International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Evaluating Arabic Lexicon Structure with Machine Learning Techniques

[1]Aya Mohammed Abdul-Samad, [2]Salma A. Mahmood

Submitted: 03/11/2023 **Revised**: 23/12/2023 **Accepted**: 02/01/2024

Abstract: In an age where big data reigns supreme, the advent of technological advancements has led to an exponential increase in digital textual data, particularly in the realm of the Arabic language. This surge has spurred the proliferation of electronic Arabic lexicons, which, while abundant, often lack the structured format required for effective use in Natural Language Processing (NLP) applications. This study seeks to bridge this gap by presenting a methodology for the extraction, structuring, and storage of lexicon data to render it suitable for NLP tools and technologies. Utilizing web scraping techniques, the study harvested lexical data from various online sources, transforming it into well-organized Excel files. The corpus encompasses a rich assembly of nouns (10,000 words), verbs (10,000 words), letters (70 words), adverbs (500 words), and pronouns (20 words), thus laying the groundwork for a comprehensive Arabic lexicon. Furthermore, the study leveraged several machine learning models to evaluate the structuring of the lexicon. The Support Vector Machine (SVM) and Random Forest models exhibited commendable accuracy (both at 0.85), underscoring the high quality of the data structuring process. Meanwhile, models like Logistic Regression and Multinomial Naive Bayes, despite lower precision and recall metrics, maintained moderate accuracy, which demonstrates the potential for further refinement.

Keywords*: Automatic Arabic lexicon generation; web scraping; web information extraction, ML.*

1. Introduction

The Arabic language is famous for its rich history and complex linguistic characteristics, and the Arabic language is the fourth most widely used language on the Internet due to the spread of Muslims, the Holy Qur'an, and Arabic culture [1]. Therefore, the need for an integrated and comprehensive Arabic lexicon arose, and this is what technological development and the spread of textual data led to, which led to the spread of existing electronic lexicons. On the Internet, but in its unstructured or semi-structured form, it is difficult to use it in natural language processing. Therefore, the need arose to build a structured and comprehensive Arabic lexicon ready for use by means of natural language processing. Here, the need arose to use this data found on the Internet and use it to create a structured and comprehensive Arabic lexicon. The method of web scraping is through scraping data from web pages by dealing with the basic language of building websites Hypertext markup language (HTML) [2].

Web scraping deals with the textual data present on the lexicon's website in a dynamic way to scrape the data, structure it, and save it in Excel form with five entries, each entry according to its division in the Arabic language (nouns, verbs, letters, adverbs, pronouns), and then a structured file is created consisting of these entries.

In [3], this research consists of building an automatic Arabic semantic lexicon by choosing an Arabic lexicon and providing it with morphological and semantic information. The aim of this research is to work on a simple method to extract lexical entries and relationships between words. In [4], the paper begins by discussing the standards for representing linguistic data. It also covers digital lexicographic resources and W3C's recommendations for linguistic data sources. It mentions initiatives like LLOD and the PanLex project. Finally, it highlights the lack of Arabic lexicons and previous efforts to represent Arabic morphological lexicons. In [5], this research focuses on conducting a semantic analysis of the data of the social networking site Twitter to collect structured information about traffic congestion. It includes the stage of data collection and analysis, revealing words and their meanings, and separating data about the place and time of congestion occurrence according to the publication of tweets on Twitter, which allows inferring the locations of traffic congestions through interaction, also with the geographical location of the tweets. In [6], this paper describes the creation of a specialized search engine that retrieves specific information. The process includes retrieval, extraction, presentation, and delivery. Data is collected from web pages and relevant information is extracted and organized for user-friendly presentation. The search engine allows efficient access to information. The paper introduces innovative techniques for each stage of the process, including representing templates, optimizing product indexing, and deploying data in a cloud-based database. Ethical and security considerations are also discussed. A practical example is provided, and suggestions for further research are given to improve cost efficiency and scalability. In summary, the paper details the steps and introduces innovative methods for building search engines,

1,2Department of Computer Information Systems, Computer Science and Information Technology University of Basrah Basrah, Iraq
itpg.aya.mohammed@uobasrah.edu.iq
salma.mahmood@uobasrah.edu.iq

emphasizing efficient resource usage and ethical considerations. In [7], the paper offers a detailed analysis of web scraping or web crawling, which involves extracting data from websites using software. Search engines like Google were among the first to develop scrapers that scan and index web pages. Web scraping mimics human browsing through web browsers or HTTP. It can convert unstructured web data to structured data. In [8], this work is on the impact of digitalization on the tourism industry and underscores the indispensability of collecting and analyzing data for a more profound comprehension of customers, competitors, and stakeholders. Additionally, the paper underscores the pivotal role of web scraping in gathering and retrieving data from the internet and furnishes information on the available tools and packages for web scraping. In [9], the present paper offers geographical research studies that employ the technique of web scraping, demonstrating its applicability across diverse research domains. The paper accentuates the benefits of web scraping, encompassing instantaneous access to geographically located data and cost-effectiveness in contrast to conventional data acquisition approaches. Additionally, it delves into the obstacles posed by web scraping, which encompass ethical and legal dilemmas pertaining to intellectual property rights, informed consent, and confidentiality, as well as technical challenges embracing data inconsistencies and partiality. Machine learning algorithms require large amounts of data to make automatic predictions, and data scraping solves this problem of collecting large amounts of data by collecting data from web pages and extracting data from them that are usually unstructured or semi-structured and transforming them into structured and organized data, which makes this The collected data is important for machine learning algorithms [10]. In [11], the importance of data scraping is defined as an indispensable tool for electronic commercial companies. Data scraping plays an important role in the process of collecting information, which enables important decisions that depend on data scraped from the web.

The paper is arranged as follows. Section 2, introduces the methodology of how to build an Arabic lexicon and ML models. Section 3, shows the results and the discussion of the proposed method. Finally, we address the conclusions in Section 4.

2. Methodology

The problem on which this study is building a comprehensive, integrated, and automated structured Arabic lexicon in the face of and overcoming many challenges. Through this proposed framework, we will review the flow of how to create an Arabic lexicon automatically using a web scraping method from web pages effectively, without effort, in less time, and more accurately and free of noise. Through this section, we will review the mechanism of data collection and processing, the creation of the lexicon, and finally how to use a mechanism to display this data in an easy-to-use and understandable way for the user.

A. Data Collection Stage

1) *Data sources:* They refer to the resources from which data is collected. The Almaany lexicon was used in this study because it is a comprehensive and integrated lexicon and is subject to many of the basic standards that were relied upon in building the lexicon, which is comprehensive, contains grammatical and morphological information, contains lexical-contextual semantic information, accuracy, continuous updates, integration with technology, intellectual property, and ethics.

2) *Generate entries seeds:* Preparing a group of words that serve as seeds for the purpose of scraping their information from the Almaany lexicon. Almaany lexicon is a semi-structured lexicon and its pages are dynamic, that is, they display content private to each user, and you need to access the words through these seeds. Each seed has its own page. These seeds were brought from the Alarabi Aljamaa Lexicon, which is a static web page. Only the words were taken using a web scraping method and collected in Excel to later send each word separately to the Almaany lexicon for completely scraping its data.

B. Data Processing Stage

The data processing stage is the most important stage in which this methodology was built, which is the method of extracting data from the lexicon webpage and converting it into structured and understandable data by using web scraping. At this stage, after making a request to the server to fetch the information of the web page of lexicon through the website URL. The website is structured based on HTML. On the basis of the dynamic page, the best way to access the lexicon's database is through the search port of the lexicon's website. One of the keywords is sent to the designated part of the search port, and one word after another is searched for, then processed and scraped according to its classification into entries.

In Fig. 2, after placing the seed in the search field, the word will go through several processing operations to scrape it.

1) Is found? When the word is sent to the website, the word is searched for on the lexicon website. If the word is found in the database of the lexicon website, this word passes several tests, including filtering and removing noise of the word, which is the treatment that distinguishes the word whether it is a noun, a verb, a letter, a pronoun, or an adverb.

2) Is suggested? If the seeds are not found in the lexicon, the lexicon gives words close to the actual word, i.e. a word suggested by the lexicon itself whether it is synonymous, opposite, or even a word that is close in letters, so we take

this word and scrape the data for the proposed word. This process is done after ensuring that the word does not exist.

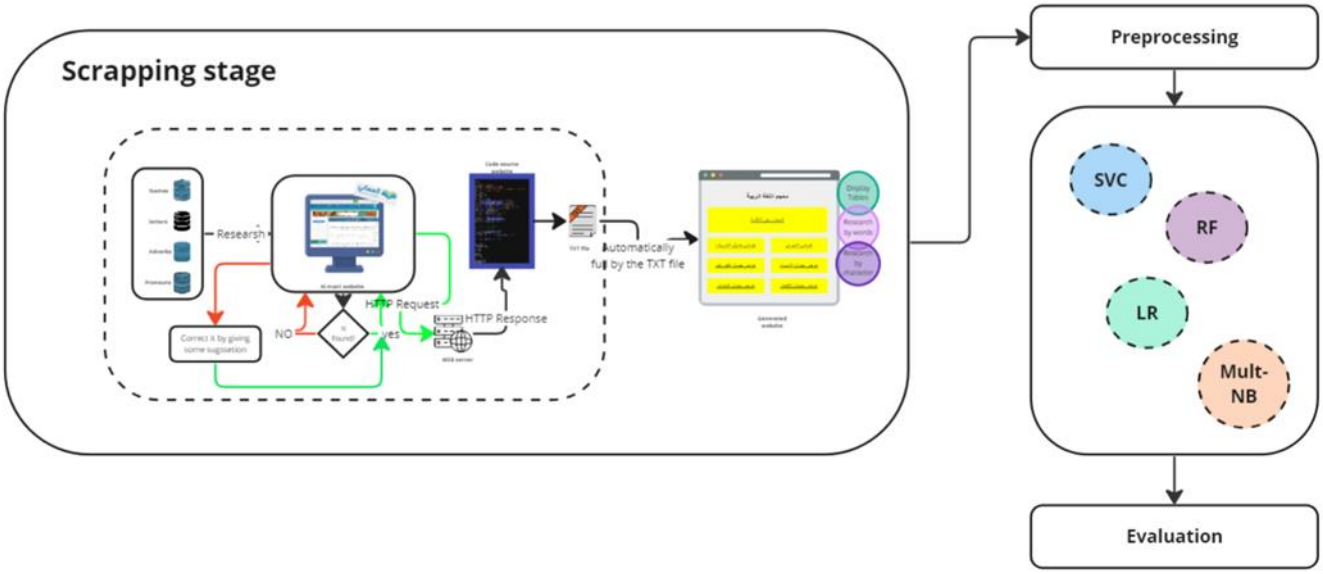

Fig. 1. The proposed framework.

3) The word not found/suggested: If the word does not exist in the lexicon and there is no suggestion for the word in this case, the system converts it to the next word in the Excel file for seeds of words and it goes through the same process to perform the scraping. When a word is identified and found, whether it exists or is suggested, it passes to the next stage, whereby the word is Scraped and then classified according to the entries of the Arabic language and its parts of speech, and it is preserved in the table for each word according to its classification. Every word that passes through the scraping is classified according to the entries, which are the five entries of the Arabic language (nouns, verbs, letters, adverbs, pronouns), following the method of filtering the website according to words. Every word found on the non-Arab website is classified according to its entry, with the information related to the word being scrapped.

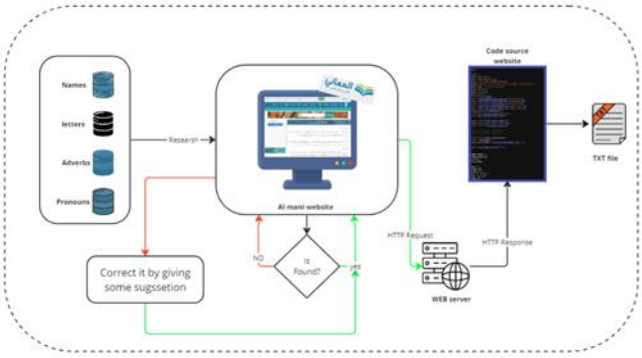

Fig 2 Web scraping process.

4) Lexicon creation: The structured Arabic lexicon is of great importance in the field of computer science and its development due to the scarcity or non-availability of an integrated structured Arabic lexicon that includes all aspects of the Arabic language. For building a lexicon, it is necessary to ensure the accuracy of the entries created according to the word Classification. Five entries were collected, including (nouns, verbs, letters, adverbs, and pronouns).

a) The indexing mechanism: For collecting entry files to create a lexicon through indexing to facilitate access to words. The mechanism used is to examine all words one by one to confirm the class to which they belong and to put the class number of the word in it. As in Fig. 3, the words and their position are listed in terms of the table in which the word is located, and its row number is given for all words in all tables.

Fig 3 Indexing.

b) Joint indexing mechanism: The common indexing mechanism means that there are words in the Arabic language that share a noun, a verb, and an adverb at the same time. It is known that the Arabic language is a morphological language, and one word can be a group of words that differ

according to the parts of speech. This problem was solved through joint indexing, in which the word is shared between the noun and the verb, or between the noun and the adverb, by developing a simple mechanism, which is sharing the same word, but finding it in two different entries, each with its own row number and entry to which it belongs as in Fig. 4. The number (1) was placed in the entry name field so that we know that this word belongs to this entry, as well as the number of the field in which the word is located, in relation to the table in which it is found.

Fig 4 Cognates indexing.

C. End-User

Creating an interface for the lexicon has many benefits, including that the interface provides an easy and quick way for users to interact with the lexicon. It allows individuals, learners, and those familiar with the language to search and access the content of the lexicon with ease without the need for any technical skills or user manual, in addition to an effective search that allows users to use the search mechanism. It can easily display the required words, especially when dealing with a lexicon that includes thousands of words. The interfaces also provide an organized presentation of the lexicon for ease of understanding. It can also be easily viewed on any computer or mobile device if it has been uploaded to the Internet. The interface also allows adding any elements required to develop the lexicon. Interfaces are a link between the user and programming data. The interface makes ease of use and ease of interaction with the lexicon of great value. Flask framework is used for building user interfaces using Python language with easy-to-understand code writing. Flask offers a code library for website creation that eliminates the need to build from scratch. The simplicity of Flask results in a lighter framework. Flask is a Python and easy-to-use web framework with a low learning curve. Its explicit design enhances readability [12]. Also uses a trie data structure is also known by other names such as radix tree, prefix tree, and digital tree. Derived from retrieval, a trie is a multi-way tree data structure implemented for the purpose of storing strings over an alphabet for searching into lexicon entries and search into indexing files [13].

1) Server block: A server ban, also known as an IP block, may occur. This block occurs when more requests are sent than the normal rate and in a very short time, and this rate is much shorter than the normal use of a successive series of requests. It looks like he requested a bot and wants to hack it. There are many solutions to avoid blocking, using the method of delaying requests between one request and another and following human behavior in the process of searching and scraping. Delaying avoids blocking, but it takes a long time to scrape all the data [14].

2) Legal web scraping: When scraping the web, you must take into account legal and ethical issues and not violate the laws when the website does not allow scraping. If scraping is done illegally, such as hacking the site or using illegal tools, the perpetrator will be subject to legal punishment. Beware of violating copyright. Blocking by the server may also lead to a legal problem, even if the website allows scraping because it is considered successive attacks on the server for the purpose of hacking [15].

D. Preprocessing Machine Learning Stage

Our preprocessing stage begins with the acquisition of data from various CSV files, each containing different parts of speech like adverbs, letters, names, pronouns, and verbs. These files are loaded into separate Pandas Data Frames. We then define a function to extract words from a specified column which could be either 'Word' or its Arabic equivalent 'الكلمة', depending on the file's structure. This function also tags each word with the file type it came from.

After extracting the relevant columns from each Data Frame, we merge them into a single Data Frame, creating a comprehensive collection of words along with their associated file names. This combined dataset serves as the foundation for further analysis and visualization.

To visualize the distribution of words across the different files, we generate a pie chart that displays the percentage contribution of each word type to the total dataset. This helps in understanding the composition of our data at a glance.

Next, we prepare our data for machine learning. We split the words into training and test sets, ensuring a mix of all word types in both. This is crucial for the model to learn and later predict accurately on unseen data.

Then, we employ a Label Encoder to convert the file names into a format that can be understood by machine learning algorithms—numerical labels. This process is done for both the training and the test sets.

Finally, we use Tfidf Vectorizer to convert our textual data into a numeric form, specifically into TF-IDF features, which are essential for modeling. TF-IDF stands for Term Frequency-Inverse Document Frequency, a numerical statistic that reflects how important a word is to a document in a collection or corpus. It helps in understanding the relevance of words in the context of their usage across different documents.

This completes our preprocessing stage, where the data is now ready for the evaluation phase, where various machine learning models will be trained and tested on this processed data.

E. Machine Learning Models

The models we used in our classification tasks represent a range of supervised machine learning algorithms, each with its unique strengths and approach to learning from data.

We started with the Support Vector Classifier (SVC)[17], a powerful algorithm well-suited for high-dimensional spaces, which is ideal for text classification problems. By fitting it to our TF-IDF transformed data, we trained the model to find the optimal hyperplane that separates the different word categories.

Next, we employed a Random Forest Classifier, which builds multiple decision trees and merges them together to get a more accurate and stable prediction. This ensemble method is good for dealing with unbalanced datasets because it provides a way of assessing feature importance.

Following that, we implemented a Gradient Boosting Classifier, another ensemble technique that builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. This model is typically more sensitive to overfitting if the data is noisy.

We also included Logistic Regression in our suite of models. Despite its name, Logistic Regression is a classification algorithm, not a regression algorithm. It is simple but effective and often serves as a baseline for binary classification problems.

Lastly, we utilized a Multinomial Naive Bayes classifier. This model is particularly well-suited for classification with discrete features (e.g., word counts or TF-IDF features for text classification). It assumes independence between the features, and it's a good choice when dealing with text data.

For each model, we predicted categories for the test set and calculated the classification report, which includes precision, recall, and F1-score. Precision measures the accuracy of positive predictions. Recall, also known as sensitivity or true positive rate, measures the fraction of positives that were correctly identified. The F1 score is the harmonic mean of precision and recall and gives a combined idea of the two metrics.

We also computed the overall accuracy for each model, which is the percentage of correct predictions out of all predictions made. To visualize the performance metrics, we generated bar plots for precision, recall, F1-score, and accuracy, allowing us to compare the efficacy of each model at a glance.

By appending the performance metrics of each model to a Data Frame, we created a consolidated overview of our models' performances, which is instrumental in the selection of the best-performing model based on the given dataset and problem statement.

3. Results and Discussion

In this section, we will delve into the essence of our endeavors in this work; the fruits of the work of building an automatically structured Arabic lexicon using the web scraping method. We have so far delved into the complexities of the Arabic language, as well as into the challenges of extracting data from the web (Linguistic complexity and morphological diversity, lexicon structure and format, resources for rare and technical terms, handling new vocabulary, the cultural importance and diversity of dialects, semantic ambiguity and contextual nuances, …etc.) and now begin the journey of exploring data that is collected and structured [16]. The Arabic language is a rich, morphological, and complex language that contains many words and is subject to many challenges and complexities. The Arabic language is divided into several entries: nouns, verbs, letters, and auxiliary words that give the sentences and words of the Arabic language coherence and suitability. The entries of the Arabic language are divided into five entries, including (nouns, verbs, letters, pronouns, and adverbs), all entries will be detailed in this section.

F. Nouns Entrance

Nouns are considered one of the main and important parts of the Arabic language. It has an essential role in speech, in constructing sentences, and in conveying meaning. Nouns in the Arabic language represent the names of people, places, and concepts. 10,000 words of nouns were collected with their information in a structured manner. Among the most important things that were collected was the collection of information related to the nouns (word, meaning, context, example, root, participle noun, passive participle noun, adjective, synonyms, antonyms, singular, plural, masculine, and feminine). This information is scraped from the web and structured under names in the Arabic language.

G. Verbs Entrance

The verb is an important part of the Arabic language, as is the case with nouns because they are complementary to the Arabic language and speech. Verbs play an essential role in constructing sentences and conveying meaning. Verbs in Arabic describe an event or situation. 10,000 verb words were collected with their information in a structured form from the information related to the verbs (word, meaning, context, example, root, subject, object, past, present, intransitive verb, and transitive verb). All of this information is scraped from the web and structured under the name of verbs in the Arabic language.

H. Letters Entrance

Letters, which play a crucial role in the Arabic language, serve the purpose of shaping speech and attributing meaning to it. They are interconnected with nouns and verbs to build

sentences. The true meaning of the letters remains hidden until they are accompanied by nouns or verbs. 70 letters were collected with their information in a structured form from information related to the letters (word, meaning, context, example, letters that enter nouns, letters that enter verbs, letters that enter both a noun and a verb, and type of letter). All of this information is scraped from the web and its structure under the name of letters in the Arabic language.

I. Adverbs Entrance

Adverbs are words that modify verbs or adjectives to provide more information to a sentence or speech, such as where, when, and how it happened. They indicate time and place or describe a situation. Adverbs are often derived from adjectives. 500 adverbs were collected with their information in a structured form from the information related to the letters (word, meaning, context, example, singular, plural, feminine, masculine, synonym, opposite, and type of adverb). All of this information is scraped from the web and structured under the name of adverbs in the Arabic language.

J. Pronouns Entrance

Pronouns in Arabic, as in other languages, are a word used to replace nouns to avoid repetition and make the sentence less complex, depending on gender, number, case, and grammatical case. 20 pronouns were collected with their information in a structured form from information related to the letters (word, meaning, context, example, singular, plural, feminine, masculine, type of pronoun). All of this information is scraped from the web and structured under the name of pronouns in the Arabic language.

Table 1 Result Summarization

Category	**Number Collected**	**Key Information Collected**	**Example**
Nouns	10,000	Word, Meaning, Context, Example, Root, Participle noun, Passive participle noun, Adjective, Synonyms, Antonyms, Singular, Plural, Masculine, Feminine	Word: فعل (Verb), Meaning: Action or state, Root: ف-ع-ل, Singular, Masculine
Verbs	10,000	Word, Meaning, Context, Example, Root, Subject, Object, Past, Present, Intransitive verb, Transitive verb	Word: يفعل (To do), Meaning: To perform an action, Root: ف-ع-ل, Present tense, Transitive verb
Letters	70	Word, Meaning, Context, Example, Letters that enter nouns, Letters that enter verbs, Letters that enter both noun and verb, Type of letter	Word: ف, ع, ل (F, 'A, L), Meaning: Root letters of verbs, Type of letter: Root letters
Adverbs	500	Word, Meaning, Context, Example, Singular, Plural, Feminine, Masculine, Synonym, Opposite, Type of adverb	Word: بشكل فعال (Effectively), Meaning: In an effective manner, Type of adverb: Manner
Pronouns	20	Word, Meaning, Context, Example, Singular, Plural, Feminine, Masculine, Type of pronoun	Word: هو يفعل (He does), Meaning: He is performing an action, Type of pronoun: Personal, Singular, Masculine

All of these entries are saved in Excel format, a user interface has been created for each of these Excel entries, and a lexicon has been created from these entries.

The main goal of this project is the construction of a new, valuable dataset accompanied by the development of a website specifically designed to assist in data scraping. The vision behind this initiative is to create a resource that is both comprehensive and user-friendly, significantly contributing to various fields requiring such data. However, the project is not without its challenges. One of the primary concerns is the potential for inaccuracies that might arise from automated analysis processes. These processes, while efficient, often struggle with nuances and complexities, particularly when dealing with varying dialects. This aspect is crucial as dialectal variations can significantly impact the accuracy and applicability of the data.

Moreover, there's a recognition of the dynamic nature of data and its contexts. To remain relevant and useful, the dataset and the scraping tools will require constant updates. This necessity stems from the ever-evolving nature of language, usage, and the contexts in which data is applied. Therefore, a significant part of the project will be dedicated to developing mechanisms for regular updates and maintenance, ensuring that the dataset remains a valuable resource for its users.

In summary, while the project is ambitious and holds the promise of creating a significant resource, it is also grounded in a realistic understanding of the challenges involved in such an endeavor. The aim is to navigate these challenges effectively to provide a dataset and scraping tools that are not only useful in the present but continue to be so in the future.

K. Machine Learing Results

The results of the ML models illustrate the performance metrics for each classifier used in the analysis. Let's go through each figure:

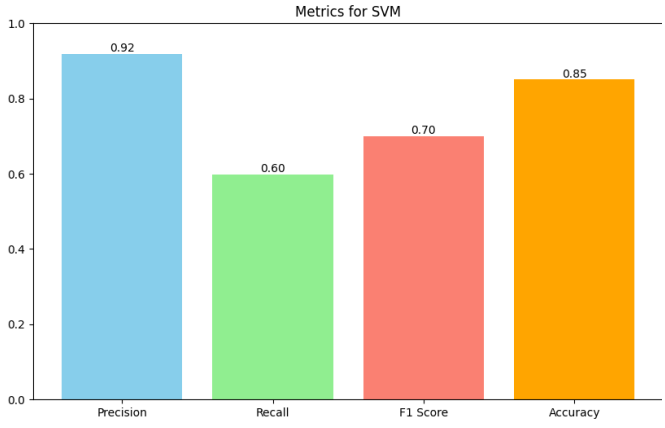

Fig 5 SVM Results.

Fig 5 shows the performance metrics for the Support Vector Machine (SVM) model. It has a high precision of 0.92, indicating that it is very effective at producing relevant results. However, the recall is 0.60, which means it is less effective at identifying all relevant instances. The F1 Score, which balances precision and recall, is at 0.70, and the overall accuracy of the SVM model is quite high at 0.85.

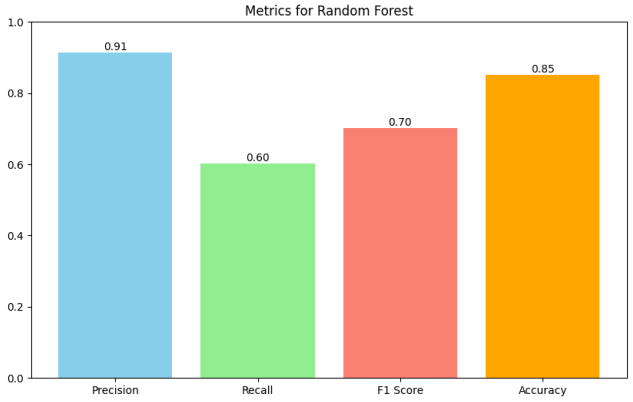

Fig 6 RF Results

Fig 6 presents the metrics for the Random Forest model. This model has a slightly lower precision than the SVM at 0.91 but matches in recall and F1 Score, with values of 0.60 and 0.70, respectively. The accuracy is also at 0.85, indicating that it performs almost identically to the SVM in terms of these metrics.

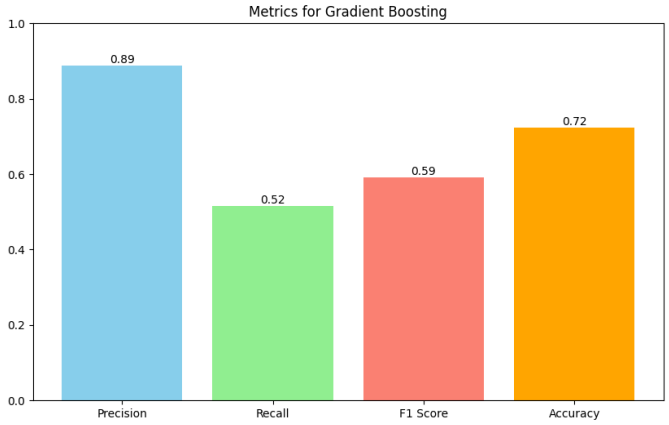

Fig 7 GB Results.

Fig 7 depicts the performance of the Gradient Boosting model. The precision here is slightly lower than the previous two models at 0.89. The recall drops significantly to 0.52, and the F1 Score is at 0.59, which suggests that Gradient Boosting is less balanced in terms of precision and recall compared to SVM and Random Forest. The accuracy is also lower at 0.72.

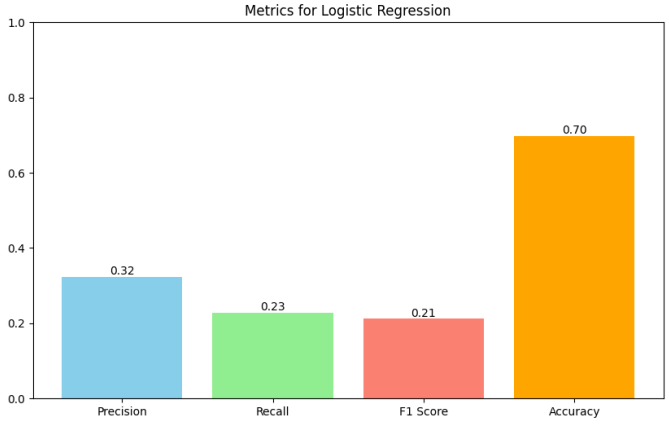

Fig 8 LR Results.

Fig 8 shows the Logistic Regression model's metrics. This model has a precision of 0.32, which is significantly lower than the other models, indicating a high number of false positives. The recall is at 0.23, and the F1 Score is 0.21, both of which are low, suggesting that Logistic Regression is not performing well on this dataset. However, the accuracy is at 0.70, which is interesting given the low precision and recall, suggesting it may be more accurate for specific classes or that accuracy is not the best measure of performance for this model and dataset.

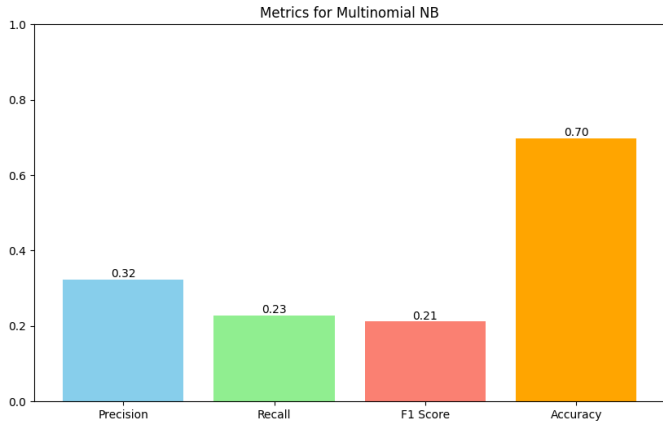

Fig 9 NB Results.

Fig 9 illustrates the metrics for the Multinomial Naive Bayes (NB) classifier. It mirrors the performance of the Logistic Regression with a precision of 0.32, recall of 0.23, and F1 Score of 0.21, which are all low. The accuracy is the same as Logistic Regression at 0.70, presenting the same discrepancies in performance as noted with Logistic Regression.

In conclusion, while SVM and Random Forest show a balanced and robust performance across all metrics, Gradient Boosting falls short on recall and F1 Score, and both Logistic Regression and Multinomial NB show lower performance across precision, recall, and F1 Score but maintain a decent accuracy rate. These results can inform which models are better suited for the dataset and objectives at hand.

L. Comparasion

The table presents a detailed comparison of five different machine learning models based on their performance metrics. Each model is evaluated on precision, recall, F1 score, and accuracy.

The SVM model exhibits the highest precision among the models at 0.918, suggesting that when it predicts a label, it is correct most of the time. Its recall rate is 0.590, which means it correctly identifies 59% of all relevant instances. The F1 score, which is the harmonic mean of precision and recall, is 0.690, indicating a good balance between precision and recall for the SVM model. The accuracy stands at 0.85, showing that the SVM correctly predicts the label 85% of the time, making it a strong performer in this lineup.

The RF model closely matches the SVM in terms of precision with a score of 0.913 and has a slightly better recall at 0.601. Its F1 score is 0.702, marginally better than that of the SVM, suggesting a slightly better balance between precision and recall. The accuracy of the Random Forest model is also 0.85, which means it performs on par with the SVM in overall prediction correctness.

The GB model shows a decrease in performance with a precision of 0.880. Its recall of 0.516 is lower than that of the SVM and Random Forest models, leading to an F1 score of 0.590. The accuracy for Gradient Boosting drops to 0.72, indicating that it is less effective in correctly predicting the overall labels compared to the SVM and Random Forest models.

LR has significantly lower precision at 0.32 and recall at 0.227, resulting in the lowest F1 score of 0.210 among the models. Despite these lower scores, its accuracy is 0.69, which suggests that while it may not be as precise or as good at recall as the other models, it still manages to predict the correct label with moderate success.

The Multinomial NB classifier has a precision similar to Logistic Regression at 0.323 and an identical recall of 0.227. Its F1 score is slightly higher at 0.211, and it shares the same accuracy rate of 0.69 with Logistic Regression. This indicates that Multinomial NB and Logistic Regression have comparable performances, with neither standing out in terms of precision or recall.[18]

In summary, while SVM and Random Forest lead the group with high accuracy and a strong balance of precision and recall, Gradient Boosting follows with moderate scores across the board. Logistic Regression and Multinomial NB, while sharing the lowest accuracy, demonstrate a need for improvements in their ability to provide precise and reliable classifications for this particular dataset.

Table 2 Result ML Summarization

Model	Precision	Recall	F1 Score	Accuracy
SVM	0.918	0.590	0.690	0.85
RF	0.913	0.601	0.702	0.85
GB	0.880	0.516	0.590	0.72
LR	0.320	0.227	0.210	0.69
Multi-NB	0.323	0.227	0.211	0.69

Figure 10 displays a bar chart of the accuracy for each model, offering a quick comparison. SVM and Random Forest are the top performers with accuracies of 0.85, followed by Gradient Boosting at 0.72, and both Logistic Regression and Multinomial NB at 0.69.

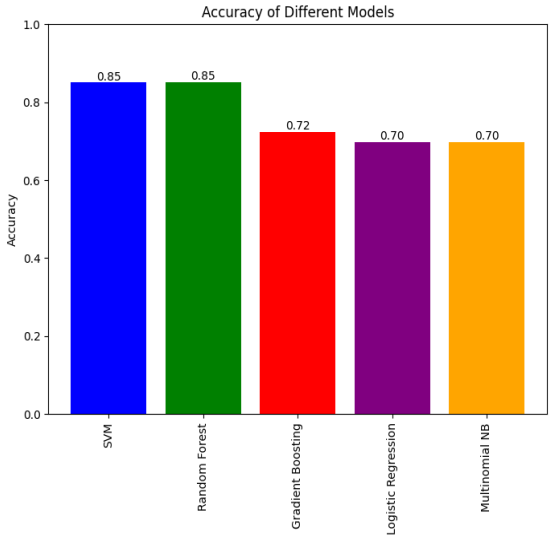

Fig 10 Accuracy Results.

This comparison reveals that while SVM and Random Forest are the best models for this dataset according to accuracy, the precision, recall, and F1 Score are crucial for understanding the models' performance comprehensively. Gradient Boosting performs moderately in accuracy but has a lower recall, and both Logistic Regression and Multinomial NB, despite having decent accuracy, perform poorly in precision and recall.

4. Conclusions

Our project to construct an automated and structured Arabic lexicon through web scraping has reached its conclusion. It has been an enlightening journey, filled with challenges that have pushed us to explore uncharted territories in the realm of Arabic linguistic studies. The task of assembling such a lexicon was no small feat, yet it has proven to be a fulfilling endeavor, highlighting the strides made in computer science and technology, particularly in the processing and analysis of the Arabic language.

The challenges we faced were multifaceted, stemming from the inherent linguistic intricacies and morphological diversity of Arabic. Yet, with strategic approaches, we not only navigated through these complexities but also celebrated the rich cultural and historical tapestry that the Arabic language weaves. The lexicon categorizes words into distinct sections: nouns, verbs, letters, pronouns, and adverbs, creating a comprehensive resource for language processing.

The final outcome is a structured Arabic lexicon, finely tuned and ready for automated processing in various natural language processing applications. The construction of this lexicon showcases the precision and efficiency achievable through automated web scraping and advanced data retrieval techniques. It bridges the linguistic heritage of the past with the technological advancements of the present, paving the way for a future where the Arabic language can be analyzed and utilized in new and innovative ways.

The results from the machine learning models add a quantitative testament to the lexicon's robustness. Models like SVM and Random Forest stood out with high precision and accuracy, reinforcing the quality of the lexicon's structure. Even as some models like Logistic Regression and Multinomial Naive Bayes showed room for improvement, they still maintained a reasonable level of accuracy. These outcomes underscore the lexicon's readiness for deployment in automated systems, promising a significant leap forward in Arabic linguistic studies and applications.

References

[1] O. Hamed, S. Salah, and A. A. Freihat, "ALRT: Cutting edge tool for automatic generation of arabic lexical recognition tests," in Proceedings of the Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022) co-located with ICNLSP 2022, Trento, Italy, 2022, pp. 43-49.

[2] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra, and K. R. Bodke, "Analysis of different web data extraction techniques," in 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, India, 2018, pp. 1-7.

[3] O. Batarfi, M. Dahab, and A. Ezz, "Building an arabic semantic lexicon for hajj," International Journal of Computer Applications, vol. 181, pp. 9-15, 2019.

[4] M. Jarrar and H. Amayreh, "An arabic-multilingual database with a lexicographic search engine," in Natural Language Processing and Information Systems, Cham, 2019, pp. 234-246.

[5] S. Subhan, E. Sediyono, and F. Farikhin, "The semantic analysis of twitter data with generative lexicon for the information of traffic congestion," Journal of Advances in Information Systems and Technology, vol. 1, pp. 45-54, 2019.

[6] A. Alexandrescu, "Optimization and security in information retrieval, extraction, processing, and presentation on a cloud platform," Information, vol. 10, p. 200, 2019.

[7] M. Khder, "Web scraping or web crawling: state of art, techniques, approaches and application," International Journal of Advances in Soft Computing and its Applications, vol. 13, pp. 145-168, 2021.

[8] R. Egger, M. Kroner, and A. Stöckl, "Web scraping," in Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications, R. Egger, Ed. Cham: Springer International Publishing, 2022, pp. 67-82.

[9] A. Brenning and S. Henn, "Web scraping: a promising tool for geographic data acquisition," arXiv preprint arXiv:2305.19893, 2023.

[10] S. D. S. Sirisuriya, "Importance of web scraping as a data source for machine learning algorithms - Review," in 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 2023, pp. 134-139.

[11] S. Shreekumar, S. Mundke, and M. Dhanawade, "Importance of web scraping in e- commerce business," NCRD's Technical Review, vol. 7, pp. 1-14, 2022.

[12] M. R. Mufid, A. Basofi, M. U. H. A. Rasyid, I. F. Rochimansyah, and A. rokhim, "Design an MVC model using Python for flask framework development," in 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 2019, pp. 214-219.

[13] S. S. Chawathe, "Data structures for ordered short character-sequences," in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, 2021, pp. 1370-1376.

[14] A. Haque and S. Singh, "Anti-scraping application development," in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 2015, pp. 869-874.

[15] A. Luscombe, K. Dick, and K. Walby, "Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences," Quality & Quantity, vol. 56, pp. 1023-1044, 2022.

[16] K. Shaalan, "A survey of arabic named entity recognition and classification," Computational Linguistics, vol. 40, pp. 469-510, 2014.

[17] Muaad, Abdullah Y., et al. "Arabic document classification: performance investigation of preprocessing and representation techniques." Mathematical Problems in Engineering 2022 (2022): 1-16.Alsaleem, Saleh. "Automated Arabic Text Categorization Using SVM and NB." Int. Arab. J. e Technol. 2.2 (2011): 124-128.