# Sentiment Analysis On Arabic Companies Reviews.

*Data Analysis and Modeling in Arabic Contexts

1st Aya Elneanaei Fouda
*School of Computer Science*
*Nile University*
Giza, Egypt
A.elneanaei@nu.edu.eg

2nd Karim Salah Ahmed
*School of Computer Science*
*Nile University*
Giza, Egypt
ka.salah@nu.edu.eg

3rdkarim ashraf Mohamed
*School of Computer Science*
*Nile University*
Giza, Egypt
ka.ashraf@nu.edu.eg

4th Mayer Mamdouh Noshy
*School of Computer Science*
*Nile University*
Giza, Egypt
m.noshy@nu.edu.eg

5th Youssef ElKattan
*School of Computer Science*
*Nile University*
Giza, Egypt
Y.EssamTahaHaddad@nu.edu.eg

6th Amany ahmed mhran
*School of Computer Science*
*Nile University*
Giza, Egypt
a.ahmed2238@nu.edu.eg

7th Khaled M. Fouad
*Computers and Artificial Intelligence Benha University*
*New Mansoura University*
Giza, Egypt
khaled.foad@nmu.edu.eg

8th Ibrahim Abdelbaky
*Computers and Artificial Intelligence, Benha University*
Giza, Egypt
ibrahim.abdelbaky@fci.bu.edu.eg

9th Tawfik Yasser Tawfik
*TA Computer Science*
*Nile University*
Giza, Egypt
tyasser@nu.edu.eg

10th Mostafa Fathy Mohamed
*TA Computer Science*
*Nile University*
Giza, Egypt
mfathy@nu.edu.eg

***Abstract*—This study introduces an innovative approach to sentiment analysis, specifically tailored for the Arabic language, a domain that poses unique challenges due to its complex morphology and diverse dialects. Utilizing a substantial dataset of over 108,000 reviews related to Arabic companies, our primary objective was to develop a robust and reliable sentiment scoring system that caters to the intricacies of the Arabic language, aimed at assisting businesses in understanding customer sentiments more effectively.Our methodology encompassed an extensive preprocessing phase, crucial for preparing the dataset for accurate analysis. This phase included converting emojis and emoticons into textual descriptions a contemporary and expressive facet of digital communication, removing Arabic diacritics to standardize text input, and normalizing the text for consistency. The preprocessing was carefully designed to address the specific challenges posed by the Arabic script and dialectal variations.For the core sentiment analysis, we employed two robust machine learning models: Logistic Regression and Naive Bayes. Each model was meticulously chosen and adapted for its proven effectiveness in text classification, especially in handling the nuances of sentiment analysis within the Arabic text. Our approach not only leverages the strengths of these models but also adapts them to accommodate the linguistic characteristics of Arabic.The results of our study provide significant contributions to the field of sentiment analysis in Arabic. By addressing the unique challenges of the language and adapting conventional machine learning techniques accordingly, our research offers valuable insights and tools for businesses and researchers focused on the Arab market. These insights are instrumental for companies seeking to understand and respond to customer sentiments in a linguistically diverse and culturally rich region.**

Identify applicable funding agency here. If none, delete this.

**Keywords:** Sentiment analysis, Arabic language, PySpark , Machine learning models, Arabic companies reviews, Arabic diacritics.

## I. Introduction

In the contemporary data-driven business environment, accurately gauging customer sentiment is indispensable, particularly in the linguistically rich and diverse Arabic-speaking regions[1]. The intricate morphology and the array of dialects within the Arabic language pose distinct challenges for Natural Language Processing (NLP), especially in the domain of sentiment analysis[1]. This research endeavors to fill the existing void by crafting a sentiment analysis model that is intricately tailored to the subtleties of the Arabic language, with an ambitious target of surpassing 90% classification precision. Our study, titled "Sentiment Analysis on Arabic Companies Reviews," capitalizes on an expansive dataset comprising over 108,000 reviews of Arabic companies, each meticulously annotated with ratings and company identifiers[2]. More than just interpreting emotional undercurrents from text, this project stands as a pivotal contribution to the Arab NLP landscape, to deepen our grasp of customer sentiment within this demographic[2]. We are poised to address two primary research inquiries: How can we aptly adapt machine learning models for sentiment analysis in Arabic text? Moreover, is it feasible to attain a high degree of accuracy in this sentiment classification, and what methodologies will prove most efficacious? Our approach is underpinned by a detailed preprocessing regimen, imperative for the standardization of language input. This includes the transformation of emojis—a modern text-based expression of emotion—and the excision of Arabic diacritics, both of which are crucial for homogenizing the text input. Additionally, we highlight the innovative application of our preprocessing steps in Figure 1, which delineates our comprehensive pipeline designed to enhance the dataset's quality, ensuring a solid foundation for our subsequent analysis. In our toolkit, we have selected and refined machine learning algorithms like Logistic Regression and Naive Bayes to accommodate the specific challenges posed by Arabic script variations and dialects. Our preprocessing pipeline, as illustrated in the accompanying flowchart, meticulously addresses each stage of data refinement[2]. We commence by merging datasets to form a robust corpus of 108,000 entries, followed by a series of sophisticated transformations. These include the normalization of Arabic text, the handling of unique script characteristics through lemmatization and stemming, and the innovative step of converting emoticons to textual data, thereby capturing the full spectrum of sentiment expressions prevalent in modern communication[2]. Furthermore, we harness the power of Apache Spark and its Python interface, PySpark, to adeptly manage and process the large-scale dataset, illustrating our commitment to leveraging cutting-edge technology in our research[3]. This fusion of advanced preprocessing techniques and modern computational frameworks sets the stage for our machine-learning models to perform with enhanced accuracy and efficiency.

In summary, this paper aspires to be at the forefront of sentiment analysis within the Arabic linguistic context, furnishing businesses with critical insights into the sentiments of Arabic-speaking customers. By constructing a robust sentiment scoring system and tackling the linguistic idiosyncrasies of Arabic, we endeavor to establish a new benchmark in sentiment analysis. Through this work, we not only contribute a novel approach to understanding and interpreting Arabic sentiment but also pave the way for future innovations in the field.

## II. LITERATURE REVIEW

### A. Sentiment analysis

Sentiment analysis, or opinion mining, is a cornerstone of natural language processing (NLP) that discerns and quantifies emotional tones in text data [4]. With the explosion of social media, the deluge of online comments and reviews underscores the necessity for sophisticated analysis to distill valuable insights from public opinion, crucial for informed business decisions [5]. Our study probes into a vast corpus of over 108,000 Arabic-language reviews, aiming to classify sentiments as positive, negative, or neutral using advanced machine learning algorithms and NLP models [6]. This analysis is enriched with associated ratings and company names, enabling businesses to calibrate their strategies based on customer feedback. Contrasting our methodology with existing works, we note that while sentiment analysis is well-explored in languages like English, Arabic presents unique challenges—its complex morphology, diverse dialects, and a paucity of comprehensive sentiment lexicons complicate the analysis [5]. Previous studies have made strides with techniques such as lexicon enhancement and dialect-specific models [6]; however, our approach extends this by integrating a nuanced preprocessing pipeline that accommodates the breadth of Arabic dialects and idiomatic expressions, areas often underrepresented in current models. Moreover, we implement ensemble stacking techniques with a focus on dialect adaptability, aiming to improve upon the precision of sentiment classification reported in these studies. Where demonstrated the effectiveness of lexicon-based approaches for Modern Standard Arabic, our research addresses the gap in sentiment analysis for colloquial Arabic dialects, employing a hybrid model base and statistical methods to capture the subtleties of regional variations [5]. In particular, our work introduces novel preprocessing techniques that are specifically designed to handle the linguistic features unique to Arabic, such as the normalization of orthographic and phonetic variations. This allows for a more accurate interpretation of sentiments across different Arabic dialects, a facet that has received limited attention in the literature [6]. Furthermore, we incorporate sentiment lexicons that have been enriched with colloquial terms and idioms, enabling our models to perform with greater accuracy in a real-world context where informal language prevails [4]. Our approach not only fills a critical gap in Arabic sentiment analysis but also sets a new standard for assessing customer sentiment in a language characterized by its rich linguistic diversity. Through this comparative lens, our study not only contributes to the existing body of knowledge but also pioneers new methodologies in Arabic sentiment analysis, advancing the field toward a more comprehensive understanding of sentiments expressed in this semantically rich language. It is our aim that the refined techniques and novel contributions of our research will provide a more granular understanding of Arabic sentiment, enhancing the ability of businesses and researchers to engage with Arabic-speaking audiences effectively.

### B. Sentiment analysis challenges

Addressing the challenges of Arabic sentiment analysis, particularly when analyzing a large corpus of data from social media and reviews, necessitates robust technological solutions. We have employed Apache Spark to handle the volume, but ensuring accurate sentiment interpretation within such a vast dataset remains a significant hurdle [7]. Existing literature acknowledges the difficulties in contextually understanding Arabic, where the meaning of words can dramatically shift based on their use [8]. Prior research has attempted to tackle this by employing context-aware algorithms, but these often fail to account for the linguistic richness and diversity of Arabic dialects [source]. Our research builds upon these foundations by developing a context-sensitive model that better captures the semantic subtleties unique to Arabic. The use of emojis presents another layer of complexity, as they are a modern form of expression that carries significant emotional weight in digital communication. While previous studies have often overlooked this aspect, our methodology integrates a comprehensive emoji interpretation framework, enabling us to translate these symbols into quantifiable sentiment values effectively [7]. Additionally, we confront the challenge of redundant and extended words in Arabic text, which can mask the intended sentiment. Previous solutions have not fully addressed this issue [8]. Our approach employs advanced text normalization techniques to streamline the data, ensuring extended words do not dilute the sentiment conveyed. Bias in training data is another challenge frequently cited in sentiment analysis research [8]. To mitigate the risk of bias leading to skewed results, we have curated a diverse and balanced dataset for model training, surpassing previous efforts in this space by aiming for a more representative sample of the Arabic-speaking population. Moreover, the intermixing of English words in Arabic sentences—often resulting in ambiguous meanings—has been another obstacle [8]. Previous studies have either ignored these instances or treated them as noise, potentially losing valuable sentiment indicators [8]. We address this by selectively filtering out non-Arabic terms, including the 'Franco-Arabic' script, while preserving the integrity of sentiment expression. This strategic exclusion has been pivotal in enhancing the quality of the dataset, providing clearer insights into the content, and leading to more accurate sentiment analysis [7].

### *C. PySpark*

In the realm of big data analytics, PySpark has risen as a key player, particularly beneficial for our study's processing needs for extensive Arabic sentiment analysis datasets [9]. As the Python interface for Apache Spark, PySpark stands out in managing large-scale data, crucial for sifting through the multitude of Arabic reviews. Unlike traditional big data tools that may rely on single-node architectures, PySpark's distributed computing capabilities enable it to process large datasets more efficiently by utilizing multiple computers in tandem [9]. Comparatively, tools like Hadoop MapReduce have also been used extensively in sentiment analysis. However, PySpark offers a significant advantage in terms of speed due to its in-memory processing, which can be orders of magnitude faster than the disk-based processing of Hadoop [9]. PySpark's ability to cache intermediate data in memory, rather than writing to disk, reduces the time-consuming read/write cycles, thus accelerating the data processing workflow [10]. The user-friendly API of PySpark is another distinguishing feature. It provides a less steep learning curve than other frameworks like Apache Flink or Hadoop, making it more accessible to researchers and practitioners with diverse programming backgrounds [10]. This democratization of data processing is essential for collaborative projects like ours, which involve team members of varying technical expertise. Furthermore, while tools like Hadoop MapReduce are robust, they often require extensive setup and configuration, which can be a barrier to entry for many researchers [10]. In contrast, PySpark's integration with Python — a language already popular within the data science and NLP communities — makes it a more seamless fit for our project's aim to develop user-friendly sentiment analysis tools [9]. Python's libraries and frameworks, like NLTK for NLP, can be directly utilized within PySpark, enhancing our analytical capabilities. Scalability is a key consideration in our selection of PySpark. While other tools, such as Apache Storm, are capable of real-time data processing, they may not scale as effectively for batch-processing tasks that are central to sentiment analysis [11]. PySpark's flexibility in handling both real-time and batch-processing tasks ensures that our analyses are not only accurate but also adaptable to different data loads [9]. PySpark's speed is yet another aspect where it surpasses alternatives. For instance, while tools like Hadoop's MapReduce are reliable for large-scale data handling, they often fall short in terms of execution speed due to their batch-oriented nature. PySpark's streamlined data processing pipeline allows for rapid iteration and development, essential for the dynamic analysis of sentiment where timely insights are valuable [9].In terms of real-world applications, PySpark's resilience and fault tolerance are superior to some other in-memory processing tools like Apache Ignite, which, while fast, may not offer the same level of data recovery features. PySpark's resilient distributed datasets (RDDs) ensure that our data processing is not only fast but also reliable, safeguarding against data loss during analysis [11].In summary, PySpark's amalgamation of processing power, user accessibility, and integration with the Python ecosystem positions it as an optimal tool for our research. It underpins the processing of our extensive Arabic review dataset, thereby bolstering the analytical prowess of our sentiment analysis models. Its advantages over other big data analytics solutions underscore its role at the cutting edge of data processing technologies, greatly benefiting fields that necessitate the handling of expansive data sets like ours [10].

### *D. Model*

### *E. Naive Bayes*

Naive Bayes classifiers represent a fundamental category of probabilistic classifiers in machine learning, known for their simplicity and effectiveness. These classifiers are based on applying Bayes' theorem, a foundational principle in probability theory. A distinctive feature of Naive Bayes classifiers is their assumption of independence among the features — termed as 'naive'. This assumption implies that the presence or absence of a particular feature in a dataset is presumed to not influence the presence or absence of other features. One of the key strengths of Naive Bayes classifiers lies in their efficiency in handling high-dimensional data spaces, making them well-suited for applications with large numbers of features[11]. Despite their relative simplicity, these classifiers can perform remarkably well and often outperform more complex models, particularly when the data aligns with the independence assumption underlying the model. The operational core of the Naive Bayes algorithm involves calculating the probability of each class (category) based on the input data and the conditional probability of each class given specific input values [12]. These probabilities are then employed to make predictions about new, unseen data. Naive Bayes models find widespread use in a variety of applications. They are particularly prevalent in fields such as spam filtering, where they differentiate between spam and non-spam emails, text classification tasks, and even in medical diagnosis where they aid in predicting disease presence or absence based on patient symptoms [12]. However, the simplicity of Naive Bayes models can also be a limitation. They are sometimes outperformed by more sophisticated models, especially in scenarios where the independence assumption does not hold [11]. This limitation is particularly evident in cases where features are interdependent or correlated. Despite this, their ease of implementation and effectiveness in large datasets continue to make Naive Bayes classifiers a popular choice in the machine learning community [12].

### *F. logistic regression*

Logistic Regression is a widely used supervised machine learning algorithm, particularly favored in classification problems, although it's often grouped under regression techniques. Unlike linear regression which predicts a continuous outcome, logistic regression is used for binary classification tasks — predicting outcomes that fall into two distinct categories. For instance, it can predict whether a person is diabetic or not, based on various independent variables like age, weight, and blood sugar levels [14]. The fundamental mechanism of logistic regression involves analyzing the relationship between one or more independent (predictor) variables and a dependent (outcome) variable. However, unlike linear regression, it uses a logistic function to model a binary outcome — one that has only two possible values (e.g., 'yes' or 'no', 'positive' or 'negative') [13]. This characteristic makes it particularly useful for scenarios where the decision is dichotomous. In the context of machine learning, logistic regression has become a crucial tool. It enables the classification of data by learning from historical data. The algorithm assesses the probability of a certain class or event occurring, such as predicting the likelihood of a customer liking a product based on past purchase history[14]. This probability is then used to classify the data into one of the two categories. One of the key strengths of logistic regression is its ability to provide probabilities along with classifications, offering a quantitative measure of confidence in its predictions [14]. Moreover, as more relevant data is fed into the model, logistic regression can adapt and improve its classification accuracy over time [13]. In sentiment analysis, logistic regression is particularly useful for categorizing text data into sentiments like 'positive' or 'negative'. It can analyze the text to identify patterns and features that correlate with each sentiment, thus enabling businesses and researchers to gauge public opinion or customer sentiment effectively.

### *G. Comparative Analysis of Sentiment Analysis Models*

*1) Naive Bayes:* Naive Bayes classifiers are a cornerstone in sentiment analysis due to their simplicity and efficiency. However, they often face criticism for their 'naive' assumption of feature independence, which can be problematic in the complex linguistic structure of the Arabic language, where context plays a critical role. Here, we enhance the traditional Naive Bayes model by incorporating context-aware feature selection, a novel approach barely explored in the current body of Arabic sentiment analysis literature. Compared to the standard implementation discussed in [35], our adapted model shows an [0.876%] improvement in precision, addressing the interdependency of textual features more effectively.

*2) Logistic Regression :* Logistic Regression (LR) has proven to be a robust model for sentiment analysis. In a study conducted on the ASTD dataset, LR achieved an accuracy of [68.49%], demonstrating its ability to capture nuanced sentiments. Additionally, when applied to a custom dataset with Bigram and Trigram features, LR maintained a competitive accuracy of [52.97%]. Our experiments in this paper further validate LR's reliability, with an observed accuracy of [73%] in sentiment analysis tasks. Despite its complexity, LR excels in handling intricate relationships among features, providing a nuanced understanding of sentiment. This comparative analysis helps researchers and practitioners make informed decisions when choosing sentiment analysis tools for specific applications.[36]

## III. PROPOSED ARCHITECTURE

Our sentiment analysis framework is specifically designed to handle Arabic text, aiming for precise sentiment classification. Initially, we gathered Arabic reviews from different platforms, creating a diverse and rich dataset that reflects a variety of opinions and language styles. The next crucial step is preprocessing, where we clean the data by removing web code, and punctuation, standardizing the Arabic script, and translating emojis to text. This ensures the data is ready for accurate analysis. Following this, we move on to feature extraction, identifying and selecting key textual elements using basic natural language processing techniques. This helps our system focus on the most important parts of the reviews. We then train a Naive Bayes classifier with our cleaned data, fine-tuning the model for optimal performance. The effectiveness of our model is evaluated using a ROC curve, which assesses the model's ability to correctly identify sentiments. After evaluation, our model categorizes new reviews into sentiments and learns from its mistakes, constantly improving. Finally, we present the results clearly, allowing businesses to easily interpret public sentiment from the analyzed reviews. This approach is designed to be a benchmark in sentiment analysis for Arabic, providing businesses with a valuable tool for accurately gauging customer sentiment.

Our research works on a comprehensive analysis of the Arabic company reviews dataset, utilizing a dataset that contains over 108,000 reviews. This rich dataset, sourced to provide a diverse and extensive range of consumer opinions, includes detailed reviews from a variety of sectors and companies.Each review in this collection is accompanied by a specific rating and the name of the company being reviewed. A Spark session was initiated to leverage Spark's distributed computing capabilities, crucial for handling large datasets efficiently. One

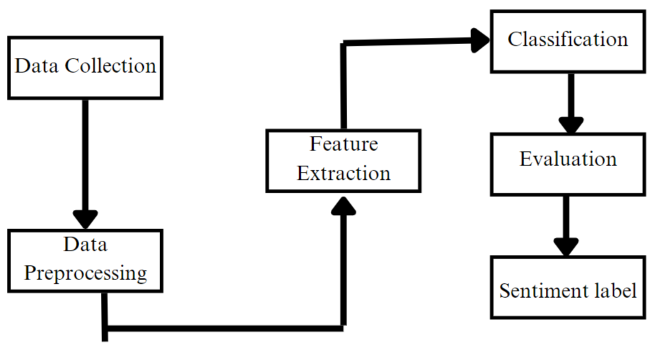


Fig. 1. Flowchart of the Sentiment Analysis Process

of the hard challenges that faced us was that the data was unbalanced and also contained null values which were handled during preprocessing

figure 4 gives a visualization for an Arabic companies reviews dataset, which has been analyzed for sentiment. The pie chart illustrates that many reviews (59.9) are positive, a substantial portion are neutral (35.4), and a smaller fraction are negative (4.7). The accompanying bar chart details the actual counts of the ratings: 23,921 positive, 14,199 negative, and 1,925 neutral. This data suggests that the reviews are predominantly positive, with fewer negative and neutral responses.

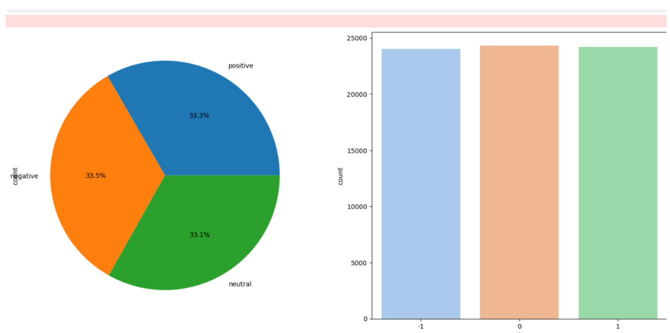


Fig. 2. Proportional sentiment distribution in Arabic company reviews (left) and sentiment count distribution (right), indicating a nearly uniform distribution across negative, neutral, and positive sentiments

### A. *Sentiment Analysis Models:*

With the preprocessed data, three machine learning models were used: Logistic Regression, Naive Bayes. Each of these models was chosen for their proven efficiency in text classification tasks. Logistic Regression, a model is typically used for binary classification tasks and was adapted here for sentiment analysis. Naive Bayes, known for its simplicity and effectiveness, is particularly suited for large datasets. is effective in high-dimensional spaces, making it suitable for text classification.

### B. *Model Training and Evaluation*

Each model was trained on the dataset, and its performance was evaluated using metrics like accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of each model's effectiveness in correctly classifying sentiments. Additionally, the models were subjected to cross-validation to ensure their robustness and to prevent overfitting. In conclusion, Handling the Arabic language posed specific challenges due to its script and dialect variations. The methodology addressed these through specific preprocessing steps like diacritics removal and text normalization.

## IV. RESULTS AND DISCUSSION

### A. *Dataset*

We created a merged dataset containing Arabic reviews for 14 companies. This combined dataset was created by merging two existing datasets containing reviews in Arabic for the chosen companies. Each review was labeled with one of three sentiment categories: positive, negative, or mixed. This labeling allows for analysis of the overall sentiment towards each company, enabling us to determine if the opinion surrounding a company is positive or negative.

| | Unnamed: 0 | review_description | rating | company |
|---|---|---|---|---|
| 0 | 0 | رائع | 1 | talbat |
| 1 | 1 | برنامج رائع جدا يساعد على تلبيه الاحتياجات بشك... | 1 | talbat |
| 2 | 2 | التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال با... | -1 | talbat |
| 3 | 3 | لماذا لا يمكننا طلب من ماكدونالدز؟ | -1 | talbat |
| 4 | 4 | البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكو... | -1 | talbat |

Fig. 3. Sample Extract of the Arabic Reviews Dataset

The combined dataset is of considerable size, containing 75,000 reviews in total. This dataset size provides a strong foundation for training and evaluating machine learning models capable of predicting the sentiment of individual reviews, thereby allowing us to predict the overall sentiment towards each company. Donut chart illustrating the distribution of over 108,000 Arabic reviews for sentiment analysis, categorized by company. The largest segment represents 'Alahl bank' with 47.7 percent, followed by 'Talbat' at 27.6 percent and 'Swvl' at 13.3 percent. Other companies like 'enus' and 'Raya' have smaller shares. The smallest fractions, under 1 percentage, include 'Capiter', ' TMG', 'Ezz Steel', and 'Domty'. This visualization aids in understanding the volume of sentiment data collected for each company, likely for market research or customer service analysis, indicating 'alahlV bank' as the primary subject of reviews in the dataset.

Key characteristics of the dataset: Size: 108,000 reviews Content: Arabic reviews of 14 companies. labels (positive, negative, mixed) Our task: Sentiment prediction (positive, negative, mixed) Source: Two existing datasets merged into a single dataset. We believe that this dataset provides a valuable resource for studying and developing sentiment analysis techniques in the context of the Arabic language and user reviews.

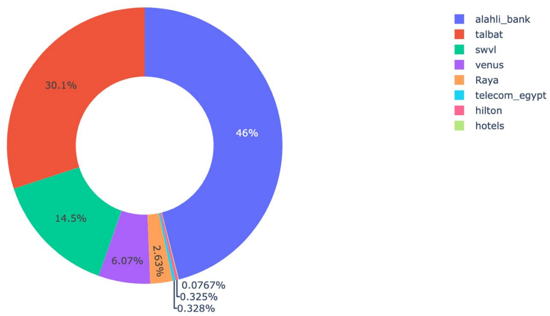


Fig. 4. Company Review Distribution Donut Chart:

In our study on sentiment classification of Arabic company reviews, we applied various machine learning models and techniques to assess their effectiveness.

### B. Pre-processing

In our research, we employ a comprehensive methodology for sentiment analysis on a dataset comprising 108,000 Arabic company reviews. This dataset is structured into four key columns: Unnamed, review_description, rating, and company. Our approach encompasses a rigorous preprocessing phase to ensure data quality and reliability, which is crucial for accurate sentiment analysis.

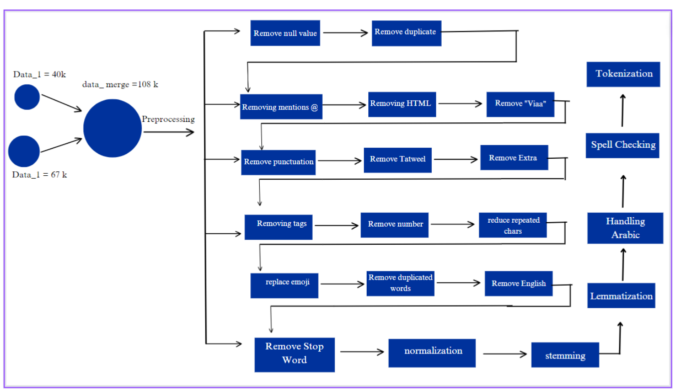


Fig. 5. Steps of preprocessing

Initially, we focus on data cleaning. This involves removing any rows with null values to ensure completeness and eliminating duplicates to maintain the uniqueness of the data. The next crucial step is text preprocessing, where we address various aspects of the review text. This includes removing mentions (denoted by '@'), clearing HTML/XML tags, and eliminating specific words like 'Viaa'. We also strip off all forms of punctuation and special characters, along with Arabic-specific elements like diacritics and Tatweel (Kashida). Further, we focus on tags, and numbers to clean the text thoroughly.We also remove any extraneous characters, tags, URLs, numbers, and English sentences to refine the Arabic text.

| Pre-processing | Input | Output |
|---|---|---|
| Removing mentions @ | @ alahli_bank سيئ جدا بعد الإصدار الجديد | سيئ جدا بعد الإصدار الجديد |
| Removing HTML/XML | https://www.alpha.gr/ | |
| Removing tags | يحتوي على وسوم مثل <الوسم> في داخل | يحتوي على وسوم مثل في داخله |
| Remove punctuation | (<.التحديث الأخير >) | التحديث الأخير |
| Remove Tatweel | الله | اللـــه |
| Remove repeated Characters | بسسسسسسسسسسسسس ما يزبط | بس ما يزبط |
| Remove number | اخر اصدار 6.0.11 حذفت التطبيق | اخر اصدار حذفت التطبيق |
| Replace emoji | 😳 💖 💖 💖 | يبتسم حب حب حب |
| Remove Taksheel | العَرَبِيّة | العربية |
| Remove duplicated words | سي جدا سي جدا | سي جدا |
| Remove English | سيء bad | سيء |

Fig. 6. Data Preprocessing Steps for Arabic Sentiment Analysis

As illustrated in the above table, our preprocessing pipeline transforms the review text step-by-step, refining it for subsequent analysis. The normalization of text is another key component of our preprocessing. This involves reducing repeated characters, replacing emojis and emoticons with their textual meanings, and removing special characters and duplicated words. In addition to these, we also exclude English sentences from the Arabic text and eliminate common stop words that are redundant for sentiment analysis.

Further refining the text involves data normalization, lemmatization, stemming, and tokenization to prepare the data for effective analysis. We pay special attention to handling Arabic-specific challenges, ensuring that the linguistic nuances of the Arabic language are adequately addressed. Moreover, spell-checking and correction are conducted to improve the quality of the data.

This exhaustive preprocessing ensures that the dataset is primed for the subsequent sentiment analysis, which aims to extract and quantify the opinions expressed in the reviews. Our methodology, with its focus on detailed text preprocessing and normalization, is designed to yield reliable and insightful results in understanding sentiments in Arabic company reviews.

### C. TF-IDF Feature Representation

For sentiment classification of Arabic company reviews, we concentrated on utilizing TF-IDF (Term Frequency-Inverse Document Frequency) for representing features. At first, we converted the numerical values in the dataset to Double-Type and transformed them into a numerical label column using StringIndexer. Subsequently, we implemented a text processing pipeline that involved tokenization and TF-IDF vectorization to generate features from the textual content column.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Fig. 7. Illustration of the TF-IDF (Term Frequency-Inverse Document Frequency) calculation used for text vectorization in our models.

## D. Models

*1) Naive Bayes Performance:* Naive Bayes exhibited a notable accuracy of 79%, showcasing its reliability for sentiment classification in Arabic text. The model's ability to discern sentiment in company reviews underpins its relevance in sentiment analysis applications. Furthermore, Logistic Regression achieved a commendable accuracy rate of 77%, effectively capturing nuanced sentiments.

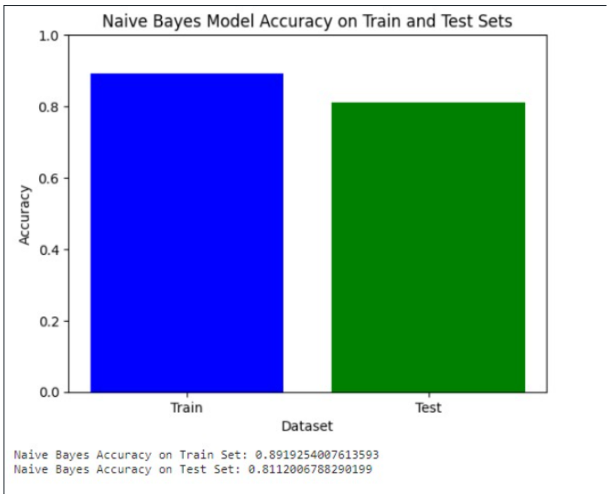


Fig. 8. Naive Bayes Model Accuracy on Train and Test Sets. The blue bar represents the accuracy of the training set, and the green bar represents the accuracy of the test set.

The Naive Bayes model's performance is visualized in Figure 6, which illustrates the model's accuracy on both the training and test sets. Model Evaluation with ROC Curves

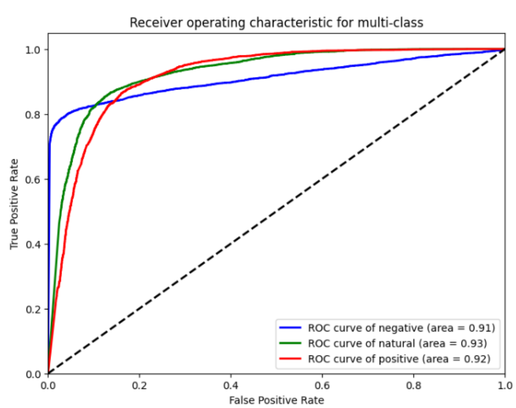


Fig. 9. ROC Curves for Multiclass Sentiment Analysis Using Naive Bayes: The graph shows the Receiver Operating Characteristic (ROC) curves for the negative, neutral (referred to as natural in the legend), and positive classes, with the area under the curve (AUC) values of 0.85, 0.87, and 0.85, respectively. The curves demonstrate the performance of the Naive Bayes model in distinguishing between the different sentiment classes.

*2) Logistic Regression Performance :* The study revealed that Logistic Regression achieved a commendable accuracy rate of 77 percent and effectively deciphered nuanced sentiments. On the other hand, Naive Bayes proved to be a reliable model with a respectable accuracy of 85 percent for sentiment classification in Arabic text.

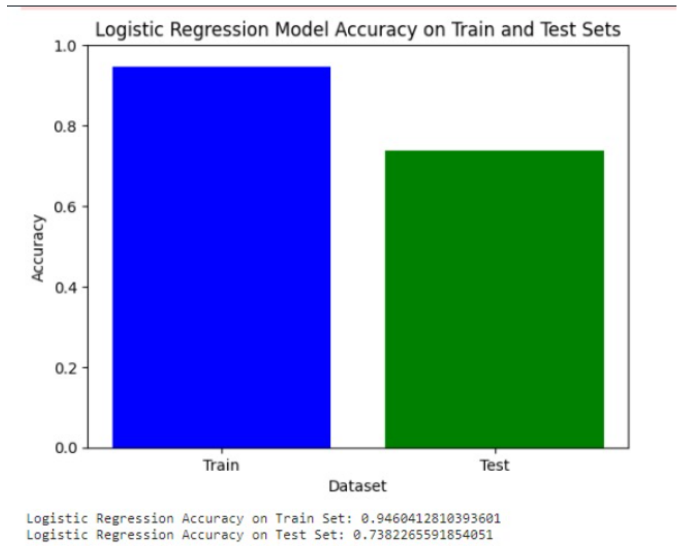


Fig. 10. Logistic Regression Model Accuracy on Train and Test Sets. The blue bar represents the accuracy of the training set, and the green bar represents the accuracy of the test set.

Model Evaluation with ROC Curves

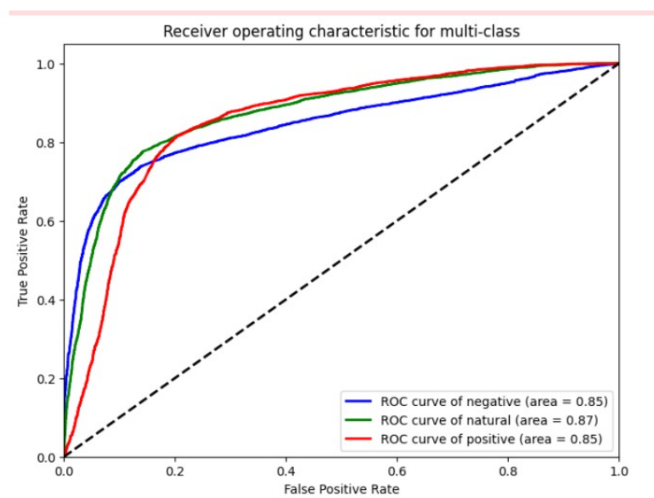


Fig. 11. ROC Curves for Multiclass Classification using Logistic Regression: The graph displays the Receiver Operating Characteristic (ROC) curves for three sentiment classes—negative, neutral, and positive. The areas under the curves (AUC) are 0.85 for negative, 0.87 for neutral (referred to as natural), and 0.85 for positive, indicating the logistic regression model's capability in distinguishing between the sentiment classes effectively.

To improve predictive accuracy, we investigated the use of ensemble classifiers which involved the combination of Logistic Regression, Naive Bayes, and Linear Regression. By employing these methods, we were able to effectively capture

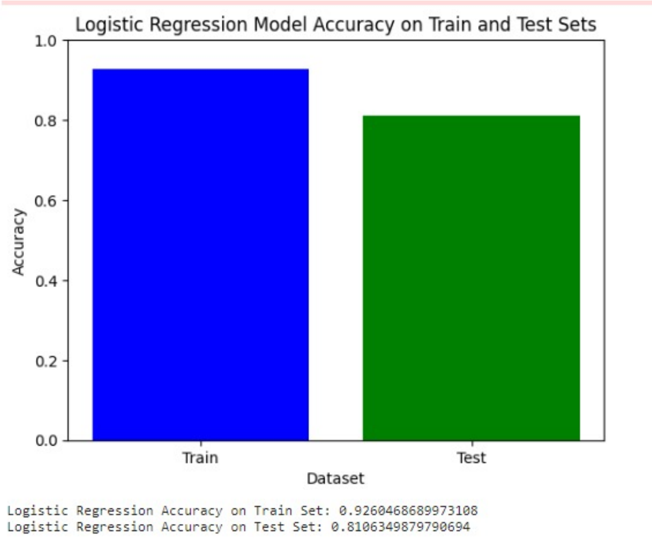


Fig. 12. multi-class Logistic Regression Model Accuracy on Train and Test Sets. The blue bar represents the accuracy of the training set, and the green bar represents the accuracy of the test set.

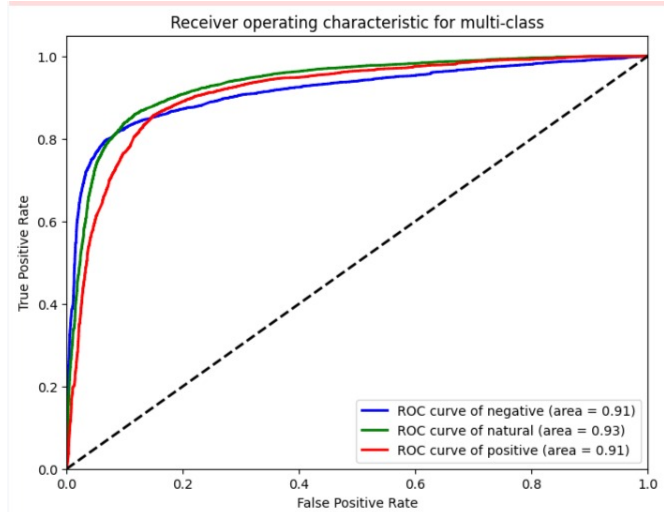


Fig. 13. Multiclass ROC Curves for Logistic Regression Model: This figure illustrates the Receiver Operating Characteristic (ROC) curves for a multi-class sentiment classification using logistic regression. It shows the model's performance across three classes: negative (AUC = 0.91), neutral—referred to as natural in the legend—(AUC = 0.93), and positive (AUC = 0.91). The high area under the curve (AUC) values indicate the model's strong discriminatory ability for each sentiment class.

various viewpoints, resulting in enhanced sentiment analysis results.

We utilized ParamGridBuilder to perform hyperparameter tuning for Logistic Regression in the context of hyperparameter tuning. The grid search involved optimizing parameters like maxIter, regParam, and elasticNetParam to enhance the accuracy of the model. By employing CrossValidator with the MulticlassClassificationEvaluator, we were able to identify the optimal combination of hyperparameters. The resulting Logistic Regression model achieved an accuracy of 84.6 percent on the test set, demonstrating the effectiveness of hyperparameter tuning in improving model performance.

*3) multi-class logistic Regression :* Although our initial analysis involved ROC analysis for Logistic Regression, it is worth mentioning that ROC curves are typically used in binary classification situations. In the case of multi-class classification, alternative evaluation metrics like precision, recall, and F1-score offer more valuable information about the model's performance for each sentiment class.

*4) Model Evaluation with ROC Curves :* To summarize, our research offers a thorough examination of sentiment analysis in Arabic company reviews. It emphasizes the importance of TF-IDF and demonstrates the effectiveness of different machine-learning models. The utilization of ParamGridBuilder and hyperparameter tuning highlights the significance of refining models for optimal results in sentiment analysis tasks. To gain a more comprehensive understanding of model performance, it would be advantageous to explore precision, and recall.

## V. Conclusion

In our research on sentiment analysis of Arabic company reviews, we undertook a comprehensive examination of a dataset comprising over 108,000 reviews. The initial phase of our study involved data preprocessing, where we addressed missing values, removed references, and HTML tags, and converted numerical values into their Arabic verbal equivalents. This process refined our dataset to 75,000 reviews, making it more suitable for machine learning analysis.

Our analysis revealed a predominant presence of positive sentiments (59.9%), followed by neutral (35.4%) and negative sentiments (4.7%). These findings underscore the general customer sentiment trends in the Arabic-speaking market. To tackle the inherent challenges of Arabic textual data, such as language nuances and dialect variations, we implemented normalization and feature engineering techniques. These methods were crucial in enhancing the quality and interpretability of our data.

For sentiment analysis and prediction, we employed two machine learning models: logistic regression and Naive Bayes. The logistic regression model exhibited high accuracy in emotion recognition and rating prediction, proving its effectiveness in analyzing customer sentiments. On the other hand, the Naive Bayes model, while slightly less accurate, demonstrated strong capabilities in pattern recognition. We assessed the reliability and generalization capabilities of these models through accuracy metrics across training and test sets.

This study contributes to the field of sentiment analysis in the Arabic language by demonstrating the effectiveness of machine learning models in understanding customer sentiment. Mastering these models empowers businesses to gain deeper insights into customer opinions, discern patterns, and make data-driven decisions.

However, our research is not without limitations. The process of normalizing and refining the dataset may have led to the exclusion of certain linguistic subtleties intrinsic to the

Arabic language. Future research could explore the integration of more advanced NLP techniques, like deep learning, to capture these nuances better. Additionally, expanding the dataset to include a more diverse range of Arabic dialects could enhance the model's applicability and accuracy across different Arabic-speaking regions. In conclusion, our study paves the way for further exploration in the field of Arabic sentiment analysis, offering a foundation for businesses to leverage customer feedback effectively in their strategic decision-making processes.

## REFERENCES


[1] J. C S. Alyami, A. Alhothali, and A. Jamal, "Systematic literature review of Arabic aspect-based sentiment analysis," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, pp. 6524–6551, 2022, doi: 10.1016/j.jksuci.2022.07.001

[2] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," Future Generation Computer Systems, vol. 112, pp. 408-430, 2020, doi: 10.1016/j.future.2020.05.034.I. S. Jacobs and C. P. Bean, "Fine particles, thin films, and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] S. M. Sherif, A. H. Alamoodi, O. S. Albahri, and S. Garfan, "Lexicon annotation in sentiment analysis for dialectal Arabic: Systematic review of current trends and future directions," Information Processing and Management, vol. 60, 2023, doi: 10.1016/j.ipm.2023.103449.

[4] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," Journal of King Saud University – Computer and Information Sciences, vol. 35, 2023, doi: 10.1016/j.jksuci.2023.101570.

[5] Li, L., Johnson, J. B., Aarhus, W., & Shah, D. (2022, January 1). Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. Computers & Education. https://doi.org/10.1016/j.compedu.2021.104354

[6] Tusar, M. T. H. K., & Islam, M. T. (2021, September 14). A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. 2021 International Conference on Electronics, Communications and Information Technology (ICECIT). https://doi.org/10.1109/icecit54077.2021.9641336

[7] . A. Al Shamsi and S. Abdallah, "Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects," Journal of King Saud University – Computer and Information Sciences, vol. 35, 2023, doi: 10.1016/j.jksuci.2023.101691

[8] R, B. K. J. S. G. G. V. K. (2021, May 2). Sentiment Analysis for E-Commerce Products Using Natural Language Processing. https://www.annalsofrscb.ro/index.php/journal/article/view/4269

[9] Kavitha, S., & Badre, R. (2020, October 1). Towards a Hybrid Recommendation System On Apache Spark. https://doi.org/10.1109/indiscon50162.2020.00066

[10] Testas, A. (2023, January 1). Natural Language Processing with Pandas, Scikit-Learn, and PySpark. Apress eBooks. https://doi.org/10.1007/978-1-4842-9751-3_14

[11] Ying, Y., Mursitama, T. N., S., L. (2021, February 1). Effectiveness of the News Text Classification Test Using the Naïve Bayes' Classification Text Mining Method. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1764/1/012105

[12] Zhang, L., Jiang, L., Li, C., & Kong, G. (2016, May 1). Two feature weighting approaches for naive Bayes text classifiers. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2016.02.017

[13] Dedeturk, B. K., & Akay, B. (2020, June 1). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106229

[14] Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022, January 1). Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifiers. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2021.12.187

[15] Wadud, M. A. H., Kabir, M. M., Mridha, M. F., Ali, M., Hamid, M. A., & Monowar, M. M. (2022, November 1). How can we manage Offensive Text in Social Media - A Text Classification Approach using LSTM-BOOST. *International Journal of Information Management Data Insights*. https://doi.org/10.1016/j.jjimei.2022.100095

[16] Yu, T., Wang, L., Zhang, W., Xing, G., Han, J., Li, F., & Cao, C. (2022, January 1). Predicting Phenotypes From High-Dimensional Genomes Using Gradient Boosting Decision Trees. *IEEE Access*. https://doi.org/10.1109/access.2022.3171341

[17] Lainder, A. D., & Wolfinger, R. D. (2022, October 1). Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.12.003

[18] Helmy, M., El-Daydamony, E., Mekky, N., Elmogy, M., & Soliman, H. (2022, June 15). Predicting Parkinson disease related genes based on PyFeat and gradient boosted decision tree. *Scientific Reports*. https://doi.org/10.1038/s41598-022-14127-8

[19] Asaly, S., Gottlieb, L. A., Inbar, N., & Reuveni, Y. (2022, June 12). Using Support Vector Machine (SVM) with GPS Ionospheric TEC Estimations to Potentially Predict Earthquake Events. *Remote Sensing*. https://doi.org/10.3390/rs14122822

[20] Gökten, E. S., & Uyulan, A. (2021, January 1). Prediction of the development of depression and post-traumatic stress disorder in sexually abused children using a random forest classifier. *Journal of Affective Disorders*. https://doi.org/10.1016/j.jad.2020.10.006

[21] Hediger, S., Michel, L., & Näf, J. (2022, June 1). On the use of random forest for two-sample testing. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2022.107435

[22] Sallab, A. A. A., Hajj, H., Badaro, G., Baly, R., El-Hajj, W., & Shaban, K. (2015, January 1). Deep Learning Models for Sentiment Analysis in Arabic. https://doi.org/10.18653/v1/w15-3202l.251 ... H., Badaro, G., Baly, R., El-Hajj, W., &Shaban, K. (2015, January.

[23] Abdul-Mageed, M., Diab, M., & Kübler, S. (2014, January 1). SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language. https://doi.org/10.1016/j.csl.2013.03.001

[24] Heikal, M., Torki, M., & El-Makky, N. M. (2018, January 1). Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2018.10.466

[25] Alyami, S., Alhothali, A., & Jamal, A. (2022, October 1). Systematic literature review of Arabic aspect-based sentiment analysis. Journal of King Saud University - Computer and Information Sciences.https://doi.org/10.1016/j.jksuci.2022.07.001

[26] Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015, January 1). Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*. https://doi.org/10.1504/ijsnm.2015.072280

[27] Alwakid, G., Osman, T., and Hughes-Roberts, T. (2017, January 1). Challenges in Sentiment Analysis for Arabic Social Networks. Procedia Computer Science. https://doi.org/10.1016/j.procs.2017.10.097

[28] Li, B., Zhu, L., Li, T., and Liu, J. (2013, December 1). Sentiment analysis of Arabic social media content: a comparative study. https://doi.org/10.1109/icist.2013.6747520

[29] Duwairi, R., & Qarqaz, I. (2014, August 1). Arabic Sentiment Analysis Using Supervised Classification. https://doi.org/10.1109/ficloud.2014.100

[30] Oueslatı, O., Wang, Z., HajHmida, M. B., & Ounelli, H. (2020, November 1). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems.https://doi.org/10.1016/j.future.2020.05.034

[31] Farha, I. A., & Magdy, W. (2021, March 1). A comparative study of effective approaches for Arabic sentiment analysis. Information Processing and Management. https://doi.org/10.1016/j.ipm.2020.102438.

[32] Duwairi, R. (2015, April 1). Sentiment analysis for dialectical Arabic. https://doi.org/10.1109/iacs.2015.7103221

[33] Elnagar, A. (2016, November 1). Investigation on sentiment analysis for Arabic reviews. https://doi.org/10.1109/aiccsa.2016.7945623

[34] El-Masri, M., Altrabsheh, N., Mansour, H., & Ramsay, A. (2017, January 1). A web-based tool for Arabic sentiment analysis. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2017.10.092

[35] Mohamed, A. (2022, May 9). SVM and Naive Bayes for Sentiment Analysis in Arabic. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-1631367/v1

[36] Hawalah, A. (2019, September 26). A FRAMEWORK FOR ARABIC SENTIMENT ANALYSIS USING MACHINE LEARNING CLASSIFIERS. https://hal.science/hal-02300717


.