of the most important methods for identifying an individual's general emotional state or opinion toward a particular topic is to do a sentiment analysis.

Objectives: This study evaluates the effectiveness of the SauDiSenti lexicon in conducting sentiment analysis of Saudi dialect tweets. SauDiSenti is the only one that is available publicly for the Saudi dialect to look into the issue of insufficient linguistic resources for conducting Arabic sentiment analysis.

Methods: A total of 27,000 tweets were collected and preprocessed from the discussions around the STC Pay platform and then analyzed through manual annotation, lexicon-based, and machine-learning approaches. To evaluate the performance of these methods Metrics such as accuracy, precision, recall, and F1 score were computed.

Results: The lexicon-based approach achieved an overall accuracy of 92% and F1 scores of 89.7% and 83.8% for positive and negative sentiments, respectively. While the annotation-based approach provided more accurate sentiment classification with approximately 50.27% of neutral tweets, 31.13% of positive, and 18.6% of negative tweets. Furthermore, the Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR) classifiers were developed and trained on the documents created by each of the methods. Compared with the rest, SVM outperformed the other two classifiers yielding 93% and 96% accuracy with the lexicon-based corpus and annotation-based model.

Conclusions: These results show that even though the SauDiSenti lexicon is compact, it still serves as a reliable tool for analysis across various topics. There exists a trade-off where manual annotation guarantees specialized precision, but the lexicon offers a practical alternative for domain-specific projects. The study recommends expanding the lexicon's size and diversity to its applicability for broader datasets.

Keywords: Sentiment analysis, Saudi dialect, Annotation-based, lexicon-based, STC Pay, machine learning.

INTRODUCTION

The implementation of sentiment analysis for opinion measurement in different fields has been enhanced in the last decade. This has resulted from the occurrence of global social media platforms, also known as opinion mining. It identifies positive and negative sentiments and attitudes regarding specific topics from internet data using methods such as natural language processing, machine learning, and artificial intelligence. Sentiment analysis refers to the extent to which an individual has a positive or negative attitude toward a particular subject matter or even a person's overall thematic orientation at the document, sentence, topic, and other related levels (Rahab, Haouassi, & Laouid, 2022; Al-Ajlan & Alshareef, 2023).

The polarity of a particular text can be identified using one of three sentiment analysis approaches. The first was an annotation-based approach that refers to the methodical process of adding pertinent contextual information to a dataset to facilitate data-mining operations (Leech, 1993). Hinze, Heese, Luczak-Rösch, and Paschke (2012) described annotation as the process of labeling a text, sentence, or word using predefined classes. There are two approaches to annotation: manual and automatic approaches. Manual annotation relies on human labor and is more accurate, whereas automatic annotation utilizes an annotation tool (Almuqren & Cristea, 2021).

The second approach involves utilizing an existing lexicon comprising terms with either positive or negative meanings. Moreover, neutral terms and their corresponding polarities can be incorporated into this lexicon. In this method, the process of ascertaining the polarity of a given text entails an examination of each word within the text and its comparison with the terms included in the lexicon. Subsequently, the cumulative scores of the words that were matched between the text and lexicon were computed. Typically, lexicons are constructed either manually or automatically (Al-Thubaity, Alqahtani, & Aljandal, 2018). Manual lexicon construction makes use of linguistic tools (Abdul-Mageed & Diab, 2014; Al-Twairesh, Al-Khalifa, & Al-Salman, 2016), whereas automatic lexicon construction uses statistical techniques such as pointwise mutual information (PMI) (El-Beltagy & Ali, 2013; Mohammad, Salameh, & Kiritchenko, 2016).

The third strategy is the machine learning approach, where machine learning classifiers are qualified using subjectively labeled datasets known as corpora to construct the model for classifying the testing dataset. However, this strategy has two major flaws. The first is the difficulty of generating large enough data sets worthy of use in developing machine learning algorithms. This is a very common problem in the development of useful linguistic resources. The second problem for these sets is that the fields of application considerably influence these data sets. While a trained model may perform well in one domain, it may significantly underperform in another domain. This is a well-known problem in tweet sentiment analysis because subjects are constantly changing as users' interests shift over time, making it challenging to maintain high classification accuracy (Bifet & Frank, 2010; Refaee & Rieser, 2014).

The mechanisms of polarity detection of a given text can be accomplished using a combination of a machine-learning approach and a lexicon-based approach. This approach may be supplemented by the speed and stability of the lexicon approach and the much higher accuracy of Machine learning algorithms (Alruily, 2021).

At present, most of the literature has predominantly focused on English. However, it is crucial to emphasize that other languages, notably Arabic, are rapidly gaining significance in the context of social network communities. Arabic is positioned as the fourth most extensively used language on the Internet, with a user population of over 400 million. The Arabic language is officially recognized in 22 countries (Sherif et al., 2023). The challenge of performing Arabic sentiment analysis persists mostly due to the limited availability of Arabic sentiment resources, particularly when compared to other languages such as English. Another reason for variation in Arabic language usage among Arab nations is the incorporation of dialectical Arabic vocabulary alongside modern standard Arabic (MSA) (Almuqren & Cristea, 2021; Alali, Mohd Sharef, Azmi Murad, Hamdan, & Husin, 2022; Laifa & Mohdeb, 2023).

Twitter is the prevailing social media site in Saudi Arabia, boasting a user base of more than 15 million individuals as of 2023 (Statista, 2023). In contrast to other Arabic dialects, the availability of Saudi dialect corpora and lexical resources is limited despite having a large user base. This scarcity is notably evident from the lack of freely accessible datasets (Almuqren & Cristea, 2021). The creation of a sentiment vocabulary called SauDiSenti was undertaken by Al-Thubaity, Alqahtani, and Aljandal (2018) with the express purpose of making the sentiment analysis of tweets produced in the Saudi dialect easier. The lexicon consists of a comprehensive collection of 4,431 lexical items, which include both words and phrases derived from modern standard Arabic (MSA) and Saudi dialects. These lexical units were extracted through manual efforts from a larger set of tweets obtained from popular hashtags in Saudi Arabia. However, the dataset underwent annotation procedures before extraction.

Currently, it is understood that SauDiSenti is the only publicly available digital lexicon expressly created for sentiment analysis of Twitter data in the Saudi dialect. In addition, the lexicon creators assessed the efficacy of the lexicon using a limited dataset that included 1,500 randomly selected tweets from a pole of tweets encompassing both the Saudi dialect and MSA. None of the previous studies have used this unique lexicon in the context of sentiment analysis within a specific field. Also, it remains uncertain whether this lexicon can be deemed reliable for analyzing Saudi dialects across several domains and a substantial amount of data. Therefore, the main objective of the study is to evaluate the effectiveness of applying the SauDiSenti lexicon to a randomly chosen topic comprising a sizable number of tweets written in the Saudi dialect to conduct sentiment analysis. To address this, the study aims to: (1)

evaluate the accuracy and reliability of the SauDiSenti lexicon in sentiment classification, (2) compare the performance of the lexicon-based approach with manual annotation, and (3) assess the effectiveness of machine learning classifiers (Naïve Bayes, Support Vector Machine, and Logistic Regression) trained on both lexicon-based and manually annotated corpora. This study would help assess the suitability of employing this lexicon across diverse topics that Saudi people trade on Twitter.

LITERATURE REVIEW

Saudi dialect sentiment analysis lacks appropriate resources, including lexicons and corpora. However, not much research has been done in this field. For example, Al-Twairesh, Al-Khalifa, and Al-Salman (2016) built AraSenTi, an Arabic sentiment analysis lexicon including the modern standard Arabic (MSA), the Saudi dialects, and other Arabic dialects. This lexicon is freely accessible online and is said to be the largest of its kind. The AraSenTi dataset compiled in the present study was based on an automated process and comprised 225,329 entries. The study's outcome showed that the use of AraSenTi in sentiment analysis of tweets was more positive than the other available Arabic dialect dictionaries, despite its efficiency and compactness, AraSenTi is not personalized for the Saudi dialect; instead, it includes several other Arabic dialects.

Al-Thubaity, Alqahtani, and Aljandal (2018) created SauDiSenti, a sentiment vocabulary specifically designed to assess the sentiment of tweets produced in the Saudi dialect. It comprises 4,431 lexical units, encompassing both words and phrases derived from modern standard Arabic (MSA), as well as several Saudi dialects (Al-Thubaity, Alqahtani, & Aljandal, 2018). SauDiSenti was precisely extracted by manual efforts in Saudi Arabia by employing a dataset of tweets taken from current hashtags. Additionally, the dataset was annotated formerly to ensure its reliability and accuracy. The results showed that when associated with the larger autonomously created lexicon, AraSenTi and SauDiSenti established promising results in sentiment analysis for Saudi dialect tweets, despite their smaller size.

Khoo and Johnkhan (2018) presented the WKWSCI Sentiment Lexicon and evaluated its efficiency against five established lexicons for sentiment classification. They found that Hu & Liu Opinion Lexicon performs better in product reviews while WKWSCI Sentiment Lexiconis more effective for news headlines sentiments, yielding an accuracy rate of 69%. This study suggested that the WKWSCI lexicon can be used for non-review texts and Hu & Liu Lexicon for review product texts.

Regarding creating Saudi dialect corpora, several studies focused on creating corpora of Saudi dialect tweets. For example, Al-Twairesh, Al-Khalifa, Al-Salman, and Al-Ohali (2017) created a large corpus of Arabic tweets written mostly in the Saudi dialect and MSA that was named AraSenTi-Tweet. The corpus encountered manual annotation to determine sentiments and comprised 17,573 tweets. These tweets were categorized into four different sentiment categories: mixed, neutral, negative, and positive. The annotation method was comprehensively elucidated and accompanied by a thorough discussion of the challenges experienced during the annotation process (Al-Twairesh, Al-Khalifa, Al-Salman, & Al-Ohali, 2017).

Al-Twairesh, Al-Khalifa, Al-Salman, and Al-Ohali (2017) created a thoroughly annotated corpus named SUAR for the Saudi dialect. This corpus was compiled from several online sources and encompasses approximately 104,079 words. MADAMIRA tool was used for the automatic annotation of the corpus, followed by manual inspection to confirm the analysis. The study conducted a comparative analysis of the linguistic characteristics exhibited in the Saudi dialect corresponding to modern standard Arabic (MSA) and many other Arabic dialects.

Furthermore, GSC AraCust known as the largest gold-standard corpus of Saudi tweets, was created for Arabic sentiment analysis. This corpus comprised 20,000 Saudi tweets. Furthermore, Elgibreen et al. (2021) produced the largest and newest Saudi dialect corpus, the King Saud University Saudi Corpus, with over 1 billion words and over 161 million sentences that covered 26 domains.

Abdulla, Ahmed, Shehab, and Al-Ayyoub (2013) used corpus-based and lexicon-based approaches to conduct sentiment analysis by manually annotated datasets due to the limitation of publicly accessible Arabic datasets and lexicon for sentiment analysis. Their findings demonstrate that corpus-based tools employing support vector machine (SVM) for the classification of light-stemmed datasets perform with high accuracy. Moreover, it also reveals that the performance of the lexicon-based tool improves as the size of the lexicon increases.

Assiri, Emam, and Al-Dossari (2018) presented a novel approach based on lexicons that applied to other fields. This tool was developed to conduct sentiment analysis of the Saudi dialect. Initially, researchers created a large dataset of annotated samples from the Saudi dialect. Subsequently, they developed an extensive lexicon specifically tailored to the Saudi dialect. An algorithm was then created using a lexicon incorporating weighting. The developed algorithm aimed to extract the relationships between words that express polarity and those that do not in a given dataset (Assiri, Emam, & Al-Dossari). Subsequently, it assigns weights to the words according to the identified links. Several tests were carried on for the evaluation of the algorithm's efficacy.

Similarly, Al-Ayyoub, Essa, and Alsmadi (2015) adopted lexicon-based approach to sentiment analysis and constructed sentiment lexicon for approximately 120,000 Arabic terms and designed sentiment analysis tool based on predicate calculus. Their results exhibited superior predictive accuracy compared to the keyword-based method. In addition, they also highlight the need to address challenges associated to dialects and their particular areas.

Alqarafi, Adeel, Hawalah, Swingler, and Hussain (2018) presented a semi-supervised methodology to create a sentiment dataset annotated exclusively for tweets published in the Saudi dialect. Word embedding methods, such as word2vec, have been utilized in this method. The proposed protocol annotated a substantial corpus from the Twitter platform. They employed classification methods to conduct a corpus validation and evaluate their effectiveness.

Moreover, Alahmary, Al-Dossari, and Emam (2019) introduced the application of deep learning techniques for sentiment categorization in texts written in the Saudi dialect. The impetus for this study arises from the notable success achieved by deep learning methods in accurately performing sentiment analysis in several languages, including English, Thai, Persian, and Tamil. The researchers used two deep learning techniques, long short-term memory (LSTM) and bidirectional long-short-term memory (Bi-LSTM), to analyze sentiment on a dataset including 32,063 tweets. Data sentiments were classified using the support vector machine (SVM) algorithm to facilitate comparison. The study showed that deep learning techniques performed better than the SVM strategy (Alahmary, Al-Dossari, & Emam, 2019).

Catelli, Pelosi, and Esposito (2022) performed sentiment analysis for a dataset in the Italian language while comparing the performance of two methods such as lexicon-based and model based on deep neural networks (BERTbase). This study emphasizes the potential strengths and drawbacks of these two approaches to linguistic framework and composition of particular terms. They observed that the lexicon-based approach is more appropriate for small datasets with limited computational power, however they show slightly reduced performance. Conversely, a language-based approach provides more favorable outcomes, specifically addressing unresolved issues like identifying mixed sentiments within the same text. Consequently, this study highlights that deep learning models outperform the lexicon-based model.

METHODS

The SauDiSenti lexicon was evaluated through a two-phase process, employing three sentiment analysis approaches. In the first phase, the selected dataset underwent analysis using two methods: manual annotation and a lexicon-based approach. Subsequently, the results obtained from the lexicon-based approach were compared against the outcomes of the manual annotation.

In the second phase, the corpora generated during the initial phase were utilized for training three machine learning classifiers: Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LR), which were evaluated using key metrics such as accuracy, recall, precision, and F1-score of classifiers. The experiments were implemented using Python scripts.

Data Collection

According to Saudi Vision 2030, digital payment solutions are emerging and expanding rapidly in Saudi Arabia because of the exponential rise in technologies and infrastructure (Al-Razgan et al., 2021). The Saudi Arabia digital payment platform STC Pay continues to attract seven million users who trust its services. (Almuhammadi, 2020; Analysys mason, 2023). Therefore, the present study aimed to examine a dataset of tweets in the Saudi dialect about customer satisfaction with the STC Pay application. Thus, the dataset was developed by retrieving tweets from Twitter. Initially, fifty thousand tweets from Twitter between October 7, 2019, to December 26, 2021. These tweets gathered were about STC Pay in Saudi Arabia and contained the keyword "@stcpay_ksa". However, the process of annotations exhibited several issues within retrieved tweets. They contain a large number of duplicate tweets which may be due to retweeting. Subsequently, retweets and tweets encompassing non-Arabic words were all removed from the datasets. Moreover, tweets that were not related to user experience with the STC Pay application were also removed. Therefore, approximately 27,000 tweets were retained in the dataset.

Data Pre-processing

The following actions were performed on each tweet in the dataset using Python code: First, all repeated characters, hashtags, punctuation, @usernames, and "HTTP" and "WWW" links were eliminated. This step minimizes the noise in the dataset as they do not contribute to sentiment-related information and may disrupt text processing. The most common Arabic stopwords were then eliminated, including "لقد," "من," "إلى," "على," and "ما." Stopwords are commonly occurring words that do not bring substantial meaning in sentiment analysis and can minimize model effectiveness by adding excessive computational load. In addition, the process of normalization was applied, which involves converting multiple synonymous text elements into a single unified version. The Arabic letters (و,ي, ة,أ) are normalized to maintain uniformity, ensuring all variations are converted into a single form. For instance, different versions of alif (إ,آ,أ) are converted to (ا), the letter kaaf ك)) is converted to (ک), the letter ya'a (ى) is converted to (ي), the letter ta'a (ة) converted to (ه) and he letters (ؤ,ئ) are converted to]8,18). [ء(. Ultimately, the tokenization process was implemented, which involved breaking down phrases, sentences, and paragraphs into smaller units, words, phrases, or letters. Each small unit was considered as a token [5, 27], which will lead to removing non-letter characters such as *, #, {}, ... etc. In our tests, each token consists of a single word. This step enables the sentiment model to process text more effectively by concentrating on single words as significant units of analysis.

Annotations Guideline

Annotations guideline were demonstrated to annotators through one-hour training session. The detailed guideline presented below with reasoning behind them.

1. News: News should be categorized as neutral regardless of whether it delivers negative or positive information, as it is not intrinsically subjective.

2. Perspective: Sentiments must be evaluated based on the author's perception rather than the annotators' interpretations.

3. Context: Labels should be assigned by the contextual understanding of the text.

4. Ambiguity: When the opinion is ambiguous, select it as neutral, rather than making assumptions.

5. Mixed: The mixed label should be assigned after thorough evaluation.

6. When a tweet encompasses a smiley emoticon but contradicts with textual statement, choose it as mixed.

7. The sentiment of the tweet can still be determined even if the subject is occasionally unclear due to the removal of mentions and hashtags.

This study followed guidelines 1 and 2 on the principle of Abdul-Mageed and Diab (2011) to ensure consistency in annotating news and perspective determination, considering that annotators' opinions may differ from authors. For guideline 3, annotators were instructed to drive sentiments from the tweet's context rather than relying on their background knowledge of the topic when labeling the tweets. For guideline 4, considering the intrinsic ambiguity of the Arabic language and the briefness of tweets, annotators were directed to avoid guessing sentiment and instead mark unclear sentiments as neutral. This approach facilitates avoiding feeding ambiguous instances into the classifier, thus maintaining the accuracy of classification results. For guideline 5, annotators were advised to use mixed labels cautiously. The rationale for this is that a tweet may include both negative and positive words while still expressing a single sentiment. Lastly, since the hashtags and mentions were excluded during the pre-processing of tweets, annotators might be confused about the subject of the tweet. Nevertheless, we ensured during the selection of manual tweets that sentiment analysis was not compromised.

Inter-Annotation Agreement

It is essential to ensure the reliability of annotations. If annotators provide similar labels, it indicates that they interpret the annotation guidelines similarly and have consistency in their work. The trustworthiness of annotations schemes and guidelines. The reliability test assessed the trustworthiness and credibility annotation framework and guidelines. Inter-annotation agreement is employed to measure the difficulty of a task. It can be asserted that if multiple annotators fail to agree on a particular annotation, it is anticipated that the machine learning classifier will also struggle to classify the instance accurately. Fleiss's Kappa (1971) is employed to evaluate IAA, if there are more than two annotators. We estimated Fleiss's Kappa for 27,000 tweets which were annotated by three annotators using the following three sentiment categories: positive, negative, and neutral. Based on the research of Landis and Koch (1977), a Kappa value of 0.60 implies a moderate level of agreement.

Ethical Consideration

This study examines publicly available tweets while following ethical standards for social media research. To maintain ethical integrity, no personally identifiable information was collected or stored, and all data processing complied with relevant guidelines. In addition, sentiment analysis was conducted at an aggregate level, and individual user identities were not analyzed or disclosed.

RESULTS AND DISCUSSION

Annotation-Based Approach

This study recruited three annotators who are CS graduates to manually annotate the dataset in this experiment. These annotators were selected based on their linguistic proficiency, particularly their fluency in the Saudi dialect and their familiarity with the annotation process. The dataset was split equally between two annotators to ensure a balanced workload and reduce individual bias in sentiment classification. Each annotator labeled half of the dataset independently. Each annotator examined the labels designated by the other annotator to confirm that the work completed by the first annotator on half of the corpus was of high quality. A third annotator reviewed a sample of the dataset to draw definitive conclusions about sentiments and confirm the meticulousness of the manual annotation process. The annotation criteria were developed based on previous studies conducted by Al-Twairesh, Al-Khalifa, and Al-Salman (2016) and Cambria, Das, Bandyopadhyay, and Feraco (2017), who thoroughly pre-processed Arabic tweets to guarantee consistency and excellence. We observed that 56% of tweets were classified by all three annotators with the same class while 34% were classified by two annotators with the same class and the remaining 10% had no initial consensus and were ultimately classified by the third annotator.

The annotation results are shown in Figure 1 which shows that approximately 50.27% of tweets were neutral. 31.13% were positive, compared to 18.6% were negative.

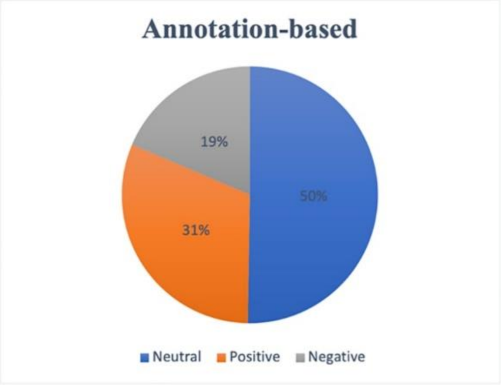

Figure 1. Percentage Distribution of Sentiment Using Annotation.

Lexicon-Based Approach

The second experiment used a lexicon-based approach. SauDiSenti comprises 4,431 words and phrases, of which 1,079 are positive words and phrases (24%), and 3,351 are negative words and phrases (76%) (Al-Thubaity, Alqahtani, & Aljandal, 2018).

VADER, a commonly used Python library specifically designed for sentiment analysis, is known for its lexicon-based approach, which is refined for social media language. When used with the SauDiSenti lexicon, it assigns a polarity score of +1, 0, or −1 to each word in a given record (tweet). The sentiment score was assigned +1 for each positive word in a tweet, -1 for each negative, and 0 for each neutral word. This aligns numbers with good, balanced, and poor scores. Sentiment scores are aggregated at the tweet level, where the tweet is classified as positive sentiment if the overall score exceeds zero, negative if the score is below zero, and neutral if the total score is exactly zero.

Figure 2 shows the results of the lexicon-based experiment. The bulk of the data in this experiment were also neutral (43%), positive (34%), and negative (23%).

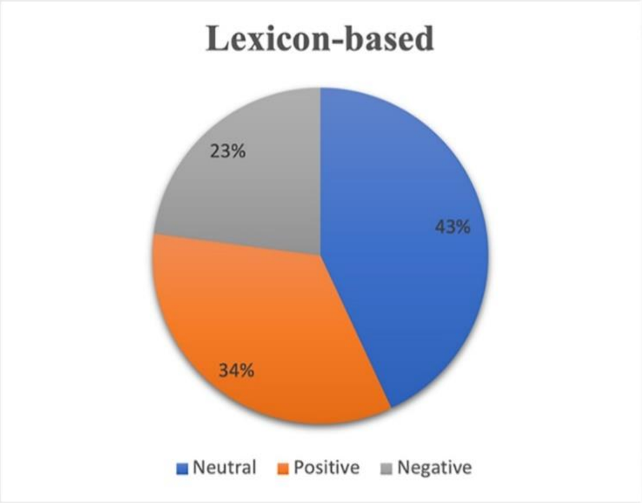

Figure 2. Percentage Distribution of Sentiment Using a Lexicon.

Evaluation of the First Phase

Table 1 and Figure 3 show the comparative outcomes of experiments conducted using lexicons and annotations that were evaluated in the first phase of this study. The utilization of the SauDiSenti lexicon for sentiment analysis yielded an accuracy rate of 92% (Table 2). Additionally, the F1-score for positive polarity classes were 89.7% and for negative polarity classes were 83.8%. This provides a commendable indication of the application of sentiment analysis to tweets written in the Saudi dialect using the SauDiSenti lexicon.

Table 1. Comparison of Results of Annotation-Based and Lexicon-Based Experiments.

Class	Annotation		Lexicon	
Neutral	13,571	50.27%	11,627	43%
Positive	8,404	31.13%	9,186	34%
Negative	5,023	18.60%	6,185	23%
Total	26,998	100%	26,998	100%

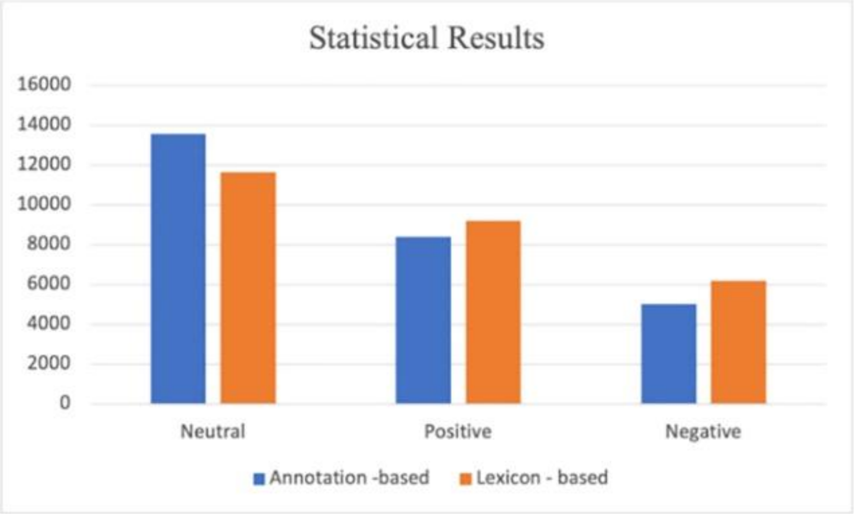

Figure 3. Comparison of Results of Annotation-Based and Lexicon-Based Experiments.

Table 2. Performance of Lexicon-Based Approach.

Class	Accuracy	Precision	Recall	F1-score
Positive	92.8%	91.5%	87.9%	89.7%
Negative		81.2%	86.5%	83.8%
Neutral		100%	85.3%	92.1%

Machine Learning Approach: Second Phase of Evaluation

The second phase of this work involves assessing the SauDiSenti lexicon using various machine-learning classifiers and their effectiveness is demonstrated in the previous literature (Aldayel & Azmi, 2015; Nhu et al., 2020; Aljameel et al., 2020). This was achieved via a comparison of the two experiments. The first stage involved feeding the three machine-learning classifiers with the training data from the lexicon-based experiment corpus. Subsequently, the remaining data, referred to as testing data, were classified to determine customer satisfaction with the STC Pay application. In the second experiment, the training data obtained from the corpus of the annotation-based experiment were utilized to train the classifiers. In contrast, the test data were classified for the same aim.

Experimental Setup

For our experiment, we used Python programming language with PyCharm serving as our IDE to enhance ease of use and workflow efficiency. Data manipulation is performed by employing libraries such as pandas, numpy, and string. However, sklearn is employed for machine learning tasks. Excel was also employed in the construction of the dataset, primarily for classifying and filtering tweets.

To evaluate sentiment classification performance, this study used three broadly used machine learning classifiers named, Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR). The selection of SVM, NB, and LR as benchmark classifiers because these machine learning methods demonstrate high performance on textual data analysis tasks. Consequently, SVM is highly effective in high-dimensional spaces like text data, NB was chosen due to its probabilistic nature while LR was included due to its strong performance in binary classification tasks. Moreover, the neutral data from both corpora were removed to begin the tests, given that our script is only relevant to positive and negative classifications. Each resulting corpus was then split into training and testing sets at 70:30 ratios, as shown in Table 3, and then fed to each classification model for training. Three classifier models were constructed for testing using a tenfold cross-validation approach to minimize the distribution bias. Moreover, features represent numerical values derived from text and serve as inputs for machine learning classifiers. It embodies an important procedure for optimizing machine learning models. To prevent overfitting, this technique streamlines the data by decreasing its dimensionality. Therefore, an effective classification depends on selecting the most appropriate features. In this study, feature extractions are performed using the term frequency-inverse document frequency (TF-IDF) method to ensure the machine learning model effectively captures the sentiments articulated in tweets. The TF-IDF method was employed to transform the two corpora into vectors and extract the compulsory features for training. This approach not only evaluates word occurrences but also assigns a score reflecting the balance between term frequency (TF) and inverse document frequency (IDF), giving more weight to words that appear less frequently in the dataset for better classification.

Table 3. Size of Training Dataset and Testing Dataset for Each Corpus.

Type	Size	Training	Testing
Annotation corpus	13,427	9,398	4,029
Lexicon Corpus	15,371	10,759	4,612

The classification results obtained from the two tests are listed in Table 4. Classification models utilizing an annotation-based corpus achieved greater accuracy than those utilizing a lexicon-based corpus experiment, which is expected given that annotation is inherently more accurate. Equally, the efficacy of organized models developing lexicon-based corpora was deemed adequate, with the SVM validating a striking and pleasing accuracy of 93%. In contrast, the NB and LR models attained 86% and 90% accuracy, respectively. In both experiments, SVM outdid the other classifiers in terms of accuracy and F1 score.

Moreover, as the lexicon-based experiment was the primary focus of our study, the three machine learning classifiers demonstrated acceptable overall performance, with SVM surpassing them all with an accuracy of 93%. It attained a precision of 93% for positive tweets and 92% for negative tweets; whereas, a recall of 95% for positive tweets and 90% for negative tweets. The SVM received a high F1 score for positive (94%) and negative tweets (91%). This evidences that applying the SauDiSenti lexicon to sentiment analysis in the Saudi dialect is viable across various subject domains. However, Belal, She, and Wong (2023) investigated the potential of ChatGPT in generating labeled datasets for sentiment analysis across tasks. They found that ChatGPT performs better than lexicon-based unsupervised methods, achieving a significant rise in the accuracy of 20% and 25% for tweets and Amazon datasets respectively. In a similar vein, Al-Thubaity et al. (2023) conducted a comparative analysis of three large language models (LLM) for Dialectal Arabic Sentiment Analysis such as ChatGPT based on GPT-3.5 and GPT-4, and Bard AI. The findings demonstrate that GPT-4 performs better than GPT-3.5 and Bard AI in sentiment analysis classification.

Table 4. Comparison of Results of Machine Learning Classification.

Experiment	Classifier	Accuracy	Recall		Precision		F1-score		Support	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Annotation-based	NB	95%	97%	94%	96%	95%	97%	94%	2,536	1,493
	SVM	96%	95%	96%	98%	93%	96%	94%		
	LR	95%	94%	96%	98%	91%	96%	93%		
Lexicon-based	NB	86%	96%	71%	83%	93%	89%	81%	2,722	1,890
	SVM	93%	95%	90%	93%	92%	94%	91%		
	LR	90%	95%	83%	89%	92%	92%	87%		

The effectiveness of machine-learning classifiers, particularly when operational with imbalanced datasets, is the area under the receiver functioning characteristic curve (AUC) another metric used to compare. The AUC procedures the classifier that can differentiate between positive sentiments and negative sentiments in the set of tests. It is dignified a liable metric when assessing divergent classification systems for both balanced and imbalanced datasets [32]. Correspondingly shown in Figure 4, in the annotation-based experiment, both SVM and LR achieved an AUC value of 0.99, signifying an exceptional classifier performance. In contrast, the lexicon-based experiment produced AUC values of 0.97 and 0.98 for LR and SVM, respectively, as shown in Figure 5. While the SVM and LR in the lexicon-based algorithm were comparatively lower than those in the annotation-based experiment, this still signifies an exceptionally strong performance in implementing the SauDiSenti lexicon.

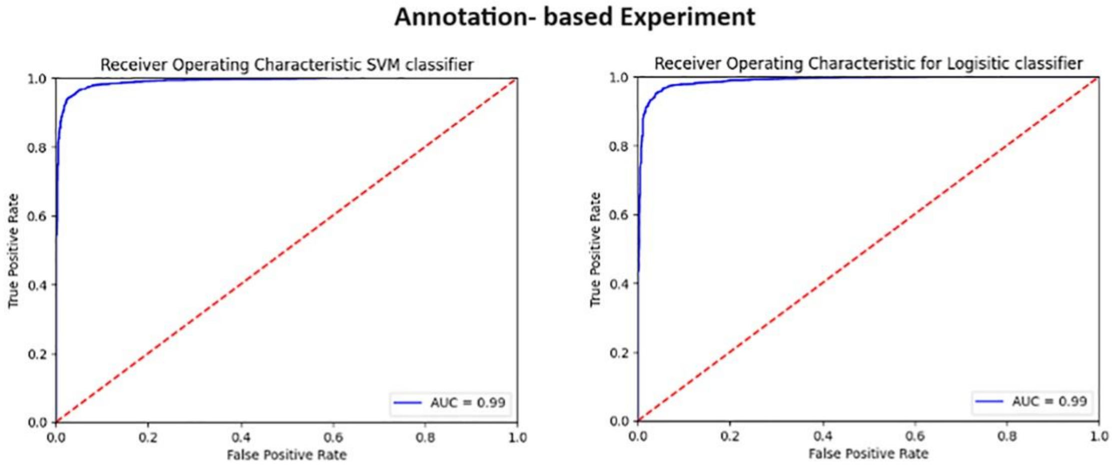

Figure 4. AUC of SVM and LR Classifiers in the Annotation-Based Experiment.

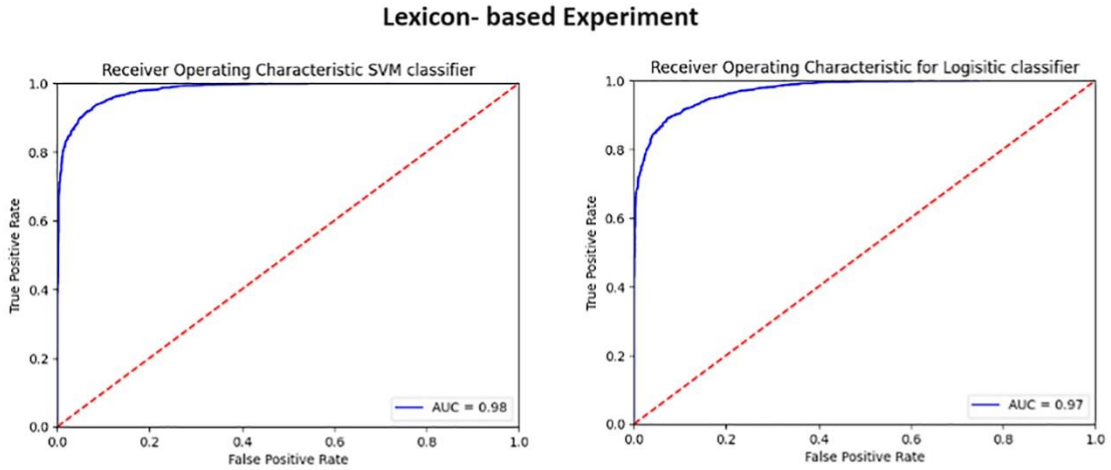

Figure 5. AUC of SVM and LR classifiers in the Lexicon-Based Experiment.

IMPLICATIONS

The present study has significantly contributed to determining the effectiveness of applying the SauDiSenti lexicon to Saudi dialect texts for sentiment analysis. Additionally, we examined the applicability of employing this lexicon to various topics discussed by the Saudi people on Twitter. The results demonstrate that the SauDiSenti lexicon implementation achieved an exceptionally high F1 Score. The lexicon is effective for articulating the sentiment of the Saudi dialect. The use of the SauDiSenti lexicon for sentiment analysis of tweets in the Saudi dialect is successful and yields excellent results. Despite the limited scale of the lexicon, the results were similar to those obtained using manual annotation. The machine learning classifiers attained good accuracy and F1-score when trained on a corpus created with the SauDiSenti lexicon. These outcomes recommend that the lexicon efficiently internments the sentiments of the Saudi dialect. The high AUC values of SVM and LR demonstrated outstanding performance, despite the fact it has a smaller size, this approach would be sufficient, especially when dealing with larger datasets or different topics. However, when analyzing sentiment, the SauDiSenti lexicon proves to be highly effective for sentiment analysis of tweets in the Saudi dialect as it generated impressive results. So, it is manifested that the lexicon must be extended in size and assortment to be more applicable to various datasets. To improve the lexicon's suitability for enormous datasets and datasets with diverse topics, it is suggested that future efforts be enthusiastic to expand its size and diversity.

This study not only demonstrates the effectiveness of the SauDiSenti lexicon for sentiment analysis of Saudi dialect texts but also provides a valuable dataset and methodology that can be built upon in future research. While the dataset and associated resources are not publicly available, researchers can apply similar methodologies using their datasets to explore sentiment trends, validate lexicon-based sentiment classification, and develop enhanced sentiment analysis models for the Saudi dialect. A README file has been provided to document the dataset's structure, preprocessing steps, and usage guidelines, ensuring clarity for those interested in replicating the approach. However, researchers interested in accessing the dataset may request it directly from the corresponding author. Future work can focus on expanding the lexicon by incorporating additional words and contextual meanings, thereby improving its applicability to a wider range of datasets and topics.

STUDY LIMITATIONS AND FUTURE DIRECTIONS

This study has some limitations. First, the limited size of the used lexicon reduces its effectiveness for broader datasets across different topics. The 27,000 tweet dataset delivers useful findings, yet the study cannot eliminate possible biases from tweet selection and manual annotation. Future studies must focus on the construction of a vast diversified lexicon database. Moreover, the lexicon relies on a predetermined set of words and phrases, potentially limiting its ability to adapt to the continuous evolution of the Saudi dialect such as contextual variation, new slang, etc. In the future, it may be possible to enhance the SauDiSenti by expanding the lexicon with contextual understanding and slang. Furthermore, while machine learning models exhibited potential performance, this study didn't incorporate deep learning models, large language models, or generative AI. Future studies should investigate the implementation of models including AraBERT, GPT-4, and LLMs explicitly trained in Arabic dialects. These models could improve accuracy by capturing detailed and contextual sentiment expressions surpassing the potentials of traditional machine-learning approaches.

CONCLUSION

This study explores the efficacy of the SauDiSenti lexicon in conducting sentiment analysis of Saudi dialect tweets related to the STC pay application. This study collected 27,000 tweets and used annotation-based, lexicon-based, and machine learning approaches to classify sentiment. The findings demonstrated that the annotation-based approach offered the most accurate sentiment classification with approximately 50.27% of neutral tweets, 31.13% of positive and 18.6% of negative tweets. The lexicon-based approach employing the SauDiSenti lexicon also exhibited significant effectiveness with an overall accuracy of 92% and F1 scores of 89.7% and 83.8% for positive and negative sentiments, respectively. Moreover, among machine learning classifiers, SVM outperforms the NB and LR classifiers across both annotation-based and lexicon-based approaches with a 96% and 93% accuracy rate for sentiment analysis tasks in Arabic dialects. However, the SVM in the lexicon-based algorithm was comparatively lower than those in the annotation-based experiment, this still signifies an exceptionally strong performance in implementing the SauDiSenti lexicon.

CONFLICT OF INTEREST

There was no conflict of interest declared by the authors.

REFRENCES

[1] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013, December). Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)* (pp. 1-6). IEEE.

[2] Abdul-Mageed, M., & Diab, M. (2011, June). Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Proceedings of the 5th linguistic annotation workshop* (pp. 110-118).

[3] Abdul-Mageed, M., & Diab, M. T. (2014). *Sana: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis*. In *LREC* (pp. 1162-1169).

[4] Alahmary, R. M., Al-Dossari, H. Z., & Emam, A. Z. (2019). Sentiment analysis of Saudi dialect using deep learning techniques. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1-6). IEEE.

[5] Al-Ajlan, A., & Alshareef, N. (2023). Recommender system for Arabic content using sentiment analysis of user reviews. *Electronics*, *12*(13), 2785. https://doi.org/10.3390/electronics12132785.

[6] Alali, M., Mohd Sharef, N., Azmi Murad, M. A., Hamdan, H., & Husin, N. A. (2022). Multitasking learning model based on hierarchical attention network for Arabic sentiment analysis classification. *Electronics*, *11*(8), 1193. https://doi.org/10.3390/electronics11081193.

[7] Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*, *2*(2), 101-114.

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

[9] Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., & Alshamrani, F. M. (2021). A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. *International journal of environmental research and public health*, *18*(1), 218.

[10] Almuhammadi, A. (2020, March). An overview of mobile payments, fintech, and digital wallets in Saudi Arabia. In *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 271-278). IEEE.

[11] Almuqren, L., & Cristea, A. (2021). AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. *PeerJ Computer Science*, *7*, e510. https://doi.org/10.7717/peerj-cs.510.

[12] Alqarafi, A., Adeel, A., Hawalah, A., Swingler, K., & Hussain, A. (2018). A semi-supervised corpus annotation for Saudi sentiment analysis using Twitter. In *Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9* (pp. 589-596). Springer International Publishing.

[13] Al-Razgan, M., Alrowily, A., Al-Matham, R. N., Alghamdi, K. M., Shaabi, M., & Alssum, L. (2021). Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by Saudi women. *Technology in Society*, *65*, 101558. https://doi.org/10.1016/j.techsoc.2021.101558.

[14] Alruily, M. (2021). Classification of Arabic tweets: A review. *Electronics*, *10*(10), 1143. https://doi.org/10.3390/electronics10101143.

[15] Al-Thubaity, A., Alkhereyf, S., Murayshid, H., Alshalawi, N., Omirah, M., Alateeq, R., Alkhanen, I. (2023, December). Evaluating ChatGPT and bard AI on Arabic sentiment analysis. In *Proceedings of Arabic NLP 2023* (pp. 335-349).

[16] Al-Thubaity, A., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia computer science*, *142*, 301-307. https://doi.org/10.1016/j.procs.2018.10.494.

[17] Al-Twairesh, N., Al-Khalifa, H., & Al-Salman, A. (2016, August). 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)* (pp. 697-705).

[18] Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for Arabic sentiment analysis of Saudi tweets. *Procedia Computer Science*, *117*, 63-72. https://doi.org/10.1016/j.procs.2017.10.094.

[19] Analysys mason. (2023). *STC pay is the most popular operator-branded mobile wallet in the Gulf region, according to our survey*. Retrieved from https://www.analysysmason.com/research/content/articles/stc-pay-gulf-rdmm0-rdmy0-rdrk0/.

[20] Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *Journal of information science*, *44*(2), 184-202. https://doi.org/10.1177/0165551516688143.

[21] Belal, M., She, J., & Wong, S. (2023). Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv*:2306.17177.

[22] Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In *International conference on discovery science* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

[23] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (Vol. 5). Cham: Springer International Publishing.

[24] Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. *Electronics*, *11*(3), 374.

[25] El-Beltagy, S. R., & Ali, A. (2013, March). Open issues in the sentiment analysis of Arabic social media: A case study. In *2013 9th International Conference on Innovations in Information Technology (IIT)* (pp. 215-220). IEEE. https://doi.org/10.1109/Innovations.2013.6544421.

[26] Elgibreen, H., Faisal, M., Al Sulaiman, M., Abdou, S., Mekhtiche, M. A., Moussa, A. M., & Algabri, M. (2021). An incremental approach to corpus design and construction: application to a large contemporary Saudi corpus. *IEEE Access*, *9*, 88405-88428.https://doi.org/10.1109/ACCESS.2021.3089924.

[27] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

[28] Hinze, A., Heese, R., Luczak-Rösch, M., & Paschke, A. (2012). Semantic enrichment by non-experts: usability of manual annotation tools. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11* (pp. 165-181). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35176-1_11.

[29] Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44*(4), 491-511.

[30] Laifa, M., & Mohdeb, D. (2023). Sentiment analysis of the Algerian social movement inception. *Data Technologies and Applications*, *57*(5), 734-755.

[31] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

[32] Leech, G. (1993). Corpus annotation schemes. *Literary and linguistic computing*, *8*(4), 275-281.https://doi.org/10.1093/llc/8.4.275.

[33] Mohammad, S., Salameh, M., & Kiritchenko, S. (2016, May). Sentiment lexicons for Arabic social media. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 33-37).

[34] Nhu, V. H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., & Ahmad, B. B. (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve Bayes tree, artificial neural network, and support vector machine algorithms. *International journal of environmental research and public health*, *17*(8), 2749. https://doi.org/10.3390/ijerph17082749.

[35] Rahab, H., Haouassi, H., & Laouid, A. (2023). Rule-based Arabic sentiment analysis using binary equilibrium optimization algorithm. *Arabian Journal for Science and Engineering*, *48*(2), 2359-2374.

[36] Refaee, E., & Rieser, V. (2014). An Arabic Twitter corpus for subjectivity and sentiment analysis. In *9th International Language Resources and Evaluation Conference* (pp. 2268-2273). European Language Resources Association.

[37] Sherif, S. M., Alamoodi, A. H., Albahri, O. S., Garfan, S., Albahri, A. S., Deveci, M., & Kou, G. (2023). Lexicon annotation in sentiment analysis for dialectal Arabic: Systematic review of current trends and future directions. *Information Processing & Management*, *60*(5), 103449. https://doi.org/10.1016/j.ipm.2023.103449.

[38] Statista. (2023). *Countries with the most X/Twitter users*. Retrieved from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.