Machine Learning-Driven Framework for Sentiment Analysis of Tweets

Dhai Eddine Salhi
1. College of Engineering and Architecture
Al Yamamah University
Riyadh, KSA
d_salhi@yu.edu.sa

Majdi Rawashdeh
1. College of Engineering and Architecture
Al Yamamah University
Riyadh, KSA
ma_rawashdeh@yu.edu.sa
2. Dept. of Business Information Technology
PSUT Amman, Jordan

Awny Alnusair
3. Indiana University KOKOMO
Indiana, USA
alnusair@iu.edu

***Abstract*—Sentiment analysis has become increasingly pivotal across diverse fields such as politics, marketing, and social sciences, driven by the profound influence of public opinion on decision-making processes. This study advances sentiment analysis for Arabic, a language marked by its rich morphological structure and high surface shape variability, which poses significant challenges in text analysis. Employing machine learning models including Recurrent Neural Networks (RNN), Support Vector Machines (SVM), and Naive Bayes (NB), alongside techniques like TF-IDF and Word2Vec for text representation, this research innovatively incorporates a comprehensive root extraction from the Holy Quran to enhance feature extraction. An extensive dataset, enriched with an augmented list of 3,000 stopwords, supports the analysis. Our findings reveal a promising accuracy of 91% with RNN based on the Word2Vec technique, underscoring the effectiveness of integrating deep linguistic features in improving sentiment analysis for Arabic text.**

***Index Terms*—Arabic language, Standard Arabic Modern (ASM), sentiment analysis, Machine learning, Text mining, Classification.**

I. Introduction

In the past few years, the expansion of the internet has been significantly propelled by the advent of social networks, which leverage straightforward and universally accessible communication methods. Platforms such as Twitter, Facebook, and LinkedIn facilitate social engagement online, catering to the desires of individuals, enterprises, and organizations to share insights, solicit advice, and communicate in a swift and uncomplicated manner. The simplicity and cost-free access to these social networks have driven their widespread acceptance among the broader populace.

This surge in popularity has been paralleled by a massive influx of opinionated data on the web, with a daily production of 2.5 billion bytes of data. Remarkably, 90% of the world's data has been created in just the last few years alone.

Social networks' increased usage highlights a growing societal need for information. Typically, individuals are keen on gathering perspectives from others, unfamiliar to them, before making decisions or forming judgments. A Pew Research Center study reveals that 20% of users on social networks have revised their stance on issues based on content encountered online, underscoring the value of sentiment analysis across various domains, including marketing, psychology, healthcare, politics, road safety, and tourism.

Consequently, there is a burgeoning interest in automated sentiment analysis within both commercial and academic circles.

Sentiment analysis is intrinsically complex, encompassing tasks such as detecting subjectivity, assessing polarity and its intensity, pinpointing the specific entity being discussed along with its attributes, evaluating the sentiment per attribute, identifying the sentiment bearer, and analyzing how opinions about an entity evolve over time. Moreover, the mode through which opinions are expressed varies, including text, speech, and multimedia. This paper narrows its focus to the analysis of textual sentiment polarity.

While sentiment analysis in Indo-European languages, particularly English, has seen extensive research, studies on the Arabic language remain sparse due to its linguistic intricacies. Arabic's notable features include agglutination and morphological richness, which complicate word formation and result in a constrained lexicon. Indeed, Arabic sentences can sometimes be encapsulated in a single word, underscoring the language's complexity.

This study aims to examine the sentiment polarity of Arabic texts on Twitter, with an emphasis on the language's complexity. It navigates through three critical stages: data preprocessing, which involves cleaning the data, removing meaningless frequent words, and identifying word roots; modeling, which employs Word2Vec and TF-IDF models for transforming text into numerical vectors, alongside machine learning techniques for analysis and prediction; and finally, the presentation phase, where the outcomes are showcased.

The structure of this paper is as follows: Section II offers

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

a background on the Arabic language. Section III reviews existing research. The methodology is elaborated in Section IV. Section V describes the implementation and discusses the findings via a case study. The paper concludes in Section VI.

II. BackGround

The Arabic language, notable for its right-to-left script known as the Arabic script, boasts a complex morphology characterized by the use of roots and patterns to derive varied meanings. As a Semitic language, Arabic predominantly utilizes triconsonantal roots as the foundation for word formation. It is also an inflected language, where the form of words alters to express different grammatical aspects like tense, gender, and number. Arabic's alphabet consists of 28 letters, and its lexicon is extensive, with over 12 million words—surpassing the vocabularies of languages like English, with 600,000 words, French, with 150,000 words, and Russian, with 130,000 words [1]. There are three primary dialects or families within the Arabic language, each distinguished by unique linguistic traits. [2]

1) **Classical Arabic (CA):** it is generally used in religious writings, namely the Koran: the sacred book of Islam.
2) **Modern Standard Arabic (MSA):** is the official language in Arab countries It is the official language in the Arab countries and is mainly used in newspapers, educational books and scientific publications. These two registers, classical and modern, constitute Standard Literary Arabic.
3) **Dialectal Arabic (DA):** used by Arab speakers in everyday life.

All three variants of Arabic can co-exist in the same utterance. It is important to note that Classical Arabic is used mainly in religious contexts, while Modern and Dialectal Arabic are the most widely used in everyday life today: MSA is used in formal communication and DA in informal discussion.

A. Structure of the Arabic language

A word in Arabic can be formed from a base (the root) to which affixes and/or clitics may be added. As far as affixes are concerned, Arabic distinguishes between prefixes (agglutinated at the beginning of the base), suffixes (attached to the end of the base) and infixes (located in the middle of the base). As for clitics, we distinguish between proclitics located at the beginning of the word and enclitics located at the end. Fig.1 shows the general structure of a graphic word in Arabic [3].

A graphic word or "maximum word" essentially contains :

- **A base:** which represents the root of the word from which the agglutination is performed.
- **A minimal word:** which corresponds to the inflected form of the base obtained by the concatenation of prefixes and suffixes to this base.
- **A maximal word** : unit decomposable into proclitics, prefixes, base, suffixes and enclitics. It can also be analysed in proclitics, minimal word and enclitics.

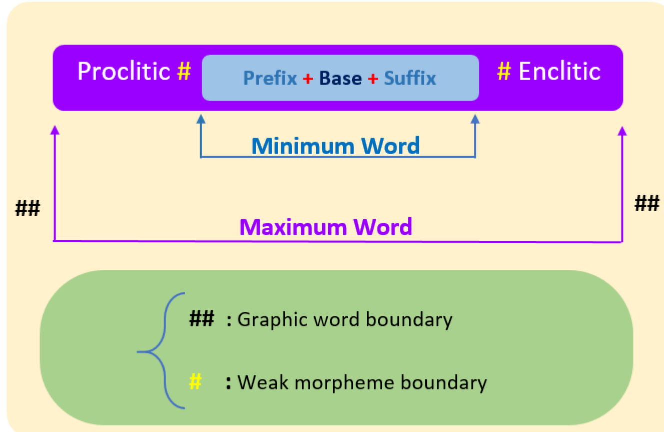

Fig. 1. General structure of a graphic word in the Arabic Language

III. Related work

Most of the research in the area of sentiment analysis has been carried out on European (especially English) and Asian (Japanese and Chinese) languages. However, very little work has been done on languages that are morphologically rich, such as Arabic.

A. Difficulties of the Arabic language

The Arabic language is complex to analyse because of the properties it has.

The following points explain the properties of the Arabic language and their impact on the analysis process [4].

1) Each country has its own version or dialect of Arabic. This means that there are different dialects of Arabic text available online that may contain different meanings. Leading to high complexities when analysing the sentiments.
   Example
   - Saudi dialect: توقعت اذا جات داريا بشوفهم كاملين بس لي للحين احس فيه احد ناقصه
   - Algerian dialect: ماعجبتنيش لبلاصة، و الماكلة عيانة بزاف
   - Tunisian dialect: البلاصة هاديك عجبتني برشا
2) The root of Arabic words can have several forms depending on the context.
3) Arabic words have the property of having the same spelling but with a different meaning depending on their punctuation. existence of words such as لكن can cause a sentence to have two opposite feelings at the same time.
   Example
   - المكان جميل لكن الخدمات فيه سيئة جدا لمْ تعجبني

Despite the many difficulties that the Arabic language contains, many challenges in the field of sentiment analysis of this language have been observed in recent years. In this section, we mention some works that we will discuss according to the approach we are dealing with:

B. Symbolic approach

Elhawary and colleagues were pioneers in creating an opinion lexicon for the Arabic language, aiming to develop a tool that classifies business stakeholders' opinions by analyzing online business reviews in Arabic. They categorized these reviews as positive, negative, or neutral, starting with a lexicon comprising over 600 positive, 900 negative, and 100 neutral words. While their evaluation showed promising precision, the recall rates were less satisfactory. Yet, their efforts enhanced local business search experiences in the Arabic-speaking Middle East.

Abdul-Mageed and associates introduced a novel manually annotated corpus of Modern Standard Arabic (MSA) along with a corresponding polarity lexicon. This corpus, consisting of annotated newswire documents, served as the basis for developing an automated system for tagging sentiment analysis in Standard Arabic (SSA). Their research explored the effects of different preprocessing levels on SSA classification, including the extent of stemming needed. They found that integrating language-specific features into morphological representation improved classification outcomes, emphasizing the utility of a polarity lexicon in enhancing performance.

In 2014, Habash and his team developed Ar-SenL, the first large-scale, publicly accessible sentiment lexicon for Standard Arabic. This lexicon was crafted using a mix of existing resources like English SentiWordnet (ESWN), Arabic WordNet (AWN), and the Standard Arabic Morphological Analyser (SAMA). They evaluated two methods for enhancing sentiment analysis—utilizing WordNet and English glosses, and discovered that methods based on English translations yielded better results than those relying solely on WordNet. Their findings suggested that the amalgamation of these resources could significantly boost sentiment analysis efficacy.

C. Digital Approach

Sghaier, M. A et al in [5], proposed an implementation of a sentiment analysis tool to detect the polarity of opinions from reviews extracted from e-commerce or product review websites in Arabic. They collected their corpus from various web resources such as Reviewzat, Jawal123, Jumia, and others. After annotating the corpus, they developed a small converter to detect emoticons and tested the impact of three types of stemmers (Arabic Stemmer, Arabic Light Stemmer, and Khoja stemmer) to perform the stemming task for the Arabic language. For classification, they used three types of algorithms: Support Vector Machines (SVM), Naïve Bayes (NB), and K-nearest neighbor (KPPV). They tested the performance using cross-validation and percentage split methods. The authors concluded that the best accuracies were achieved by using Naïve Bayes with a standard corpus plus the application of Khoja Stemmer or Light Stemmer, which gave 0.946 as precision and 0.939 as recall. Using Support Vector Machines with either the raw corpus or the corpus plus the application of Khoja Stemmer or Light Stemmer also gave 0.946 for precision and 0.939 for recall. In [6], DAHOU and others studied an Arabic Sentiment Classification scheme that evaluates and detects sentiment polarity from Arabic reviews and Arabic social media. They explored several architectures to build a high-quality Arabic word embedding model using a corpus of 3.4 billion words collected from a web corpus of 10 billion words. Additionally, they provided short, practical, and empirically informed procedures to investigate Arabic word embeddings and Convolutional Neural Networks (CNN) for sentiment classification to evaluate the quality of these word embeddings. The results demonstrate that performance improves with data quality, and high-dimensional vectors perform well on a large corpus. The CNN results for sentiment datasets show that initializing word vectors using pre-trained word embeddings leads to remarkable performance. They also indicate that a larger dataset generally achieves better performance in terms of model accuracy. In 2019, Abu Farha. I et Al in [7] proposed "Mazajak", an online system for Arabic sentiment analysis. The system is based on a deep learning model, which provides state-of-the-art results on many Arabic dialect data sets. After pre-processing the data, they used the word2vec for text representation as a learning model, they built a model based on CNN followed by LSTM. The CNN works as a feature extractor, where it learns the local patterns within the sentence and provides representative features. The LSTM works on the extracted features where the context and word order would be taken into consideration. In order to examine the effectiveness of their model they tested it on three different datasets: SemEval 2017, ASTD and ArSAS and compared it with the best existing reported performance on all three. They found that Mazajak outperformed all current state-of-the-art models on SemEval and ASTD. In addition, it achieved high performance on the ArSAS dataset including average recall and accuracy. In [8], Salhi D. E et al. presented their work on electronic reputation (E-reputation) analysis of a mobile operator in Algeria called "Djezzy" by analyzing its tweets. They conducted their study in two phases: the first phase involved pre-processing the tweets by cleaning them and removing noise, resulting in a set of 1510 tweets, including 840 positive tweets and 670 negative tweets. The second phase focused on detecting the company's reputation among Internet users in French, English, and Arabic languages using Machine Learning techniques and algorithms such as Logistic Regression and SVM. They found that both algorithms produced very similar results, with a slight advantage for SVM. Therefore, they chose SVM to continue their work to determine Djezzy's e-reputation compared to its two competitors in the Algerian market, Ooredoo and Mobilis. In [9], This paper presents a systematic review of Arabic Sentiment Analysis (ASA) literature. It aims to support research, identify future study areas, and ease researchers' access to relevant works. The review proposes a taxonomy for sentiment classification methods and highlights limitations in preprocessing, feature generation, and sentiment classification. Additionally, it suggests future research directions in both practical and theoretical aspects of ASA. In [10], Over the past decade, Arabic content on websites and social media has grown significantly, offering rich data for sentiment analysis.

Deep learning, especially recurrent neural networks (RNNs), has become a promising method for analyzing opinions due to its effectiveness with unstructured data. While 193 studies have applied RNNs to English sentiment analysis, only 24 have focused on Arabic sentiment analysis. These studies vary in focus, model performance, and dataset availability across dialects. This paper systematically examines the literature to evaluate and highlight key studies using RNNs for Arabic sentiment analysis. In [11], Arabic Language Sentiment Analysis (ALSA) operates across multiple linguistic levels, including phonetics, morphology, syntax, and semantics, but these levels often lack synergy. While sentiment analysis has been widely studied in English and Indo-European languages, Arabic's rich rhetorical and implicit meanings present unique challenges. This paper proposes a comprehensive strategy for ALSA, analyzing opinions and sentiments across all linguistic levels. The framework emphasizes the need for an annotated corpus to better understand Arabic sentences, from phonetics to rhetorical and metonymic expressions. In [12], This paper presents a comparative analysis of hyperparameter tuning techniques—Grid Search, Random Search, Bayesian Optimization, Particle Swarm Optimization (PSO), and Genetic Algorithm (GA)—to improve the accuracy of six machine learning classifiers for Arabic sentiment analysis. The classifiers include Logistic Regression, Ridge Classifier, Support Vector Machine, Decision Tree, Random Forest, and Naive Bayes. Sentiment analysis for Arabic, with its complex morphology, poses challenges. The study evaluates classifier performance on a custom dataset before and after hyperparameter tuning, revealing that Support Vector Machine achieved the highest accuracy of 95.62% using Bayesian Optimization.

D. Hybrid approach

In order to automatically extract opinions from Arabic documents, Alaa M EL-HALEES presented a combined approach [13]. They found that using a single method on Arabic opinion documents produced poor performance, so they used a combined approach consisting of three methods. Initially, the lexicon-based method was used to classify as many documents as possible. The resulting classified documents were then used as a training set for the maximum entropy method, which subsequently classified some more documents. Finally, these documents were used as a training set for the k-nearest method which ranked the rest of the documents. The authors showed that with 1143 items containing 8793 Arabic statements, the system that combines three methods achieved an accuracy of 80.29%. The results further showed that the recall and precision of positive documents were better than negative ones. The authors in [14] proposed an Arabic recommendation system based on opinion analysis and polarity detection. The system operates in three phases. In the first phase, articles are collected and the corpus is manually preprocessed. In the second phase, features are extracted for the representation of comments. Finally, in the third phase, the classification module is implemented and recommendations are generated. The authors combined the random subspace method and the support vector machine (SVM) and showed that the results obtained with the hybrid approach performed well. In [15], the authors proposed a hybrid approach for sentiment analysis of Arabic tweets. This approach combines semantic orientation and machine learning techniques to identify the polarity of Arabic tweets. The lexical-based classifier is used to classify the tweets in an unsupervised manner, i.e. it processes the unlabelled tweets (labeling the training data). The output of the lexical classifier will be used as training data for the SVM machine learning classifier. They showed that none of the individual classifiers, lexical or SVM based, can achieve the results of this hybrid approach, as this approach improved the F-measure of the lexical classifier from 5.76% to 84%, while the accuracy jumped from 16.41% to 84.01%. In their work, Al-Twairesh and Al developed a hybrid method for sentiment analysis of tweets for the Saudi dialect. They used three different classifiers for different levels of classification and integrated knowledge from a lexicon-based method as features into the corpus-based method to develop the hybrid approach. They found that the feature sets extracted from the AraSenTi lexicon worked best for the classification models and that other feature sets, such as emoticons, did not show a clear impact on sentiment classification. Overall, their hybrid approach showed promising results for sentiment analysis of Arabic tweets in the Saudi dialect.TABLE I summarises the related works based on digital and hybrid approaches.j

IV. Sentiment analysis strategy in Arabic Language

In this paper, we propose a new and original approach to . The proposed architecture is divided into four layers. We start by collecting data (which are comments or tweets) from Twitter, then we pre-process these tweets, in order to apply Machine Learning algorithms, at the end we choose the best algorithm that gives the best results

A. Data collection

Social networks are one of the most important sources of data, especially opinions or feelings, as it involves a large number of people expressing their opinions and to express their opinions and feelings.

Data collection can be carried out using several techniques and helps the researcher to understand the phenomenon, fact, or subject he or she is studying. One such technique is web scraping, which allows a wide variety of data, texts and images to be scraped in a relatively short time to automatically feed machine learning models.

1) Web scrapping: Web scraping is a technique for extracting large amounts of data from websites via a program. Most of this data is unstructured data in HTML format which is then converted into structured data in a spreadsheet or database for use in various applications. The collected information is exported in a more useful format for the user (a **JSON/CSV/Excel** file).

TABLE I
RELATED WORKS BASED ON DIGITAL AND HYBRID APPROACHES

Authors	Approach	Language	Algorithm	Dataset
Sghaier, M et Al	Digital	Modern Standard Arabic (MSA)	SVM, NB, RNN	Web Form
DAHOU. A	Digital	Modern Standard Arabic (MSA)	CNN	Web Form
Ibrahim. A et Al	Digital	Arabic dialects	CNN with LSTM	Twitter
Salhi. D et Al	Digital	English, French and Arabic	SVM, LR, RNN	Twitter
Farha, I. A. et al	Digital	Arabic Dialects	CNN with LSTM	SemEval 2017, ASTD, ArS
Salhi D. E. et al	Digital	Arabic, French, English	Logistic Regression, SVM	Twitter (Djezzy operator)
Ghallab, A. et al	Digital	Modern Standard Arabic (MSA)	Various Sentiment Classification Methods	Systematic Review
Alhumoud, S. O. and Al Wazrah, A. A.	Digital	Arabic	RNN	Twitter, Custom Data
EL-HALEES et Al	Hybrid	Modern Standard Arabic (MSA)	Lexicon-based + Maximum Entropy + KNN	Documents
Ziani. A et Al	Hybrid	Modern Standard Arabic (MSA)	Random Subspace Method + SVM	Articles
ALDAYEL, H et Al	Hybrid	Saudian Dialects	Lexicon-based + SVM	Twitter
Al Twairesh. N and Al	Hybrid	Modern Standard Arabic (MSA)	Lexicon-based + RNN	Twitter
EL-HALEES et Al	Hybrid	Modern Standard Arabic (MSA)	Lexicon-based + Maximum Entropy + KNN	Arabic Documents
Aldayel, H. K. et al	Hybrid	Arabic Dialects	Lexicon-based + SVM	Twitter
Al-Twairesh, N. et al	Hybrid	Saudi Dialect	Lexicon-based + SVM	Twitter

2) *Creating the CSV file:* CSV is the acronym for « Coma Separated Values ». The CSV format is a simple text format that is used in many contexts where large amounts of data need to be merged without being directly connected to each other [16].

3) *Example of extracting a CSV file from a website with web scrarping:* To extract data using web scraping, follow these basic steps:

- Choose the data you want to extract: In our case fig. 2 shows the tweet we want to scrape.

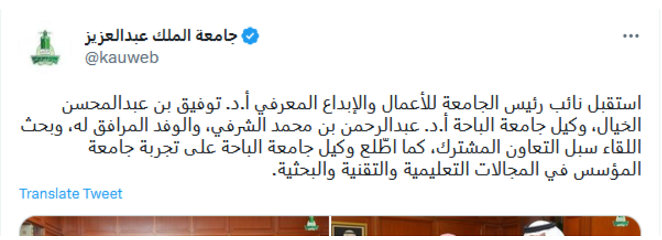

Fig. 2. Tweet in the web

- Find the URL : Fig. 3 shows the URL of the site to be scrapped.

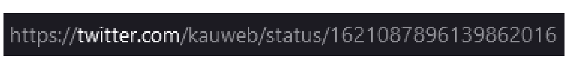

Fig. 3. URL on the web

- Inspect the page and find the data you want to extract: FIG. 4 shows the html code of the page.

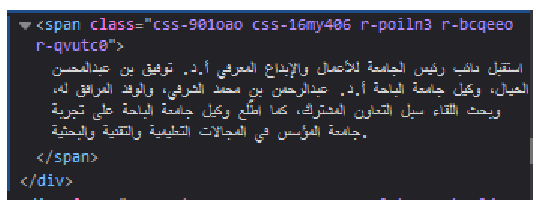

Fig. 4. Html code

- Extract the data and store it in CSV format: Fig. 5 shows the final extracted file.

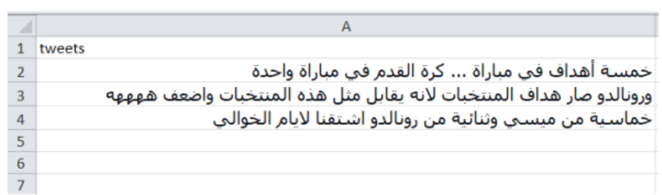

Fig. 5. Final extracted CSV file

B. Pre-processing

This step is considered one of the most important steps prior to the learning process, as it extracts only the important data. It conditions the quality of the models established in data

TABLE II
LABELING

Label	Tweet
Positive	ممتاز نوعا ما. النظافة و الموقع والتجهيز و الشاطيء.
Negative	كتاب سخيف. قصة فاشلة. ممل .و مخيب للظنون
Mixed	جيد بشكل عام ،فيه تكرار، ولم يقنعني بكل الأحوال

mining, and allows the cleaned information within the data to emerge.

1) Cleaning: also known as standardisation process, it consists in converting the document into a standard format that is easy to handle. This step is done through a set of steps, we will use some of them in our system:

- Deleting characters: /,? ,*,!,. . . .
- Deleting numbers: 2, 5863, 8933 , etc.
- Deletion of words and non-Arabic characters.
- Deletion of @ and URL.
- Deletion of repeated letters.

Example: we consider the following comment from Twitter.

انه فندق رااااااائع.....و جمييييييل جدا ١٠٠ !!!!!!

When we apply the netoyage function on the previous comment, we will get :

انه فندق رائع و جميل جدا

2) Labelling: In order to train AI from data, it is imperative to label the data beforehand. We assign labels to the data using various tools. This is what will then allow the computer to learn to recognise the different categories, and to distinguish between them. This task does not require any particular technical skills, but it does require a lot of time. In our case we will label our dataset with three labels: positive, negative or mixed.
Example TABLE II shows three tweets, where each is labelled

3) Tokenization: Tokenisation is a process that divides an input sequence into so-called tokens. We can think of this as a useful unit for semantic processing. It could take the form of a word, a sentence, a paragraph, etc. In our case, it takes the form of a word. We will use tokenisation for the representation of word integration. **Example** we have the following cleaned up tweet :

انه فندق رائع و جميل جدا

After applying the tokenisation function, we will get the following result :

انه ، فندق ، رائع ، و ، جميل ، جدا

4) Elimination of stop words: Stop-words are common words that have no effect on classification. This step consists of removing all stop-words by comparing each known word with the elements of the stop-word list. Fig. 6 shows some stop words in Arabic language.

Relative Nouns	الأسماء الموصولة	الذي، التي، الذين
Pronouns	الضمائر	نحن، أنتم، هما
Conjunctive tools	أدوات الربط	و، ثم، ف
Adverbs of time and place	ظروف الزمان والمكان	صباح، مساء، يوم
Weekdays and months	أيام الأسبوع والأشهر	الأربعاء، الجمعة، فيفري، محرم
Proper nouns	أسماء علم	محمد، لين، الجزائر، القاهرة، قطر
Common words	كلمات كثيرة الاستعمال	جدا، قليلا، أيضا، أكثر

Fig. 6. Some stop words in Arabic language

Example : we consider the following tweet already cleaned and tokenized.

انا، لم ، يعجبني، هذا، الفندق، الخدمات، التي، يقدمها، سيئة ،
جدا، لا، انصحكم ، به

After applying the stop words elimination function we obtain this result :

لم ، يعجبني، الفندق، الخدمات، يقدمها، سيئة، لا، انصحكم

5) Stemming: Stemming generally refers to the crude heuristic process of reducing inflected or sometimes derived words and cutting off their endings in order to retain only the root of the word. The root need not be identical to the morphological root of the word; it usually suffits that related words correspond to the same root. This phase is very complicated since a word in Arabic can be formed from a base (the root) to which affixes and/or clitics may be added and it is very difficult to separate them. Figure FIG. 7 reports the list of clitics while distinguishing the two classes proclitics and enclitics [61].
In our case, we will eliminate enclitics like :

نا ،ي ،ك ،كن ،كم

and prefixes such as :

ان، ين ،ون

as well as many other proclitics and prefixes in order not to change the meaning of the word.
Example
We consider the following comment from the already cleaned, tokenized Twitter which does not contain any empty words:

لم يعجبني الفندق الخدمات يقدمها سيئة لا انصحكم

After applying the stemming function we get the following result :

لم يعجب فندق خدم يقدم سيئة لا انصح

Clitics		Translation	Grammatical category
Proclitic	ال	/Al/	Defined item
	و	/w/	Coordinating conjunction
	ف	If/	Conjunction of subordination
	ل	/I/	Preposition
	ب	/b/	Preposition
	ك	/k/	Preposition
	س	/s/	Future marker particle
	أ	/a'/	Question mark
Enclictic	ي	/y/	1st person singular relative pronoun
	ني	/ny/	1st person singular relative pronoun
	ك	/k/	2nd person singular relative (or possessive) pronoun
	كما	/kmA/	duel relative (or possessive) pronoun
	هم	/hm/	3rd person plural masculine relative (or possessive) pronoun
	هن	**/hn/**	3rd person plural feminine relative (or possessive) pronoun
	كم	/ m /	2nd person plural masculine relative (or possessive) pronoun
	كن	/kn/	2nd person plural feminine relative (or possessive) pronoun
	نا	/nA/	relative (or possessive) pronoun in the 1st person plural

Fig. 7. Some stop words in Arabic language

C. Creating vectors

Considering that the texts are not formal. Therefore, before the learning process, they have to be learning process, they must be converted into numerical vectors which are characterised by their formality. Then the next step is the word representation process, which will will lead to a set of ready-to-learn vectors on which we will apply three different models.

1) Word2vec: Word2vec is a natural language processing technique published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text.

2) TF-IDF: TF-IDF (term frequency-inverse document frequency) is a weighting method often used in information retrieval and in particular in text mining.

3) BERT: BERT stands for "Bidirectional encoder representations from transformers". It is a so-called pre-training language model based on neural networks.

D. Modelling

Once the data is usable, it is finally time to move on to the modelling stage. In the case where sentiment analysis consists of determining the polarity of a text, the problem can be considered as a problem of classification by polarity. This phase consists of two steps: learning and prediction. The learning step consists in building a model from a training corpus (training data). While the prediction step uses the trained learning model to assign a class (a polarity) to a new unlabelled twit.

1) Dataset division: Before starting the learning process, we must first have two databases available. it is therefore necessary to divide our dataset into 2 parts, the first for training and the second for testing. The training data should be labeled to increase the accuracy of algorithms. This part contains 80% of the initial data.the test data is the remaining 20% of the initial data.

2) Proposed models: this task relies on machine and/or deep learning techniques, according to related and existing work, we propose to apply several algorithms like SVM, Logistic regression, Naive Bayes, Random Forest, Decision Tree and Recursive Neural Network.

E. Display of results

once we end application of chosen algorithms, we compare and display the results of accuracy between them. Fig. 8 illustrates our proposed approach in one schema detailed.

V. Results and discussion

After presenting our proposed approach and mentioning its most important steps. In this section we will show how to apply them and we will present the most important results obtained.

A. Information Collection

The recovery of the comments is done on the social network Twitter by a web scrapping function developed by the authors of this paper, where we analyzed different pages, and at the end we recovered more than thirteen thousand tweets between a negative and positive opinion. After that, we tried to balance the dataset, **Fig.9** shows the exact number of tweets.

B. Pre-processing

To train our model, we first processed our data set by following these steps:

1) Cleaning: We cleaned up our dataset using our "cleaningText" function, which removes non-Arabic characters and words, repeated letters and extra spaces.

2) Tokenization: We tokenized the dataset using the "tokenizingText" function, which we implemented using the "nltk" library.

3) Elimination of Stop Words: This was a crucial step in our process, as we aimed to eliminate as many stop words as possible. To achieve this, we developed an algorithm that allowed us to efficiently remove a significant number of stop words. Subsequently, we saved all the removed stop words in a CSV file, enabling others to utilize this resource in their own work. The added value of this paper is to create a dataset for stopwords, where we extracted more than 3000 stopwords from Holy Quoran. The dataset is published in GitHub.

4) Stemming: we have eliminated suffixes and some prefixes using our "stemmingText" function, so as not to lose the meaning of the words. **Fig. 10** shows the results of the tweets cleaned after the execution of the different functions mentioned before. Holy Quoran has more than 77000 words, it's very important source of vocabulary, for that we used this words to extract roots of words, where we have find more than 7000 roots of words. This list is published in GitHub.

C. Classification

Before classifying the feelings in positive or negative class, we made a vectorization or we applied two methods of extraction of the attributes which are: the TF-IDF and WORD2VEC. then we made a division (train-test) on the database into 20% of the test data, and 80% of the data for the training.

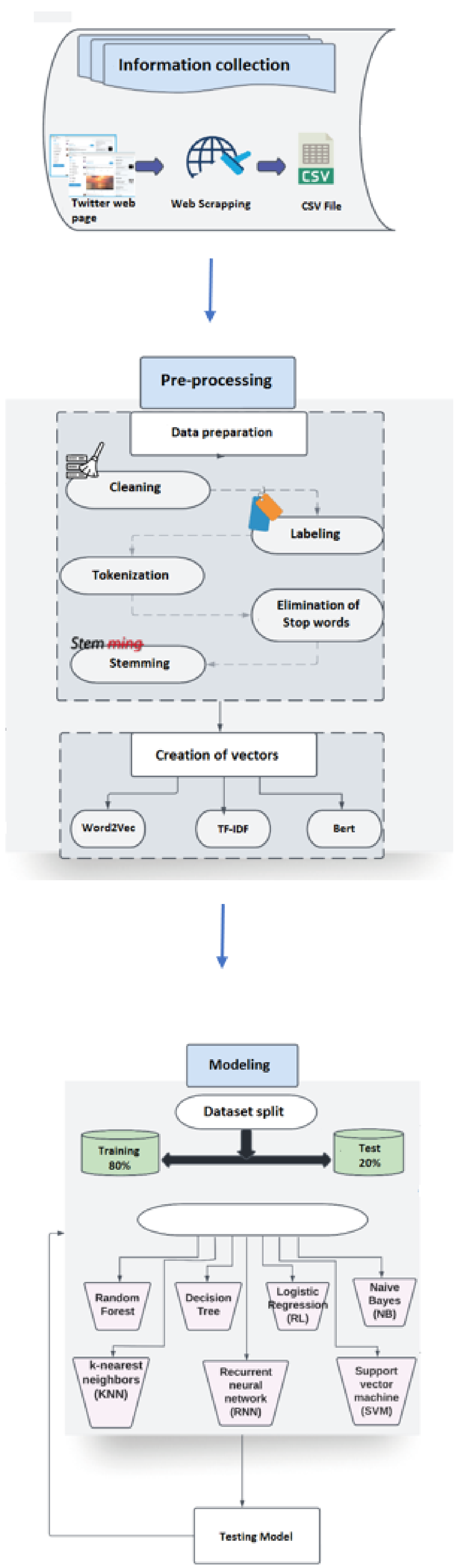

Fig. 8. Proposed Approach

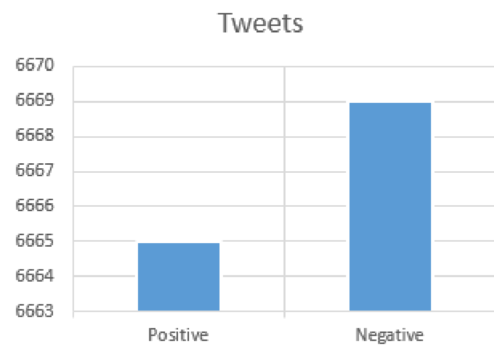

Fig. 9. The number of tweets divided between positive and negative opinions

	label	text	text_clean
0	Positive	الكتاب أكثر من رائع. الكاتب له وجهة نظر إبداعي...	أكثر, رائع, كاتب, وجه, نظر, إبداعي, نرا, عادي]...
1	Negative	ممل جدا و كثير التكرار ... هو اسلوب فاروق جويد...	ممل, تكرار, اسلوب, فاروق, جويد, كد, لول, اضطر]...
2	Negative	يبدو ميسو مغرما بهذا النمط :شخصيات رئيسية للرو...	يبدو, ميسو, نمط, شخصي, رئيسي, للرواي, يجمع, ش]...
3	Negative	رواية بقلم تونسي، تروي تاريخ تونس الحديث و ذلك...	رواي, بقلم, ترو, تاريخ, حديث, حدث, تصلح, قريب]...
4	Positive	جميل ،، يعيبه البعد عن الحرم . النظافةالموظفين...	جميل, يعيب, حرم, اسواق, ومطاعم, حرم, دقيق, سح]...
...	...	...	...
66661	Positive	استثنائي. موقع الفندق نظافة المكان طاقم العمل ...	[استثنائ, موقع, فندق, نظاف, طاقم, عمل, لا, يوجد]
66662	Positive	جيد. النظافه والموقع. تأخر خدمة الغرف	[جيد, نظاف, موقع, تأخر, خدم, غرف]
66663	Negative	مخيب للأمل. حجم الغرف مناسب. استعمال اكثر من م...	مخيب, للأمل, حجم, غرف, مصعد, للوصول, للغرفةال]...
66664	Positive	كل الفندق جميل جذاب مريح الاستقبال العماله	[فندق, جميل, مريح, استقبال, لاشئ]
66665	Negative	مخيب للأمل. الموقع. السعر مبالغ فيه جداالسرير ...	مخيب, للأمل, موقع, سعر, مبالغ, مزدوج, غير, مر]...

66666 rows × 3 columns

Fig. 10. DataSet of tweets cleaned

1) Modeling with TFIDF: We applied with this method six different classifiers (SVM, Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), KNN and RNN), we detail the results obtained in **Fig.11**. The dimension of these confusion matrices is 2*2, we have two classes negative and positive. Row values represent accurate predictions, while column items are incorrect predictions.

Table 3 contains the details of the experimental results conducted in our project using the TFIDF and the 6 classifiers mentioned before to classify feelings into positive and negative classes.

TABLE III
RESULTS OF ALGORITHMS ACCURACIES USING TF-IDF

Metrics	SVM	RNN	LR	NB	RF	KNN
Accuracy	0.83	0.83	0.82	0.82	0.66	0.76
Positive Precision	0.83	0.83	0.82	0.83	0.67	0.77
Positive Recall	0.82	0.84	0.81	0.81	0.62	0.75
Positive F1 Score	0.82	0.83	0.81	0.82	0.65	0.76
Negative Precision	0.82	0.84	0.81	0.81	0.65	0.76
Negative Recall	0.83	0.83	0.83	0.84	0.70	0.77
Negative F1 Score	0.83	0.83	0.82	0.82	0.67	0.76

For the results we obtained with the TF-IDF, we note that the precision is high in the negative class with the RL classifier compared to the positive class and the other classifiers with a value equal to 0.84.
On the other hand we note that the recall is high in the

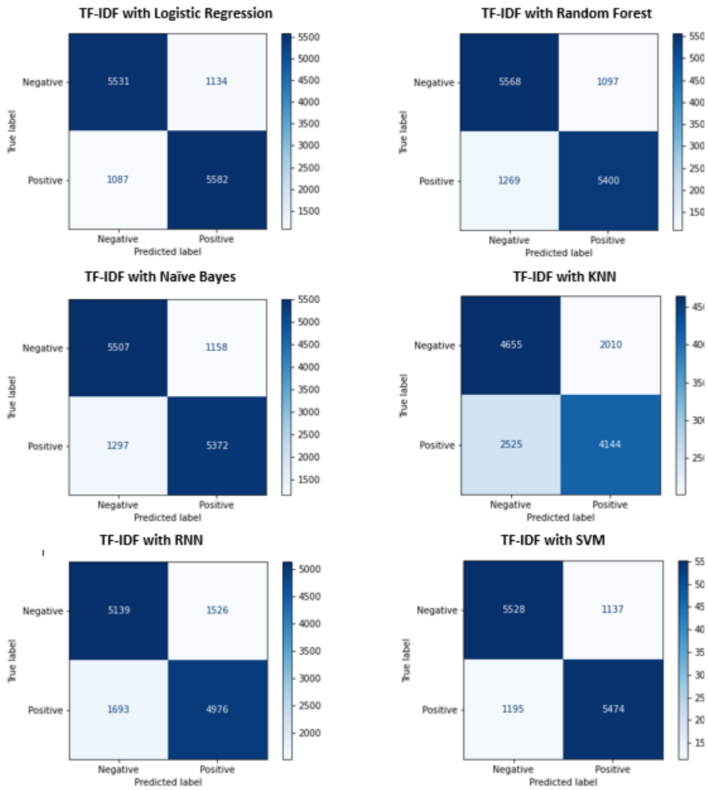

Fig. 11. Application of Machine Learning algorithms with TF-IDF

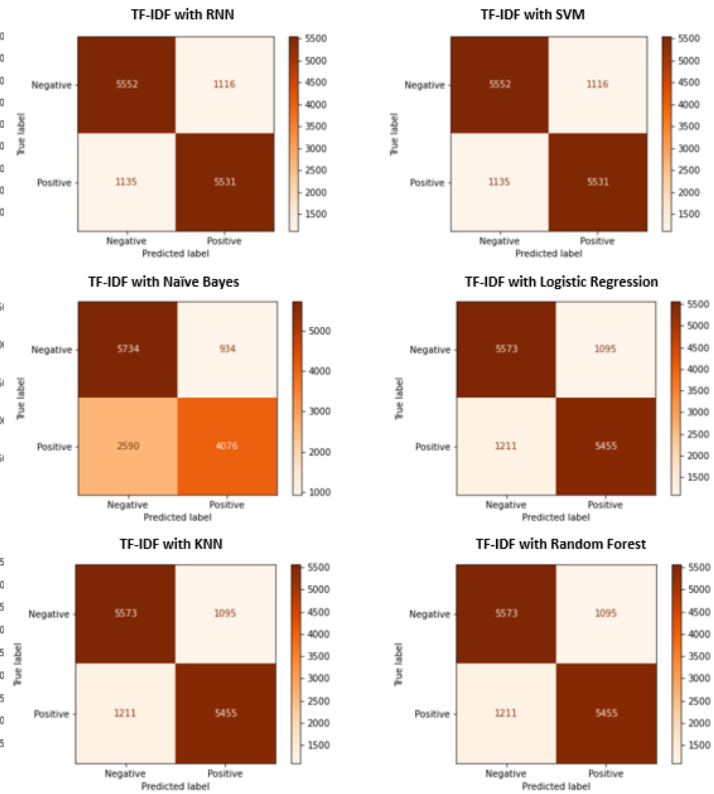

Fig. 12. Application of Machine Learning algorithms with WORD2VEC

positive and negative class with the same classifier which is RL compared to the other classifiers with a value equal to 0.84 for both classes.

The F1-score is high in the negative and positive class with the RL classifier and the negative class with the SVM classifier compared to the other classifiers with a value equal to 0.83

2) Modeling with WORD2VEC: We now apply WORD2VEC in the same way of TF-IDF with the six algorithms mentioned before, **Fig.12** shows the confusion matrices of the different algorithms applied. **Table 4** mention different results of performance measures like accuracy, Precision, Recall and F1 Score of algorithms applied with Word2Vec.

TABLE IV
RESULTS OF ALGORITHMS ACCURACIES USING WORD2VEC

Metrics	SVM	RNN	LR	NB	RF	KNN
Accuracy	0.85	0.91	0.83	0.74	0.71	0.80
Positive Precision	0.84	0.90	0.83	0.81	0.71	0.81
Positive Recall	0.86	0.92	0.83	0.61	0.71	0.78
Positive F1 Score	0.85	0.91	0.83	0.70	0.71	0.79
Negative Precision	0.86	0.92	0.83	0.69	0.71	0.79
Negative Recall	0.84	0.90	0.83	0.86	0.72	0.82
Negative F1 Score	0.85	0.91	0.83	0.76	0.71	0.80

For the results we obtained with WORD2VEC, we note that the accuracy is high in the negative class with the RNN classifier compared to the positive class and other classifiers, with a value of 0.92.

On the other hand, we note that the recall is high in the positive class with the same classifier which is RNN, and in the negative class with the NB classifier compared to the other classifiers, with a value of 0.91.

The F1-score is also high in the negative and positive class with the same classifier which is RNN compared to the other classifiers, with a value of 0.90.

3) Comparison: According to the results we obtained, we observed that the best classification was with word2vec using the RNN classifier.

VI. CONCLUSION

In conclusion, this paper provides a comparative analysis of digital and hybrid approaches to Arabic sentiment analysis, focusing on various machine learning algorithms and hyperparameter tuning techniques. The results demonstrate that Support Vector Machines (SVM) and Recurrent Neural Networks (RNN) consistently achieve the highest accuracy across different datasets. Specifically, SVM achieved 95.62% accuracy using Bayesian Optimization, while RNN models, particularly with Word2Vec, achieved up to 91% accuracy. The hybrid approaches combining lexicon-based methods with machine learning algorithms like SVM and RNN also performed well, improving classification accuracy by capturing both explicit and implicit sentiment features.

The results suggest that while deep learning models such as CNN and RNN are highly effective in handling unstructured data, lexicon-based methods still offer valuable performance boosts, especially when combined in hybrid models. Overall, the choice of algorithm and tuning method greatly influences the model's performance, highlighting the need for careful selection and optimization in sentiment analysis tasks, particularly when dealing with the complexities of the Arabic language and its various dialects

REFERENCES

[1] C. Dhaoui, C. M. Webster, and L. P. Tan, "Social media sentiment analysis: lexicon versus machine learning," *Journal of Consumer Marketing*, vol. 34, no. 6, pp. 480–488, 2017.

[2] T. Baccouche, "Esquisse d'une étude comparative des schémas des verbes en arabe classique et en arabe tunisien," *Les cahiers de Tunisie*, vol. 22, pp. 87–88, 1974.

[3] A. Barhoumi, "Une approche neuronale pour l'analyse d'opinions en arabe," Ph.D. dissertation, Le Mans, 2020.

[4] A. Assiri, A. Emam, and H. Aldossari, "Arabic sentiment analysis: a survey," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, 2015.

[5] M. A. Sghaier, H. Abdellaoui, R. Ayadi, and M. Zrigui, "Analyse de sentiments et extraction des opinions pour les sites e-commerce: application sur la langue arabe," in *5th International Conference on Arabic Language Processing (CITALA)*, 2014.

[6] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for arabic sentiment classification," in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2418–2427.

[7] I. A. Farha and W. Magdy, "Mazajak: An online arabic sentiment analyser," in *Proceedings of the fourth arabic natural language processing workshop*, 2019, pp. 192–198.

[8] D. E. Salhi, A. Tari, and M. T. Kechadi, "Using e-reputation for sentiment analysis: Twitter as a case study," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 11, no. 2, pp. 32–47, 2021.

[9] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Applied Computational Intelligence and Soft Computing*, vol. 2020, no. 1, p. 7403128, 2020.

[10] S. O. Alhumoud and A. A. Al Wazrah, "Arabic sentiment analysis using recurrent neural networks: a review," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 707–748, 2022.

[11] A. Alsayat and N. Elmitwally, "A comprehensive study for arabic sentiment analysis (challenges and applications)," *Egyptian Informatics Journal*, vol. 21, no. 1, pp. 7–12, 2020.

[12] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," in *Informatics*, vol. 8, no. 4. MDPI, 2021, p. 79.

[13] Z. Amel, "La recommandation via l'analyse d'opinions," Ph.D. dissertation, Université Badji Mokhtar–Annaba, 2018.

[14] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis–a hybrid scheme," *Journal of Information Science*, vol. 42, no. 6, pp. 782–797, 2016.

[15] N. Al-Twairesh, H. Al-Khalifa, A. Alsalman, and Y. Al-Ohali, "Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach," *arXiv preprint arXiv:1805.08533*, 2018.

[16] L. Souag, "Clitic doubling and contact in arabic," *Zeitschrift für arabische Linguistik= Journal of Arabic linguistics= Journal de linguistique arabe*, vol. 66, pp. 45–70, 2017.