Investigate the Impact of Stemming on Mauritanian Dialect Classification using Machine Learning Techniques

Mohamed El Moustapha El Arby CHRIF[1], Cheikhane Seyed[2], Cheikhne Mohamed Mahmoud[3]
EL BENANY Mohamed Mahmoud[4], Fatimetou Mint Mohamed-Saleck[5], Moustapha Mohamed Saleck[6]
Omar EL BEQQALI[7], Mohamedade Farouk NANNE[8]
Department of Mathematics and Computer Science, Nouakchott University, Nouakchott, Mauritania[1,3,4,5,6,8]
Higher School of Polytechnic of Dakar, University Cheikh Anta Diop, Dakar, Senegal[2]
Department of Computer Science, University Sidi Mohamed Ben Abdallah, Fes, Maroc[7]

***Abstract*—Despite the plethora and diversity of research on Natural Language Processing (NLP). As a technique allowing computers to understand, generate, and manipulate human language; It still remains insufficient, especially with regard to the processing of Arabic texts and their dialects which are widely used. The proposed approach focuses on the application of machine learning techniques taking into account evaluation criteria such as training to comments expressed in Mauritanian dialect, published on social media notably Facebook, and compares results generated by three algorithms which we applied such as the Random Forest (RF), Naïve Bayes Multinominal (NBM), and Logistic Regression (LR) algorithm. Additionally, We then study the effect of machine learning techniques when different stemmers are combined with other features such as the tokenizers used to process the dataset. Although major challenges exist such as the morphology of Arabic is completely different from Latin letter languages, and there is no pre-existing dataset or dictionary to train the algorithms, the result we obtained after the experiments carried out on Weka shows that the RF and NBM algorithms are more efficient when applied with ArbicStemmerKhoja giving results respectively 96.37% and 71.40%; However, Logistic gets better performance results with Null Stemme is 81.65%. Results obtained by the three techniques applied with a light Arabic stemmer were more than 70%. This article presents a contribution to NLP based on Machine learning, descript also an important study that can determine the best Arabic classifier.**

Keywords*—Machine learning; Natural Language Processing; Arabic text classification; HASSANIYA dialect; Weka; stemming

I. Introduction

Mauritania, like other countries around the world, has been invaded by new technology, which has given rise to exchange platforms commonly known as social media, through which inter-family exchanges on the one hand, and inter-governmental and two-way exchanges between government agencies and the public (citizens) on the other, can take place. Thereafter a data stream in dialect Mauritania and Arabic language will be generated reflecting the citizens' sentiment. Whereas, Sentiment analysis is an approach that uses natural language processing (NLP) [1], machine learning analysis methods [2] [3], or other lexicon-based methods [4] to extract, convert, and interpret opinions from a text and classify them into positive, negative, or neutral sentiments. However, the emergence of new technologies will allow governments and companies to take into account the opinions of their public via social media, which would help them make better decisions. Thus, artificial intelligence (AI) within other technologies has solved the challenges of business practice and introduced the application of Business Intelligence (BI) that has promoted the transformation of information. In this sense, several processing techniques have been used to classify large volumes of data (Big Data) for example regression analysis, Naïve Bayes (NB), Support Vector Machine (SVM), and Neural Network (NN) [5]. So far, most research work has been done to classify text using Machine Learning for various languages like English more than Arabic sometimes when Arabic native speakers are more than non-Arabic according to [4].

We consider dialectal Arabic to be a new field of research in the field of text classification, for several reasons: firstly, dialect is widely used in social networks, which generates a large amount of data; secondly, dialect, whether HASSANIYA or others, is generally more widely used than the main language, even if this is not an official case. Moreover, the Mauritanian dialect has an alphabet and script that are those of Arabic, which means that dialect has become an important area of research; Regarding the complexity of HAASANIYA, justified by the reasons listed above, in this work, we propose an approach that gives a clear view of the classification of HASSANIYA text using machine learning algorithms and comparing the archived result. To implement the proposed approach, we use WAKE, which implements several filters and classifiers from machine learning algorithms [6].

The differences between Standard Arabic and dialectal Arabic are minimal in terms of derivation and grammar, as well as termination. On this basis, we decided to study the classification of Mauritanian dialectal texts taking into account the effect of the stemmer method. The goal of this work is to identify the best Machine Learning for dialect classification and the effect of stemming and tokenization on text classification, particularly the HASSANIYA dialect; nevertheless, we fusion deference filters for building our property models.

More specifically, a HASSANIYA dataset was collected on Facebook and contains comments posted on popular pages in Mauritania such as bloggers' pages or government pages (the Ministry of Hydraulics, the Ministry of the Interior and Decentralization, Ministry of Housing and Urban Development) that we prepared in order to prove the stemmer method on dialectal text. To experiment, three types of stemmer were adopted in this work such as light stemming, null stemming, and heavy stemming in this case we use khoja, and every one of these three types is fusionned with another filter to make a new method. We tested three Machine Learning algorithms individually on the models built mentioned above. Machine Learning techniques applied are NVM, RF, and logistic regression.

The main contributions of this paper are:

- scraping data from Facebook pages to build a HASSANIYA dialect dataset;
- proving the classification of the HASSANIYA dialect text with Machine Learning techniques;
- applying and comparing different types of Stemmer to improve text classification.

The paper is organized as follows: an introduction followed by a state of the art and literature review gives an overview of the Arabic language and its dialects, then we explain our research methodologies followed in this paper and finally, we discuss the results obtained and give a conclusion.

II. Related Works

Text classification presents an amazing field in the data analysis area, and still rich in terms of scientific research, increasing domain due to what we let know. Many researchers studied these cases and realized more articles, but the Arabic language and its dialects still need more work.

In this context, several studies have been carried out such as the approach [4] gives an approach for the classification of Arabic texts using various algorithms, and showed an enhancement in the accuracy of classifier models.

Authors in the article [7] explored a comparative system on two different datasets based on the machine learning technique, classification models are compared in terms of accuracy for each dataset.

Another study [8] applied six variations of the Bayes classifier on Arabic data, after analyzing their results were compared, and showed that the best values were generated successively from Naïve Bayes and Naïve Baye Update, in another way Naïve Bayes Multinomial Text generated the worst results.

Proposed [9] a contribution to big data processing which is considered a challenging stage of data analysis ax, so a solution proved for the challenges in four stages: data collection, cleaning, enrichment, and availability. they looked to convert social media data to computation-based data after it was source-based.

The author [10] Proposed a model for text dialectal classification, they prepared a dataset of Marocain dialect scraped from Twitter comments and a combination of extraction(n-grames), weighting schema(Bow, TF-IDF), and word embedding was applied in order to prove the Marocain dialect classification and get the best classify model. the Machine Learning techniques which they applied are following: Naive Bayes, Random Fest, support Vector Machine, Logistic Regression, and a Deep Learning Model such as Long Short term Memory (LSTM). the experimental work showed that the SVM achieved an accuracy equal to 70%.

This paper [11] proposed a new algorithm to generate all potential derivation roots of an Arabic word, without deleting initial affixes. the author seeks to address the weaknesses and errors of existing algorithms in order to improve the accuracy of Arabic Natural language Processing. they used in this study a data set that includes a collection of roots, patterns, and affixes. by matching the derived word to identify the root. and then, they get an average accuracy rate of 96%.

This study [12] proposed a model as a novel assembly of CNNs for analysis of the task of Arabic dialect classification from spontaneous Arabic speech dataset. this model is based on a fusion of linguistic and acoustic features and uses pre-trained bidirectional encoder representation from the transformed (BERT) Model. the proposed approach achieves an accuracy of 82.44% for the identification task of five Arabic dialects.

The author [13] proposes an approach to improve P-Stemmer by combining it with various classifiers such as Naïve Bayes, Random Forest, Support Vector Machines, K-Nearest Neighbor, and K-Star. In this study they used a data set synthesized from various online news pages and did the experience on Weka tools, which is achieving the result showed that the P stemmer has Improved when using NB.

III. Review of Arabic and ITS Dialects

Arabic is one of the major languages used in the world, it is used by all Muslims because is the language of the holy book of Islam [6], [14], as well, Arabic divided into three categories according to [14] as follows: First, Classical Arabic (CA) is considered the oldest type it is the Arabic literature, The Holy Quran; Second, Modern Standard Arabic (MSA) can be defined as a simplified version of Classical Arabic to be comprehensible by whole people and be became largely used; and then, exists a third type called Arabic Dialect use the same Arabic characters for writing, this one used more than two above types in daily life.

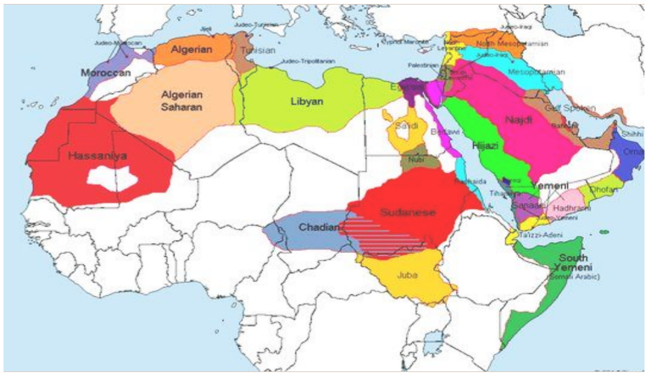


Fig. 1. Arabic dialect Map.

The Arabic dialect is divided into several types, as shown in Fig. 1; for instance:

- Moroccan Dialect: The Moroccan dialect commonly Diarj [10];

- Egyptian Dialect: The Egyptian dialect is spoken in Egypt;
- North African Dialect: The North African dialect is spoken in Algeria, Libya, and Morocco [12];
- Tunisian Dialect: Tunisian [15], and others are also Arabic dialects [16];
- Mauritania Dialect: This is named HISSANIYA dialect and is spoken mainly in the middle Mougreb region, more specifically in Mauritania country.

Mauritania Dialect: The Mauritanian dialect named HASSANIYA is a local dialect and a variety of Maghrebi Arabic spoken by Mauritanian Arabs widely used in daily life not only to change between families but also to indicate or share feelings and opinions on social media and to interact with others' posts. The operation of HASSANIYA text classification is becoming increasingly complicated, for three reasons: firstly, HASSINIYA is an Arabic dialect that has the same letters of the alphabet for writing, with changes in pronunciation and meaning depending on their diacritics, and ambiguity between words' root and their derivation; secondly, it is an unstructured language; thirdly, there is no data set or dictionary pre-exists.

In Table I, we segment an example of a Hassaniya word into sub-segments that show its basic construction; as mentioned above, this dialect uses Arabic letters and can be conjugated with all subjects and tenses; as shown in the following table, the word HISSANIYA has an affix such as prefix, suffix, and postfix determined by usage.

TABLE I. Example of a HASSANIYA „ ماخَلَّيْتُو „ Word which has Different Affixes Attached to a Root Word

Word	تُو	يْ	خَلَّ	ما
Meaning	Pronoun ‘‘ You ”	Termination of conjugation	let	Negation like ”Don’t”
Affixes	Postfix	Suffix	Root	Prefix

IV. Materials and Methods

The main stages in our proposed Methodology are data collection, preprocessing, building technique, classification, and Evaluation stages will be described in the following. This approach was applied using Weka tools. For the sake of a better selection of dialectal words, we adopt in this work a methodology consisting of phases shown in Fig. 2.

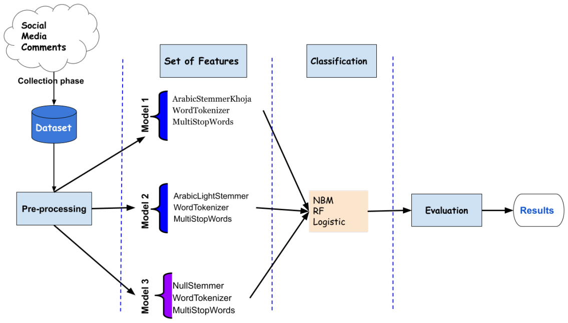


Fig. 2. Proposed architecture

A. DataSet Description

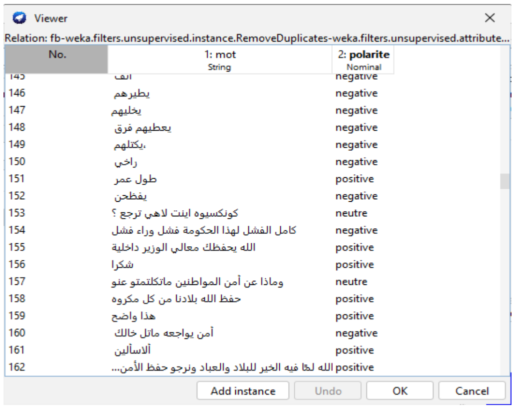
Social media data is the main source of our Data set. We have built our own Data set which contains words and sentences in HASSANIYA by gathering hundreds of comments from Facebook pages using scrap tools that present cytosine's reaction to government activities and then annotating them according to their polarity. We annotated each comment extracted according to his opinion hidden behind the writing.

The corpus of the dataset is present in the below Table 2.

TABLE II. DataSet

Class	Comment
Positive	321
Negative	348
Neutre	337
Total	1006

Based on our knowledge of the local language, we divided the corpus into three categories looking at opinions reflected as positive, negative, and neutral as well as shown in Fig. 3. Moreover, we loaded comments on Interim storage as a CSV file after converting it to ARFF format for use on Weka.



Relation: fb-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute...

No.	1: mot String	2: polarite Nominal
145	الف	negative
146	يطيرهم	negative
147	يخليهم	negative
148	يعطيهم فرق	negative
149	يكتلهم،	negative
150	راخي	negative
151	طول عمر	positive
152	يفظحن	negative
153	كونكسيوه اينت لاهي ترجع ؟	neutre
154	كامل الفشل لهذا الحكومة فشل وراء فشل	negative
155	الله يحفظك معالي الوزير داخلية	positive
156	شكرا	positive
157	وماذا عن أمن المواطنين ماتكلتمتو عنو	neutre
158	حفظ الله بلادنا من كل مكروه	positive
159	هذا واضح	positive
160	أمن يواجعه ماتل خالك	negative
161	ألاسألين	positive
162	الله لنا فيه الخير للبلاد والعباد ونرجو حفظ الأمن...	positive

Fig. 3. DataSet example annotated.

The data was balanced by the Smote method as well as shown in the following Fig. 4.

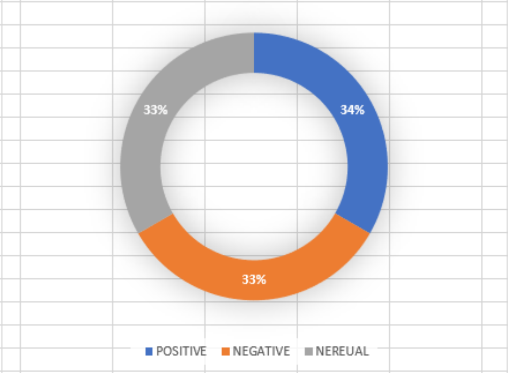


Fig. 4. DataSet balanced.

B. Pre-processing

Pre-processing is the first step in the data analysis process and that is a crucial step when dealing with Arabic documents [17]. In order to convert input data to a performed text clearly and useful for machine analyses, we were using in this work a process consisting of various steps presented in the following. thus, these steps and filtering are offered by the Weka tool.

Tokenization : Tokenization is a technique that divides and transforms the word into tokens while preserving the meaning of the words by removing spaces, punctuation, and non-Arabic words [18]; in this case, the document is also reduced to words.

Normalisation : Word normalization means giving a format where some letters appear differently [19] for instance, ا, can appear in different forms like أ, آ and إ .

StopWordsHandler : stop word is used to eliminate everything not part of the word's root.

Stemming : The streaming method is an essential step in Natural Language Processing or text classification, which converts the word into its corresponding root or stem. stem is the combination of a root and its derivation which is a suffix prefix and postfix [16]. There are two main types of stemming in Arabic, namely Stem or Light Stemming and root-based stemming, the first one can be explained by removing the suffix and prefixes from the word in order to obtain its root; The second type is divided into three sub-categories. according to [17] such as (i) Dictionary Based when using a file dictionary; like khoja. (ii) no-dictionary bases, and hybrid that is shown in Fig. 5. There are several Stemming approaches applied to the Arabic language the following is a non-exhaustive type.

Light stemmer : Light stemmer is one category of stemming approach that aims to reduce words to their stem by means of removing the most frequent word's prefix and/or suffix [20], [21], [19] .

Heavy Stemmer : Heavy stemming is the process of eliminating affixes and changing certain letters in words to obtain the root word [22].

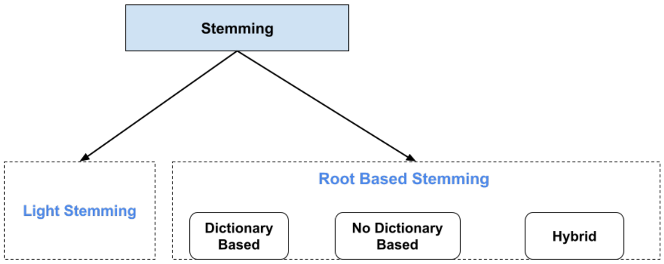


Fig. 5. Stemming structure approach.

SetMinTermFreq() : The SetMinTermFreq() method is used to define the minimum frequency threshold of a term (word) to be taken into account in the feature vector; In Weka, the StringToWordVector filter allows us to convert a collection of text documents into a set of numerical features, where each feature represents the frequency of a specific word in the document. In this study, you have used a minimum frequency equal to 2.

C. WEKA tools

WEKA is a machine learning framework with a graphical user interface, making it easy to use for beginners. it also includes a large collection of machine learning models such as Neural Networks, Decision Trees, and K-means. provides implementations of learning algorithms that can be applied for data analysis purposes [23].

Weka covers tools for transforming datasets, such as discretization and sampling algorithms for pre-processing a dataset, integrating it into a learning scheme, and analyzing the resulting classifier and its performance.

D. Machine Learning Algorithms

1) Naive Bayes Multinominal : Naïve Bayes (NB) is a data mining algorithm dedicated to data classification [24]. It is used to deduce the probability of a datum belonging to a class, based on the assumption that all attributes are independent of each other given the class [25]. In this work, we use Multinomial Naive Bayes to assign texts to classes based on statistical analysis of their content. This algorithm offers an alternative to the often cumbersome semantic analysis based on artificial intelligence and considerably simplifies the classification of textual data. It aims to classify by assigning text fragments to classes while determining the probability of a document belonging to the class in other documents with the same subject.

2) Random Forest : According to [26] RF is a set of decision trees where each tree is built from a bootstrap version of the training data set. Each tree is built according to the principle of repetitive partitioning: starting from the root node, the same node-splitting procedure is applied repeatedly until certain stopping rules are met. Its predictive power comes from the aggregation of many weaker learners (decision trees). Performance is particularly good if correlations between forest trees are low.

3) Logistic : Logistic regression is an important technique in the field of artificial intelligence and machine learning for data analysis that uses mathematics to find relationships between two data factors. It then uses this relationship to predict the value of one of these factors as a function of the other. The prediction usually has a finite number of outcomes, such as yes or no. Logistic regression belongs to the family of supervised machine learning models. It is also considered a discriminative model, meaning that it attempts to distinguish between classes (or categories) [27].

E. Evaluation Metrics

Text classification models are evaluated using well-defined essential criteria. This set was used to evaluate our models [28]. To evaluate the accuracy of our Models', a confusion metric is defined by [10] as a tool to evaluate the accuracy of ML models' predictions and compare their predictions to reality. Since We have three classes to be classified, six important terms will have come into the evaluation process as shown in Fig. 6. Results obtained are assessed using the F1 score, precision, accuracy, and recall.

Tp: here is true Positive, where the prediction is positive, and the actual values are positive also.

Fp: here is a False positive, where the prediction is positive, but the actual values are Negative or Neutral.

Tng: here is true Negative, where the prediction is negative, and the actual values are negative.

Fng: here is a false negative, where the prediction is negative, but the actual values are positive or neutral.

Tn: here is true Neutral, where the prediction is neutral, and the actual values are Neutral also.

Fn: here is a false Neutral, where the prediction is neutral, but the actual values are positive or negative also.

Precision : Precision (P) measures how many of the “positive” predictions are made correctly by the model. The mathematical formula is as follows :

$$\mathcal{P} = \frac{\mathcal{T}_{\mathcal{P}}}{\mathcal{T}_{\mathcal{P}} + \mathcal{F}_{\mathcal{P}} + \mathcal{F}_{\mathcal{P}}} \tag{1}$$

Recall : Recall(R) measures how many of the positive class samples present in the dataset were correctly identified by the model. calculated by the following mathematical formula:

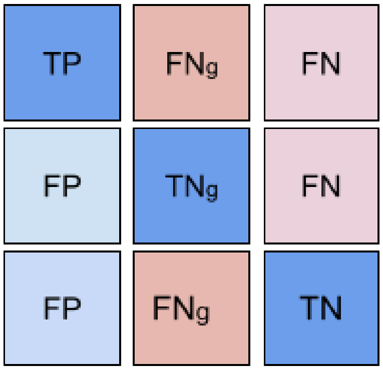
$$\mathcal{R} = \frac{\mathcal{TP}}{\mathcal{TP} + \mathcal{F}_{\mathcal{N}e} + \mathcal{F}_{\mathcal{N}}} \tag{2}$$

F-Measure : F-Measure or F1 score is a machine-learning evaluation metric that measures a model’s accuracy. It combines the precision and recall scores of a model. given by the formula:

$$\mathcal{F}1 = 2 * \frac{\mathcal{P} * \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{3}$$

Accuracy : The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$Accuracy = \frac{\mathcal{T}_{\mathcal{P}} + \mathcal{T}_{\mathcal{N}} + \mathcal{T}_{\mathcal{N}g}}{\mathcal{F}_{\mathcal{N}} + \mathcal{T}_{\mathcal{N}} + \mathcal{T}_{\mathcal{P}} + \mathcal{F}_{\mathcal{P}} + \mathcal{T}_{\mathcal{N}g} + \mathcal{F}_{\mathcal{N}g}} \tag{4}$$



TP	FNg	FN
FP	TNg	FN
FP	FNg	TN

Fig. 6. Confusion metrics for three classes.

V. Results and Discussion

There are three experimental works carried out using Weka tools are shown in Table III, in order to investigate the stemmer method effect in Mauritania dialectal classification and to compare the performance of the Machine Learning techniques applied. In the first EXP (i), we combined the ArabicStemmerkhoja, the MultiStopwords, and the word tokenizer in order to construct an appropriate feature; Exp (ii) is the result of a combination of ArabicLightStemmer, multistop-word, and word tokenizer; The last one EXP(iii) was done of null Stemmer combined with multi Stop Words, and WordTokenizer. The accuracy of the three experimental works is illustrated in Fig. 7 and Table VII, which shows that three machine learning techniques (Random Forest, Logistic Regression, and Naive Bays Multinomial) were tested using training data at three different stages, with the result changing according to features used.

TABLE III. Combined Features

Exp	Feature set
i	WordTokenize + ArbicStemmerKhoja + MultiStopWords
ii	WordTokenize + Arabic light Stemmer + MultiStopWords
iii	WordsTokenize + Null Stemmer + MultiStopWords

Tables IV, V, and VI shows the results obtained by the Random Forest, NBM, and logistic techniques on the basis of the training data. It can be seen that the three classifiers managed to classify the positive class more than the others with better data by RF with Ligth StemmerArabic equal 98,5%; moreover, RF gets better results than others classified in three cases.

TABLE IV. EXP(I) Classification Results of each Class using Stem-Based (Ligth StemmerArabic)

		Class		
		Positive	Negative	Neutral
NBM	Precision	0,768	0,685	0,694
	Recall	0,731	0,726	0,686
	F-Measure	0,749	0,705	0,690
RF	Precision	0,985	0,974	0,934
	Recall	0,974	0,938	0,979
	F-Measure	0,980	0,956	0,956
Logistic	Precision	0,841	0,761	0,768
	Recall	0,840	0,776	0, 755
	F-Measure	0,841	0,768	0,762

The results obtained from exp(i) with Light Stemmer Arabic are shown in Table IV; this shows the performance evaluation measure for each selected class or sentiment (positive, negative, and neutral), so positive sentiment was ranked higher by RF.

Table V shows the results obtained when using Arabic Stemmer Khoja. This experience shows RF arrives at a significant number classified in all classes, followed by Logistic which is better for the positive, and negative classes than the neutral.

The results obtained during the exp(iii) indicated in Table VI show that RF and Logistic in terms of classification than NBM. However, the correctly classified number of the neutral class is less important here than the other classes.

TABLE V. EXP(II) CLASSIFICATION RESULTS OF EACH CLASS USING ROOT-BASED (ARABIC STEMMER KHOJA)

		Class		
		Positive	Negative	Neutral
NBM	Precision	0,766	0,681	0,679
	Recall	0,705	0,749	0,665
	F-Measure	0,734	0,713	0,672
RF	Precision	0,982	0,951	0,952
	Recall	0,968	0,956	0,961
	F-Measure	0,975	0,953	0,956
Logistic	Precision	0,863	0,812	0,755
	Recall	0,796	0,951	0,801
	F-Measure	0,945	0,804	0,777

TABLE VI. EXP(III) CLASSIFICATION RESULTS OF EACH CLASS USING NULL STEMMER)

		Class		
		Positive	Negative	Neutral
NBM	Precision	0,787	0,674	0,688
	Recall	0,705	0,746	0,686
	F-Measure	0,744	0,708	0,687
RF	Precision	0,985	0,945	0,946
	Recall	0,965	0,953	0,958
	F-Measure	0,975	0,949	0,952
Logistic	Precision	0,872	0,820	0,762
	Recall	0,853	0,796	0, 801
	F-Measure	0,862	0,808	0,781

Metric in the three experiments for the three classes given with the NMB and Logistic, it shows that the technique used is good for predicting the positive class, especially in experiment (i), and bad for predicting the negative class. Unlike RF who managed to predict all classes.

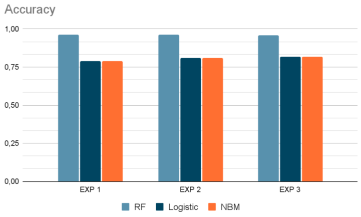


Fig. 7. Accuracy performance of the three algorithms.

Fig. 7 illustrates the Accuracy of algorithms given respectively by the three experiments applied.

TABLE VII. MODELS PERFORMANCE

	Accuracy		
EXP	RF	NBM	LOGISTIC
(i)	96.37	71.40	79.02
(ii)	96.14	70.63	80.79
(iii)	95.84	71.24	81.65

As shown in Table VII above RF and NBM algorithm was better in performance when using ArbicStemmerKhoja; however, Logistic gets better performance results with Null Stemmer. Overall, with the Light Stemmer Arabic feature, the RF algorithm had the highest accuracy rate compared to the NBM and Logistic algorithms.

TABLE VIII. COMPARISON OF EXPERIMENTAL RESULTS

Paper	**Dataset**	**Classification Algorithm**	**Best Accuracy**
[29]	Scrapping from tweets comments	SVM, NB, and K-nearest neighbour	67.19%
[30]	Moroccan sentiment analysis corpus	NB, SVM , and Maximum Entropy (ME)	82.5%
[31]	Tunisian sentiment analysis corpus	SVM, and NB	76.41%
Our approach	**Mauritania dialect analysis corpus**	**NBM, RF , and Logistic Regression**	**96.37%**

Table VIII illustrates the results obtained with previous work, which focuses on different dialects; Likewise, our approach also studies a dialect. However, the experimental study in this approach gave a result of 96.37% higher than those obtained by existing studies. Therefore it is a successful approach.

Diacritization and derivation or rootization of Arabic words are the limitations of this approach. We recommend that future research enhance algorithms by taking diacritization and all possible word lengths into account; So that the correct word meaning can be processed.

VI. CONCLUSION

This study essentially focuses on the Mauritanian dialect, looking at its morphology, structure, and meaning, with the aim of analyzing it using Machine Learning algorithms. In order to prove the classification of the Mauritanian dialect using ML algorithms, we experimented on a corpus of dialect words that gave satisfactory results, however, the study proved that the results obtained are influenced by the effect of stemmer methods;

In this article, three types of stemmer were tested with the objective of measuring and comparing their effect on the classification of dialect text, this process showed that the stemmer method "ArbicStemmerKhoja" is the most efficient with the NBM and RF algorithms in terms of prediction, unlike logistic which gives a better performance without stemmer.

These results will guide us to a deeper study of the language data in order to uncover sentiments behind his comments written in the Mauritanian dialect and find an accurate prediction.

ACKNOWLEDGMENT

First of all, I would like to thank God for allowing me to get through all the difficulties. Secondly, I would like to thank the members of the research team of which I am a member for their efforts during this work. Finally, I would also like to give special thanks to my supervisors for their guidance and advice.

REFERENCES

[1] P. P. Rokade and K. D. Aruna, "Business intelligence analytics using sentiment analysis-a survey," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 1, p. 613, 2019.

[2] M. E. M. El Arby Chrif, M. M. Saleck, A. C. M. N'Diaye, and E. B. M. Mahmoud, "Business intelligence models for e-government in mauritania: A survey," in *The International Conference on Artificial Intelligence and Smart Environment*. Springer, 2022, pp. 307–312.

[3] A. C. M. N'Diaye, M. E. M. E. A. Chrif, B. M. El Mahmoud, and O. El Beqqali, "Apply sentiment analysis technology in social media as a tool to enhance the effectiveness of e-government: Application on arabic and mauritanian dialect 'hassaniya'," in *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2021, pp. 1–5.

[4] A. Y. Muaad, G. H. Kumar, J. Hanumanthappa, J. B. Benifa, M. N. Mourya, C. Chola, M. Pramodha, and R. Bhairava, "An effective approach for arabic document classification using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 267–271, 2022.

[5] P. F. Kurnia *et al.*, "Business intelligence model to analyze social media information," *Procedia Computer Science*, vol. 135, pp. 5–14, 2018.

[6] A. Alshutayri, E. Atwell, A. Alosaimy, J. Dickins, M. Ingleby, and J. Watson, "Arabic language weka-based dialect classifier for arabic automatic speech recognition transcripts," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 204–211.

[7] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, 2022.

[8] H. Alshaer, B. Alzwahrah, and M. Otair, "Arabic text classification using bayes classifiers," *Int J Inform Syst Comput Sci*, 2017.

[9] 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.

[10] M. Errami, M. A. Ouassil, R. Rachidi, B. Cherradi, S. Hamida, and A. Raihani, "Sentiment analysis on moroccan dialect based on ml and social media content detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023.

[11] N. J. Thalji, E. Aljarrah, R. Rateb, and A. R. M. Al-Shorman, "New arabic root extraction algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.

[12] M. A. Humayun, H. Yassin, and P. E. Abas, "Dialect classification using acoustic and linguistic features in arabic speech," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, p. 739, 2023.

[13] T. Kanan, B. Hawashin, S. Alzubi, E. Almaita, A. Alkhatib, K. A. Maria, and M. Elbes, "Improving arabic text classification using p-stemmer," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 15, no. 3, pp. 404–411, 2022.

[14] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing arabic text on social media," *Heliyon*, vol. 7, no. 2, 2021.

[15] J. Younes, E. Souissi, H. Achour, and A. Ferchichi, "Un état de l'art du traitement automatique du dialecte tunisien [natural language processing of the tunisian dialect: a state of the art]," *Traitement Automatique des Langues*, vol. 59, no. 3, pp. 93–117, 2018.

[16] B. Abuata and A. Al-Omari, "A rule-based stemmer for arabic gulf dialect," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 2, pp. 104–112, 2015.

[17] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based arabic text categorization," *IEEE Access*, vol. 8, pp. 127 913–127 928, 2020.

[18] T. SSIT and B. BIT, "Document classification system using improvised random forest classifier."

[19] H. Elfaik *et al.*, "Leveraging feature-level fusion representations and attentional bidirectional rnn-cnn deep models for arabic affect analysis on twitter," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 462–482, 2023.

[20] T. Kanan, O. Sadaqa, A. Almhirat, and E. Kanan, "Arabic light stemming: A comparative study between p-stemmer, khoja stemmer, and light10 stemmer," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2019, pp. 511–515.

[21] H. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, and S. Al Muhairi, "Arabic light stemmer: A new enhanced approach," in *The Second International Conference on Innovations in Information Technology (IIT'05)*, 2005, pp. 1–9.

[22] M. G. Syarief, O. T. Kurahman, A. F. Huda, and W. Darmalaksana, "Improving arabic stemmer: Isri stemmer," in *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*. IEEE, 2019, pp. 1–4.

[23] S. Lang, F. Bravo-Marquez, C. Beckham, M. Hall, and E. Frank, "Wekadeeplearning4j: A deep learning package for weka based on deeplearning4j," *Knowledge-Based Systems*, vol. 178, pp. 48–50, 2019.

[24] L. Zhang, L. Jiang, C. Li, and G. Kong, "Two feature weighting approaches for naive bayes text classifiers," *Knowledge-Based Systems*, vol. 100, pp. 137–144, 2016.

[25] P. Langley and S. Sage, "Induction of selective bayesian classifiers," in *Uncertainty Proceedings 1994*. Elsevier, 1994, pp. 399–406.

[26] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad002, 2023.

[27] C. M. Richard, D. Poddubnyy, A. Deodhar, W. Bao, C. Parman, B. Porter, and E. Pournara, "Facteurs prédictifs de réponse au sécukinumab chez les patients atteints de spondylarthrite ankylosante: analyse par régression logistique et machine learning," *Revue du Rhumatisme*, vol. 87, p. A28, 2020.

[28] F. S. Alharithi, "Performance analysis of machine learning approaches in automatic classification of arabic language," 2023.

[29] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *2014 International Conference on Future Internet of Things and Cloud*. IEEE, 2014, pp. 579–583.

[30] A. Oussous, A. A. Lahcen, and S. Belfkih, "Improving sentiment analysis of moroccan tweets using ensemble learning," in *Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers 3*. Springer, 2018, pp. 91–104.

[31] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith, "Sentiment analysis of tunisian dialects: Linguistic ressources and experiments," in *Third Arabic Natural Language Processing Workshop (WANLP)*, 2017, pp. 55–61.