Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/18469
DOI URL: http://dx.doi.org/10.21474/IJAR01/18469

ISSN NO. 2320-5407

RESEARCH ARTICLE

AN IDENTIFICATION MODEL USED FOR ARABIC LIBYAN DIALECTS BASED ON MACHINE LEARNING APPROACH

Mohamed Abdeldaiem Mahboub[1], Tiruveedula Gopi Krishna[2] and Pyla Srinivasa Rao[3]

1. Department of Information Systems, Faculty of Information Technology University of Tripoli, Libya.
2. Adama Science and Technology University, Department of Computer Science and Engineering, Adama, Ethiopia.
3. Senior Manager, Cyber Security, Capgemini, India.

Manuscript Info

Manuscript History
Received: 25 January 2024
Final Accepted: 27 February 2024
Published: March 2024

Key words:-
Modern Standard Arabic Language, Libyan Dialects, NLP, Machine Learning

Abstract

In this research work we have especially studied both Modern Standard Arabic Language and Libyan Dialects. The focus of our study involved an in-depth analysis of both Modern Standard Arabic Language and Libyan Dialects through the lens of Natural Language Processing (NLP). Our primary objective was to assess the efficacy of a novel Machine Learning Model in enhancing performance and accurately categorizing the dataset for identifying Libyan dialects, while also addressing the preprocessing requirements of the diverse Libyan Dialects dataset. The core function of our developed Model was to transform the preprocessed Dialects dataset into its corresponding standard Arabic roots, recognizing that Libyan dialects are primarily spoken rather than written. Our identification model was specifically designed to navigate the challenges posed by these distinct Dialects, aiming to mitigate ambiguities that could potentially impact the model's effectiveness. The Arabic Libyan dialects identification model we proposed was tailored to leverage Natural Language Processing techniques to automatically determine the Arabic Libyan dialect present in a given text. This model aligns with the foundational principles of NLP, serving as a crucial initial step in a range of natural language processing applications such as machine translation, multilingual text-to-speech synthesis, and cross-language text generation. Our research paper provides a comprehensive overview of the Arabic Libyan dialects identification model, highlighting the utilization of feature representation techniques to train the proposed ML model effectively.

Introduction:-

Various Arabic dialect identification models primarily focus on word sentiment analysis through different methodologies, despite the shared cultural aspects among Arabic-speaking populations and the distinctiveness of the standard modern Arabic language, which exhibits a straightforward transition from spoken to written form once the fundamental rules of Arabic are applied. Research in Arabic linguistics, particularly in the context of Natural Language Processing (NLP), has delved into the exploration of Arabic language and its spoken variations, aiming to ascertain its potential significance within the realm of Machine Learning. Recent works have predominantly

Corresponding Author:- Pyla Srinivasa Rao
Address**:-** Senior Manager, Cyber Security, Capgemini, India.

concentrated on Arabic language morphology, cross-linguistic studies, computational linguistics, NLP, as well as broader issues like language universality and idiosyncrasies. The emphasis has been placed on the simplicity and coherence of the Arabic language, rather than its intricacies and ambiguities when juxtaposed with other languages like English. The development of an Arabic natural language processing model, potentially rooted in an Arabic dialect recognition framework, is of particular interest. Modern Standard Arabic serves as a lingua franca in news broadcasts across the Arab world, facilitating mutual comprehension among Arabic speakers from diverse regions. Nonetheless, conversational exchanges in local dialects may pose challenges in mutual understanding. Hence, the endeavor to construct a unified Libyan Arabic dialect Lexicon is pursued to enhance the efficacy of the proposed Machine Learning model.

Related Work:-

The limited research on Libyan dialects in terms of social media interaction and big data collection has sparked a growing interest in recent times. While there has been a considerable amount of research focusing on the identification of various Arabic dialects, the attention given to Libyan dialects has been relatively scarce. Upon reviewing the available literature on Arabic Libyan dialects, it becomes evident that only a few research papers have been dedicated to this specific area. Upon reviewing the literature available on Arabic Libyan dialects, it was observed that only a limited number of research papers have been dedicated to this area. For instance, J. Younes et al. [1] concentrated on the Tunisian dialect processing, aiming to develop corpora and dictionaries to identify its unique characteristics. Husien A. Alhammi and Kais Haddar [2] utilized an adjective priority scoring algorithm in their research to create a sentiment analysis system for categorizing Libyan dialect tweets into seven distinct categories. Similarly, Ashraf Elnajar et al. [3] conducted a survey following systematic review norms, focusing on computational approaches to dialectal Arabic identification and detection. Abir Masmoudi et al. [4] conducted experiments to evaluate three deep learning methods for the Tunisian Dialect within the context of Tunisian supermarkets. Fréha Mezzoudj et al. [5] introduced natural language processing on an Oranee textual corpus, integrating dialectal language models and Modern Standard Arabic. Ahmed Abdelali et al. [6] introduced QADI, an automatically collected dataset of tweets representing various country-level Arabic dialects across 18 different countries in the Middle East and North Africa region. Their dataset construction method involved applying filters to identify users from different countries based on their account descriptions and to exclude tweets in Modern Standard Arabic or containing inappropriate language. Muhammad Abdul-Mageed et al. [7] presented AraNet, a set of deep learning tools for processing Arabic social media. These tools, known as AraNet models, have the capability to predict various attributes such as age, dialect, gender, emotion, irony, and sentiment. In a similar vein, Mohamed Osman Hegazi et al. [8] focused on extracting information from Arabic text on social media. They proposed a comprehensive solution that addresses the challenges of preprocessing Arabic text in four stages: data collection, cleaning, enrichment, and availability. Youssef Fares et al. [9] investigated different techniques to measure the performance of Arabic dialect applications and resources. Abdullatif Ghallab et al. [10] conducted a systematic review of existing literature related to ASA. Mustafa Jarrar et al. [11] developed morphologically annotated corpora for Yemeni, Sudanese, Iraqi, and Libyan Arabic dialects, collected from various social media platforms, particularly twitter. Diaa Salama Abdelminaam et al. [12] introduced a framework for opinion mining of Arabic dialects on Twitter, consisting of two major components. Christoph Tillmann et al. [13] focused on improving sentence-level dialect classification between Egyptian Arabic and Modern Standard Arabic. Their approach utilized binary feature functions based on task-specific knowledge. Mohamed Elaraby Muhammad Abdul-Mageed [14] addressed limitations in the Arabic Online Commentary (AOC) and conducted a benchmark of the data. They also empirically tested six different deep learning methods. Iman S. Alansari [15] explored various studies on AD for the development and understanding of conceptual deep learning models to detect and classify Arabic dialects. Their model utilized Convolutional Neural Networks (CNNs) to capture semantic, syntactic, and phonological characteristics of the desired dialects. Mohamed Abdeldaiem Abdelhadi [16] proposed an advanced model utilizing Information Retrieval techniques to transform preprocessed Arabic dialect data into Arabic roots, ultimately converting it into Arabic Modern Standard Language. This model was designed for the purpose of constructing an Arabic Dialects Lexicon.

Methodology:-

Our methodology based on modern techniques of knowledge management systems and also AI techniques (Machine Learning) to optimize the performance of our proposed model. We have used relational data base to build up a Lexicon for Arabic Libyan Dialects by means of data science techniques. Arabic Libyan Dialects Lexicon will be used in our model as master data base containing most different Arabic Libyan Dialects words within its meaning in

Modern Standard Arabic Language. We have used the following methods to prepare the dataset collection for our propped model as follows:

Creating XML Files for Libyan Dialects Selection

We have created in our model an XML files for each Arabic Libyan Dialect containing some data types such as (City Code, City Name, and Geographical place); Type of Dialect; to organize the Libyan Dialects words which will be preprocessed into common XML data files [16]. Each template has some fields which are defined as; International Libyan Code a unique Local Code, City Name. It will be used to assign each identified Libyan Dialect to its right place in the classified Dialects Data Base as shown in (Table1).

Table1:- XML files for each Arabic Libyan Dialect.

no	International Libyan Code	Local code	City Name	Geographical place	Type of Dialects
1	00218	021	Tripoli	Western coast	Mixed (Arabic + European)
2		061	Benghazi	Eastern Coast	Arabic
3		071	Sebha	Southern	Mixed (Arabic + African)
4		0581	Waddan	Middle of Land	Arabic

Creating Libyan Dialects Dataset

The second important issue was to create a Relational Data Base to store the preprocessed Dialects words as shown Figure2. The main data fields in the Data Base design are designed dynamic relational Data Base. The Data fields are as shown in the (Table2), which includes the City name, Text in Dialect, City Code Id, and Text in Modern Standard Arabic. This Data Base example explains the idea of our Model for Arabic Libyan Dialect Identification in term of simplicity and convergence to Arabic Standard Modern Language. As we have stated in the related work of our research; most literature dealt with Natural Language Processing Models; which are generally uses techniques of Speech Recognition tools to convert speech or sometimes, Voice-IP (Audio) data files to somehow, Speech to Text data files [3, 5, 6] . Our methodology based on techniques of knowledge Base Management Systems and also Natural Language Processing approach to optimize the performance of our model. We have used RDB to build up a Lexicon for Arabic Libyan Dialects by means of Machine learning methods [16]. The Arabic Libyan Dialects Lexicon will be used in our proposed model as dataset containing most different Arabic Libyan Dialects words within its meaning in Modern Standard Arabic Language (MSA). We have an example illustrates relational Data Base system as shown in the Table2. In the preprocessed text in Dialect layout, we have used specific rules to define the Arabic Libyan Dialects Lexicon analysis and design. Each Dialect word in the proposed Lexicon will have; the following attributes: “City name”, “City code”, and “text meaning” in Modern Standard Arabic Language. To build up our data base, we have integrated our model with two different tools based on Desktop and Website, to share our scientific work with other research groups whom they may have the same interests in Arabic Libyan Dialects Identification modeling.

Table 2:- An example illustrates relational DB.

NO	CITY NAME	TEXT IN DIALECTS	CITY CODE	TEXT IN MODERN ARABIC LANGUAGE
1	Tripoli	علاش انت ديما اتجي للعمل متأخر؟	021	لماذا انت تأتي الى العمل دائما متأخرا؟ Why do you come to work always late?
2	Benghazi	كنك انت ديما اتجي للعمل متأخر؟	061	
3	Sebha	ليش انت ديما اتجي للعمل تالي ؟	071	
4	Waddan	خيرك انت ديما اتجى للخدمة موخر؟	0581	

As it was introduced in the text of Libyan dialects, علاش انت ديما اتجي للعمل متأخر؟ Which is spoken in the area of western coast of Libya (Tripoli City) has different context over other type of Libyan dialects, such as (Benghazi, Sebha, and Waddan). But after the preprocessing of text in dialect; it will be converted to MSA as one unified “Sentence of MSA”; لماذا انت تأتي الى العمل دائما متأخرا actually, we could have one to many sentences meaning related to Modern Standard Arabic Language, which can be understood in all regions of Libya. We have built our Arabic

Libyan Dialects Lexicon to convert all possible words from Arabic Libyan Dialects to Modern Standard Language. (Table3) illustrates the layout of Data Base. As shown in this example of Libyan Arabic Dialects. It is now very clear, how important to choose the best NLP methods to optimize the model performance even before getting used the main model tasks.

Arabic Libyan Dialects:-

Libyan Arabic Dialects encompass the spoken Arabic Language within the Libyan Ethnic groups, including the Arab, Amasigh, Touareq, and Tobu communities. These dialects serve as a means of informal communication among the locals. Geographically, (Figure1)[2], illustrates the distribution of Libyan Arabic Dialects across different regions. While Modern Standard Arabic (MSA) holds the status of the official language in Libya and is predominantly used in the cultural and educational systems, it is primarily a written language rather than a spoken one. On the other hand, the informal spoken dialects play a crucial role in daily life interactions, even extending to FM radio stations and Libyan Television shows. These dialects, being actually spoken and not written, are considered authentic native languages. They can be further classified into distinct groups based on their variations, as outlined in the descriptive Libyan Dialects scheme presented in (Table 3).

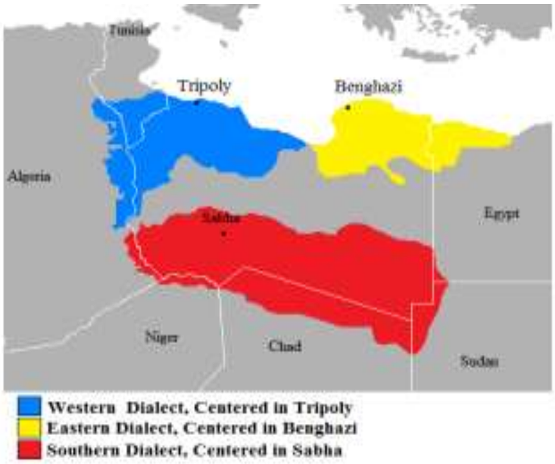

Tunisia
Tripoly
Benghazi
Algeria
Sabha
Egypt
Niger
Chad
Sudan
🟦 Western Dialect, Centered in Tripoly
🟨 Eastern Dialect, Centered in Benghazi
🟥 Southern Dialect, Centered in Sabha

Figure 1:- Distribution of Libyan Arabic Dialects geographically.

Table 3:- Descriptive Libyan Dialects scheme.

TABLE I. No	**TABLE II.** Description of Arabic Libyan Dialects Regions	**TABLE III.** Type of Dialects	**TABLE IV.** City local Code **TABLE V.**
TABLE VI. 1	**TABLE VII.** Western Coast region encompasses all cities from Misrata to Tripoli and Zwara to Jabel Nafousa, Gharian, Ghadames, Mezda, and Shwearif. **TABLE VIII.**	**TABLE IX.** Trabelsia Dialect **TABLE X.** اللهجة الطرابلسية **TABLE XI.** (Mixed Dialects)	**TABLE XII.** 021 (Tripoli) **TABLE XIII.** Western Region
TABLE XIV. 2	**TABLE XV.** Middle of Land encompasses all cities from Abugrain to Sirt and Aljufra **TABLE XVI.**	**TABLE XVII.** (Pure Arabic) **TABLE XVIII.** اللهجة العربية البدوية **TABLE XIX.** Dialects **TABLE XX.**	**TABLE XXI.** 057 (Sirt) **TABLE XXII.** Middle Region
TABLE XXIII. 3	**TABLE XXIV.** Eastern Coast region encompasses all cities from Brega to Benghazi and Tobruq to Alwahat, Kufra. **TABLE XXV.**	**TABLE XXVI.** (Pure Arabic) **TABLE XXVII.** اللهجة العربية البدوية **TABLE XXVIII.** Dialects **TABLE XXIX.**	**TABLE XXX.** 061 (Benghazi) **TABLE XXXI.** Eastern Region

TABLE XXXII. 4	**TABLE XXXIII.** Southern of Land encompasses all cities from Bouanice to Sebha and Wadi Alshati, Obari, Murzeq, Zowella, Qatroun,	**TABLE XXXIV.** Fezania Dialect **TABLE XXXV.** اللهجة الفزانية **TABLE XXXVI.** (Mixed Dialects)	**TABLE XXXVII.** 071 (Sebha) **TABLE XXXVIII.** Southern Region

Arabic Libyan Dialects Analysis Methods

Current research on Arabic Libyan Dialects has predominantly employed various machine learning techniques, information retrieval, and NLP methods to manage the vast quantities of unstructured texts in Arabic Libyan Dialects produced by online users, particularly on social media platforms. Typically, the analysis is conducted at three distinct levels: document-level, sentence-level, and aspect-based level, each offering unique insights into the language patterns and structures present in the texts. The analysis process involves the utilization of two primary approaches: the machine learning approach (ML) and the linguistics-based approach, each offering its own set of advantages and limitations in handling the complexities of Arabic Libyan Dialects texts. Figure2 illustrates the main approaches available for analysis, highlighting the flexibility in selecting the most appropriate approach based on the specific research objectives and requirements [11, 16].

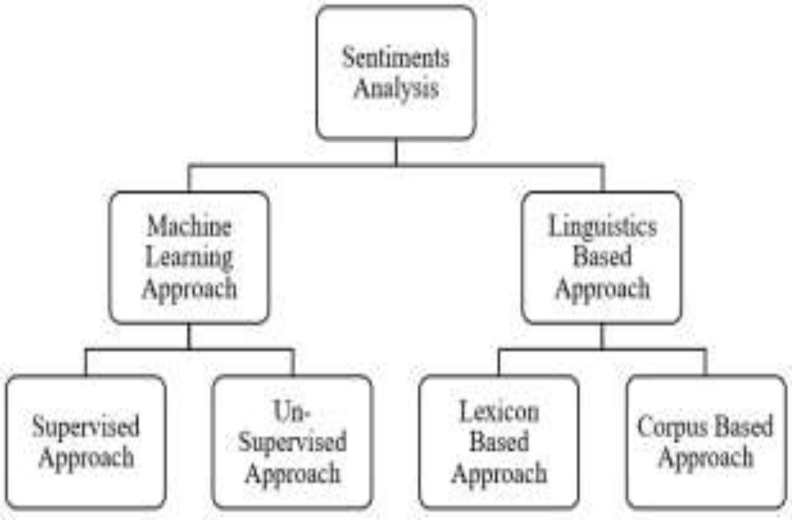


Figure 2:- Approaches for Sentiments Analysis.

Linguistics Based Approach

Linguistics based approach or Lexicon-based approach consists in building lexicons of classified words. In this respect, [2] relied on a lexicon-based approach to be able to construct and assess a very large sentiment lexicon including about 15k Arabic terms. To evaluate the lexicon we have labeled dataset of 1500 tweets and reached an accuracy rate of 82%. In our attempt to construct a new Arabic Libyan lexicon, [3]. The new lexicon is composed of 500 positive tweets, 500 negative tweets and 500 neutral tweets, for a total of 1500 tweets achieving an accuracy of 87%.In this research work, we have used our approach to implement the Lexicon of Libyan dialects as shown in (Figure3), we have used training set by a small collected corpus of described Dialects files from different resources in four local cities (Tripoli, Benghazi, Sebha, and Sirt) [16]. The corpus of Training data set composed of some different dialects which are preprocessed. As a matter of fact, we have found that our simple approach has provided good results for our model. We have also randomly selected small size of corpus, as preprocessed Test data set of Arabic Libyan Dialects, and have organized the Dialects words automatically in this phase, to investigate the model reliability.

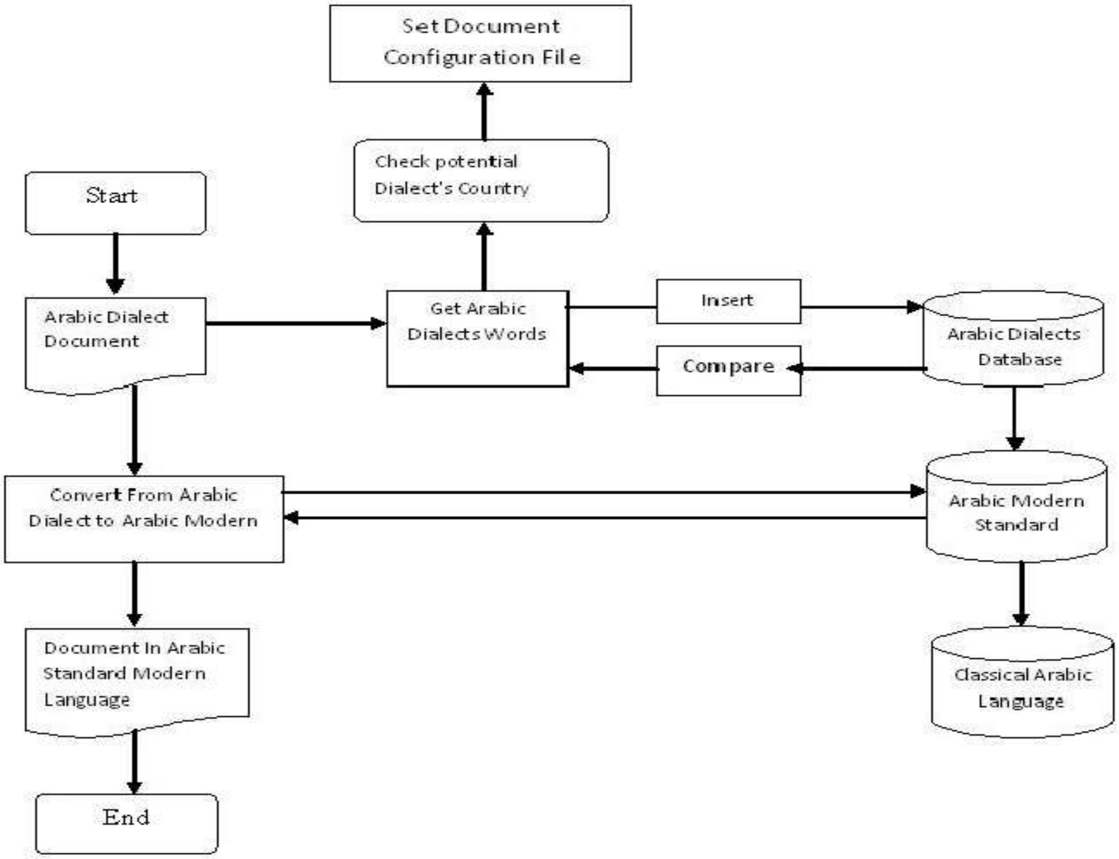

Set Document Configuration File
Check potential Dialect's Country
Start
Arabic Dialect Document
Get Arabic Dialects Words
Insert
Compare
Arabic Dialects Database
Convert From Arabic Dialect to Arabic Modern
Arabic Modern Standard
Document In Arabic Standard Modern Language
Classical Arabic Language
End

Figure 3:- Building new Lexicon of Libyan dialects.

Proposed Machine Learning Model:-

The structure of our model was simplified in order to comprehend the underlying concept of the research focused on identifying dialects. This was achieved by converting preprocessed data from Libyan dialects of Arabic into modern standard Arabic, which ultimately enhanced the performance of a model based on Arabic natural language processing. A crucial aspect of modern Arabic lies in its formalities, where words are formed from roots of three or more letters that are modified according to specific rules to create groups of words with interconnected meanings. In Arabic, the creation of extensive word families involves the addition of prefixes, infixes, and suffixes in patterns that are sequentially placed around the root of three letters. To address the diversity of Arabic dialects and their impact on modern Arabic, it is essential to introduce a unified vocabulary of Arabic dialects. The method was organized according to (Figure2), outlining the main components of the model, and the research suggests utilizing machine learning models like SVM, LR, and NB. The testing of these models with new data, along with the evaluation using performance metrics, was illustrated in Figure4 to determine the most suitable model. The model was specifically designed to efficiently identify and classify Libyan Arabic dialects, utilizing a machine learning approach and a lexicon of Libyan dialects. It was structured to capture the unique characteristics and distinctions of the Libyan dialects within the dataset, encompassing six machine learning stages as detailed in (Table4).

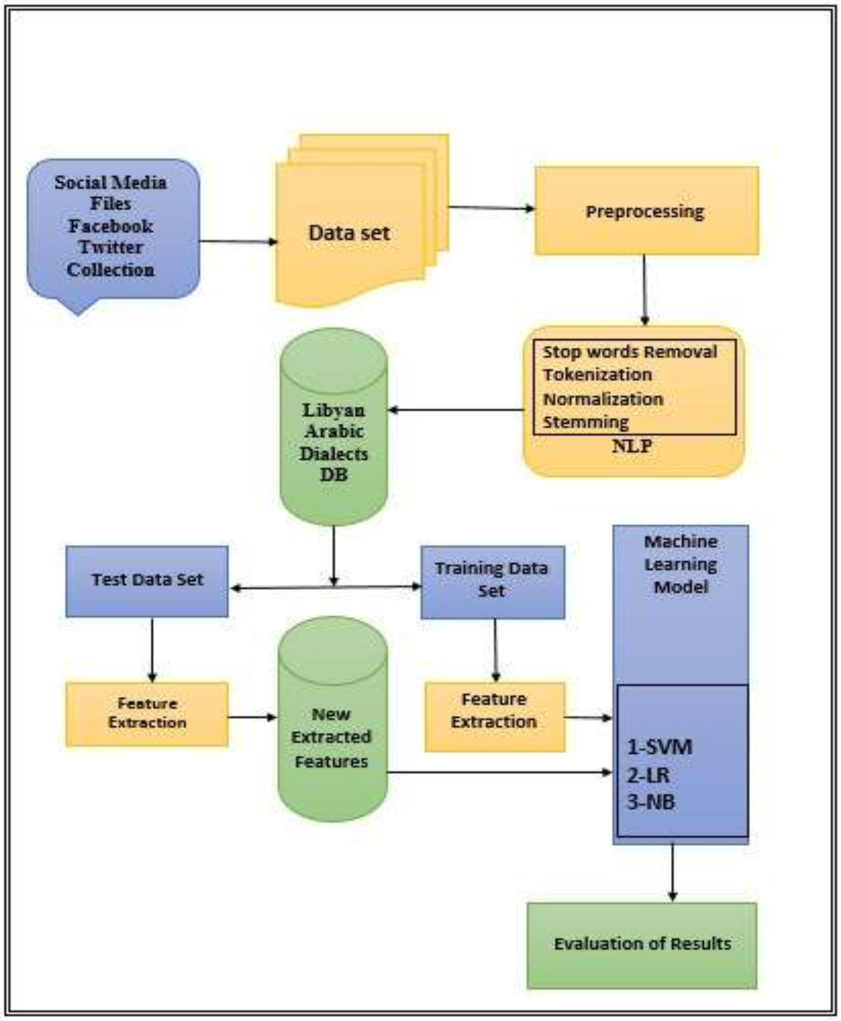


Figure 4:- Proposed ML model for Arabic Libyan Dialects.

Dataset

Social networking platforms such as Facebook and Twitter are widely used by individuals across the globe. In this study, we focused on identifying the most popular Facebook pages with over 100k followers. It was observed that Libyan community groups predominantly communicate in local dialects through their posts and comments on these pages. The Facebook pages were categorized based on the cities they represented, namely Tripoli, Benghazi, Sirt, and Sebha. A manual collection of approximately 15000 posts and comments was conducted, with the exclusion of commenters' names to maintain anonymity and privacy. Each group was then labeled, resulting in a dataset containing 5000 text terms for further analysis and classification into three equally distributed classes. [2, 16].

Preprocessing

Data preprocessing involves transforming raw data into a more refined and organized format. When dealing with data extracted from social media platforms, it is common to encounter special words, URLs, emoticons, punctuation marks, and unnecessary terms that need to be cleaned. Natural Language Processing (NLP) techniques such as Tokenization, Stop word removal, Normalization, Stemming, and part-of-speech tagging are utilized to improve the structure of text sequences and minimize irrelevant elements. The preprocessing steps also include the elimination of numbers, URLs, hashtags (by removing the # symbol), as well as special characters like emojis, Arabic diacritics, and short words with only two characters. Additionally, sentences with six words or more are separated during the data collection process to enhance the analysis. The outcomes of the various preprocessing stages conducted on the dataset are detailed in Table 5 for reference and evaluation [11]. The results of different steps of preprocessing performed on our dataset is illustrated on (Table5).

Feature extraction

Within this research, we have employed the TF-IDF method, which stands for term frequency-inverse document frequency, a widely used approach for assessing the significance of a particular term within a document. The fundamental concept of TF-IDF involves assigning a numerical value, or weight, to each word in a document based on its frequency within that document, while also taking into account how frequently the word appears across all documents. Consequently, words that are highly prevalent within a single document will be assigned lower weights, in contrast to words that are more unique and relevant to the content of the document. This methodology was introduced as a solution to the limitations of Bag of Word models, aiming to provide a more nuanced and accurate representation of the importance of individual terms within a document. By incorporating both term frequency and inverse document frequency, TF-IDF offers a more sophisticated approach to text analysis, allowing for a more precise evaluation of the significance of words based on their occurrence within a specific document and across a broader corpus of documents. [12].

Classification Methods

After extracting the features of a particular text using the methods discussed in the previous section, they are used as input to determine the dialect of the dataset using traditional Machine Learning algorithms like Support Vector Machine (SVM), Decision Rules (LR), and Naive Bayes (NB) as outlined in (Table4). These methods are typically implemented in the Machine Learning Model using the techniques specified in (Table4) [13].

Table 4:- Methods used in ML model.

[1] **No**	[2] **Techniques**	[3] **Methods**
[4] [5] **1**	[6] **Natural Language Processing (NLP)**	[7] **Tokenization** [8] **Part-of-Speech Tagging** [9] **Stop word removal** [10] **Stemming** [11] **Normalization**
[12] [13] **2**	[14] **Feature Engineering**	**I. Clustering.** **II. Classification.** **III. Vectorization**
[15] **3**	[16] **Neural Network**	**I. Support Vector Machine** **II. Logistic Regression** **III. Naïve Bayesian.**
[17] [18] **4**	[19] **Language Models**	**I. Word2Vec** **II. Glo2Ve**
[20] [21] **5**	[22] **Optimization Techniques**	**I. Gradient Descent** **II. Back propagation**
[23] [24] **6**	[25] **Evaluation Techniques**	**I. Precision** **II. Recall** **III. F-Measure**

Experiment Model:-

In our experimental framework, we employed a grid search technique utilizing three distinct machine learning classifiers, namely SVM, multinomial Naive Bayes, and Logistic Regression, as detailed in Table4. To ensure the selection of optimal hyperparameters, we relied on our expertise and experience, setting the default parameters for the Multinomial Naive Bayes classifier. Following a grid search on the TF-IDF vectorizer data from the Sklearn library1, we determined the maximum number of elements to be considered. The accuracy assessment was based on the TF-IDF feature count, with the maximum number of elements established post grid-search on the TF-IDF vectorizer from Sklearn library1 [12].. Our research methodology involved the utilization of an Experiment Model to assess the model's performance during the implementation phase. The training set was created by partitioning a small corpus of described Dialects files with the training corpus containing various dialects as outlined in Table5. Notably, our straightforward Experiment Model yielded promising outcomes for the preprocessing of Arabic Dialects data, as indicated in (Table5). Additionally, we automated the organization of Dialects words in this phase, with (Figure5) illustrating the distribution of Libyan Dialects for investigating the experiment model's efficacy.

Table 5:- Libyan Dialects corpus for Training and Test data set Sample.

No	Local City	No of text sequence	Training Data Set	Test Data Set
1	Tripoli	2466	45k	35k
2	Benghazi	3422	40k	25k
3	Sebha	4234	35k	20k
4	Sirt	4432	33k	15k

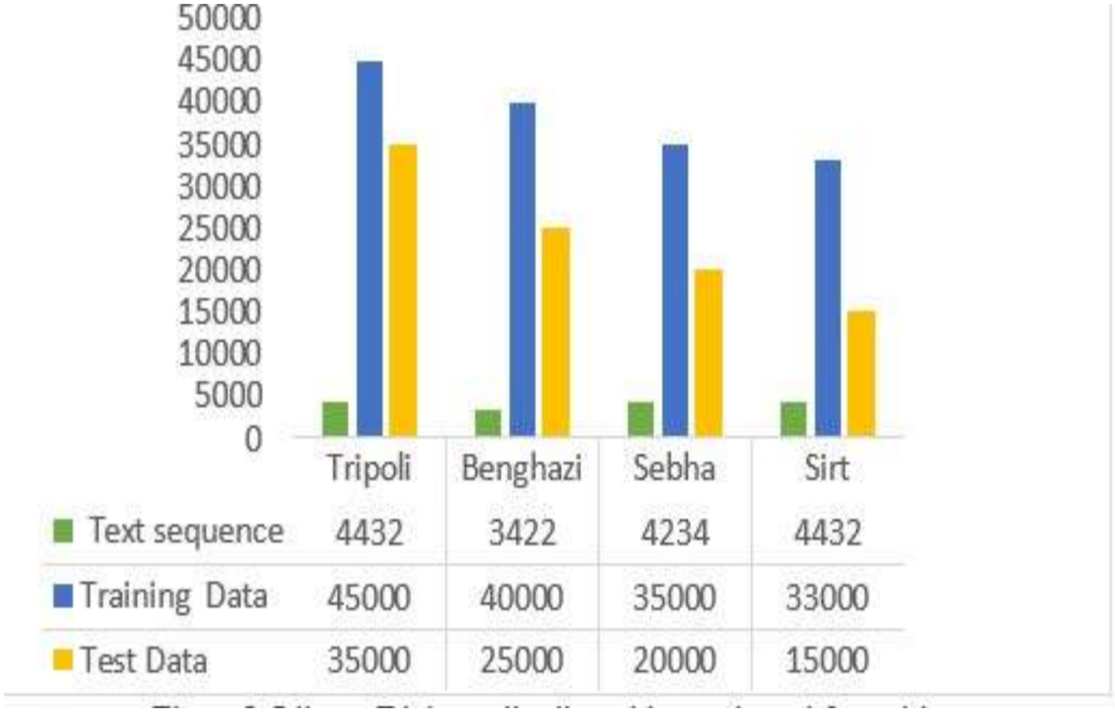


Figure 5:- Libyan Dialects distributed into selected four cities.

The study utilized a small training set, shown in (Figure3), consisting of described Dialects files from various dialects outlined in (Table5). The Experiment Model used in this research showed promising results for the proposed model. A small corpus randomly selected from preprocessed Arabic Dialects data set, as shown in (Table6), was used for experimentation. The test data set included around 15,000 posts manually annotated into seven polarities to evaluate the system's performance. The Model's accuracy wasmanually assessed compared to the baseline dataset test. The results indicate that the simple Experiment Model usedin this study had positiveoutcomes, demonstrating the effectiveness of the proposed model in analyzing Arabic Dialects data.

Table 6:- Sample of different Arabic Libyan Dialects used in experiment model.

No	Local Dialects	Text Sequence in Libyan Dialects	Dialect Words	Dialect Text conversion to MSA Language	Translated Text (MSA) to English
1	**Tripoli**	غدوة قالوا اول يوم صيام فى رمضان ,زعما الناس كلهم صايمين ,هلبا مرات فاتوا, ماصاموش الناس فى ليبيا مع بعضهم علاش أجماعة الشرق والجنوب مايبوش البلاد تتوحد؟	غدوة- زعما هلبا ماصاموش علاش أجماعة مايبوش تتوحد	قالوا غدا اول يوم صيام فى رمضان هل كل الناس صيام؟ عدة مرات سابقة لم يصم الناس فى ليبيا مع بعض لماذا اهل الشرق والجنوب لايريدون ان تتوحد البلاد ؟	**They said: Tomorrow is the first day of fasting in Ramadan. Are all people fasting? Several times in the past, people in Libya did not fast together. Why do the people of the East and South do not want the country to unite?**
2	**Benghazi**	بكرة.. قالوا.. اول يوم صيام	بكرة أزعما العرب	قالوا غدا اول يوم صيام	**They said: Tomorrow is**

		فى رمضان أزعما العرب كلهم صايمين , مرات واجدة أهلبن , ماصاموش العرب فى ليبيا مع بعضهم. كنهم عرب امغرب مايريدوش البلاد أتوحد؟	مرات واجدة أهلبن ماصاموش العرب كنهم مايريدوش أتوحد	فى رمضان هل كل الناس صيام؟ عدة مرات سابقة لم يصم الناس فى ليبيا مع بعض لماذا اهل الغرب والجنوب لايريدون ان تتوحد البلاد ؟	**the first day of fasting in Ramadan. Are all people fasting? Several times in the past, people in Libya did not fast together. Why do the people of the East and South do not want the country to unite?**
3	**Sebha**	غدوة.. قالوا اول يوم صيام فى رمضان , زعم العرب كلهم صايمين, كميين مرة فاتن ماصاموش الناس فى ليبيا مع بعضهم. خيرهم عرب الشرق والغرب مايبوش البلاد تتحد؟	غدوة زعم العرب كميين مرة فاتن ماصاموش خيرهم مايبوش تتحد	قالوا غدا اول يوم صيام فى رمضان هل كل الناس صيام؟ عدة مرات سابقة لم يصم الناس فى ليبيا مع بعض لماذا اهل الشرق والغرب لا يريدون ان تتوحد البلاد ؟	**They said: Tomorrow is the first day of fasting in Ramadan. Are all people fasting? Several times in the past, people in Libya did not fast together. Why do the people of the East and South do not want the country to unite?**
4	**Sirt**	غدوة قالو اول يوم صيام فى رمضان , زعم العرب كلهم صايمين , كميين مرة فاتن ماصاموش الناس فى ليبيا مع بعضهم. خيرهم عرب الشرق والغرب مايبوش البلاد تتحد؟	غدوة زعم العرب كميين مرة فاتن ماصاموش خيرهم مايبوش تتحد	قالوا غدا اول يوم صبام فى رمضان هل كل الناس صيام؟ عدة مرات سابقة لم يصم الناس فى ليبيا مع بعض لماذا اهل الشرق والغرب لا يريدون ان تتوحد البلاد ؟	**They said: Tomorrow is the first day of fasting in Ramadan. Are all people fasting? Several times in the past, people in Libya did not fast together. Why do the people of the East and South do not want the country to unite?**

We have used this example in our experiment to run machine learning models. As known, namely Support Vector Machines (SVC), Logistic Regression, Naïve Bayes classifiers used for multivariate Bernoulli models (Bernoulli NB), and Naive Bayes classifier for multinomial models (Multinomial NB).We have in our experiment calculated the Accuracy (A), Precision (P), Recall (R), and the normal F1-measure for the evaluation of our Model as shown in (Table7).

Table 7:- Ppecial words used in Libyan dialects.

No	Word	Pronunciation	Meaning
1	علاش	Alaash	Why?
2	هلبا	Halba	Too much
2	زعما	Zaamma	You think so
4	واجدة / أهلبن	Wajed / Gone away	Too much
5	كنهم	Kenhom	Why they are so?
6	مايريدوش	Mayeroudosh	They do not want it so
7	كميين مرة	Kammen Marah	How many times?
8	مايبوش	Mayabuosh	They don't want it so
9	غدوة	Ghedwa	Tomorrow
10	ماصاموش	MaSamoush	They don't fasten

Evaluation:-

The entire dataset was partitioned into four sets of equal size in order to assess performance. The classifier was trained on three of these sets, while the remaining set was utilized for testing purposes [11]. In relation to the fifth stage, the performance of the SVM classifier was measured using selected evaluation metrics, namely Accuracy, Precision, and Recall. These metrics were evaluated using the equations provided.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

True positives (TP) refer to correctly classified inputs in a data test that belong to the positive class, while true negatives (TN) are accurately classified inputs in a data test that belong to the negative class. False positives (FP) occur when inputs are incorrectly classified as positive when they belong to the negative class, and false negatives (FN) occur when inputs are wrongly classified as negative when they should be positive. Support Vector Machine (SVM) is recommended for handling large textual features due to its basis in structural risk minimization. In a City-level dialect identification task, we used an SVM classifier. Logistic Regression was used to assess the significance of each independent word based on its weight. Naive Bayes classifier is popular for text classification due to its speed and ease of use. We used TF-IDF with character (2-4)-grams and word (1-3)-grams as features for training our ensemble classifier. This combination of techniques and classifiers as shown in (Table8) ensures a comprehensive and efficient methods to text classification tasks. [10, 12].

Table 8:- ML Classifiers results for dialects dataset.

No	Classifier Model hyperparameters	Linear	Balanced	Balanced
1	SVM	24660	45000	35000
2	Logistic regression	34220	40000	25000
3	Naive Bayes	22340	35000	20000

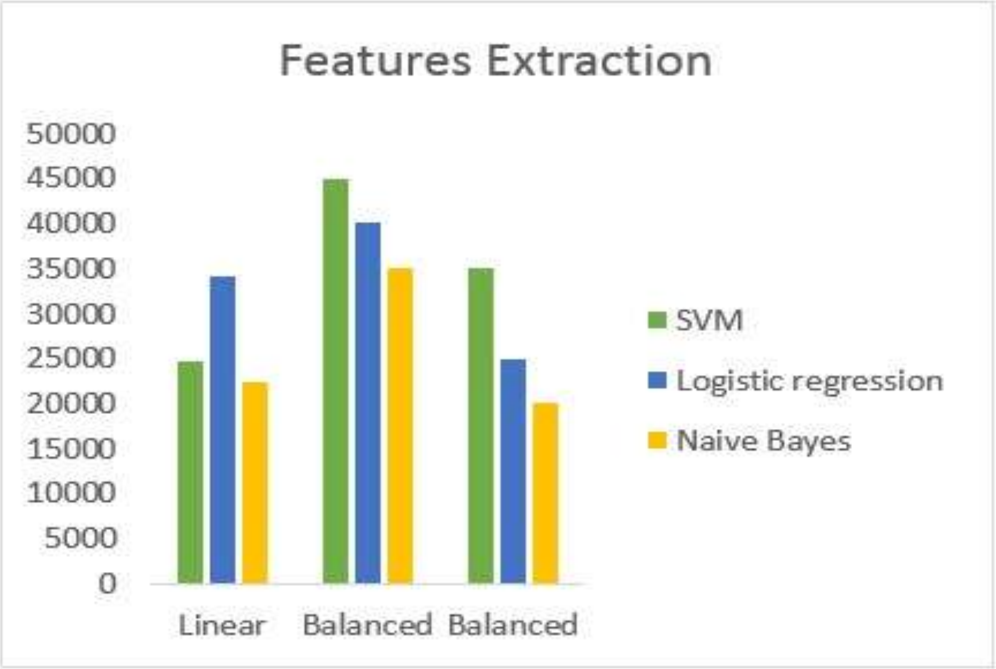


Figure 6:- Feature extraction for both Training &Test data set.

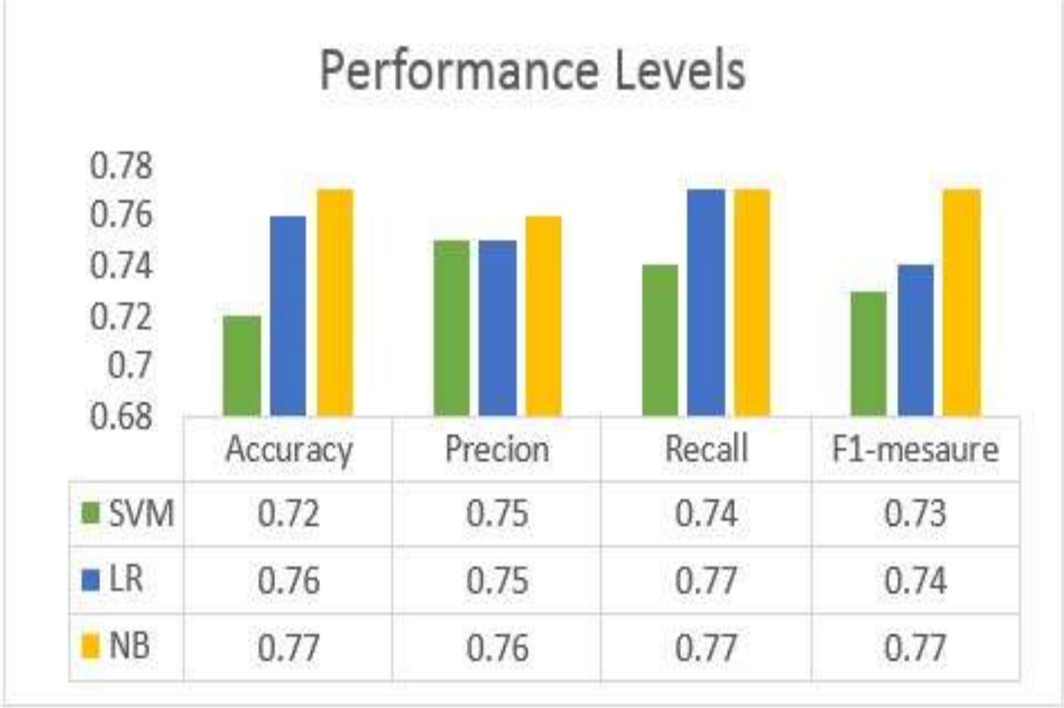
Results Analysis:-

Our study's experimental results suggest that using a machine learning approach leads to significantly better outcomes compared to current methods. We evaluated performance based on F-Measure, Accuracy, Precision, and

Recall, with the Macro Averaged F-score as the primary metric. Table9 illustrates the effectiveness of various Machine Learning models in terms of F1-Measure and accuracy. The experimental results were crucial in assessing classifier effectiveness, particularly in exploring different structures and representations within an SVM classifier framework. Comparing TF-IDF and TF weighting schemes, we found that classifiers performed better with TF-IDF in terms of accuracy and recall. The proposed model achieved the highest precision, correctly classifying a significant number of positive documents. While the results in Table9 provide valuable insights, they may not be definitive as outcomes could vary across datasets. Feature selection techniques were used in this experiment, leading to the selection of TF-IDF as the vector representation due to its 77% accuracy rate as shown in (Figure7).

Table 9:- The performance of the Libyan Arabic Dialects by ML model.

Classifier	Accuracy	Precision	Recall	F1--Measure
SVM	0.72%	0.75%	0.74%	0.73%
LR	0.76%	0.75%	0.77%	0.74%
NB	0.77%	0.76%	0.77%	0.77%



	Accuracy	Precion	Recall	F1-mesaure
🟩 SVM	0.72	0.75	0.74	0.73
🟦 LR	0.76	0.75	0.77	0.74
🟨 NB	0.77	0.76	0.77	0.77

Figure 7:- Performance level for Arabic Libyan dialects.

The results of our experimental model show that the Dialects Identification Model has successfully proven its validity. The outcomes from the Libyan Dialects Identification Model, as shown in (Table9) for both Training Data and Data Test set, are highly significant. Moreover, we have applied classifier methods with the same criteria for the identification model focusing on Modern Standard Arabic Language, as outlined in (Table10). The overall performance has significantly improved in terms of Accuracy, with a 0.88% increase as shown in (Figure8). Additionally, all other metrics have also seen a 0.87% rise, contributing to the overall improved performance of our proposed model.

Table10: the performance of the proposed ML model.

Classifier	Accuracy	Precision	Recall	F1--Measure
SVM	0.82	0.85	0.84	0.86

LR	0.84	0.87	0.86	0.84
NB	0.88	0.87	0.88	0.87

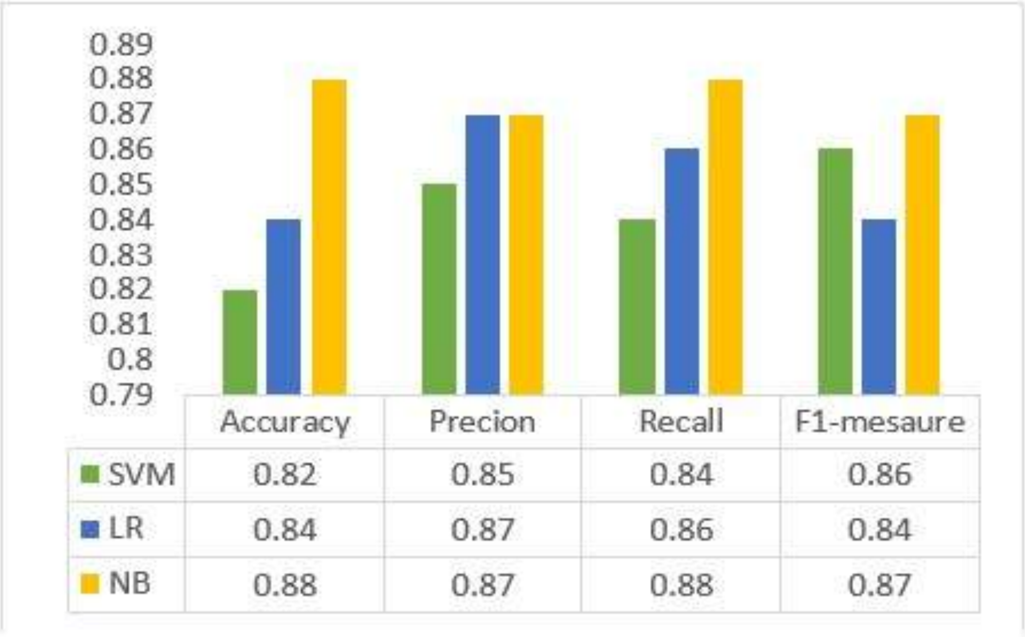


Figure 8:- Performance level based on MSA ML model.

Conclusion and Future Work:-

We have conducted different experiments in which we tried different preprocessing procedures and several feature combinations for model training and combined different machine learning approach such as (Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes) to build a strong Arabic Libyan Dialects identification model based on new lexicon for Libyan Arabic Dialects. For future work, we will explore the word embedding features in which sentiment analysis can be performed. No works on Libyan dialect sentiment analysis has been used to investigate a fine-grained sentiment analysis model for classifying Libyan Dialects. Our proposed model has a great result in Accuracy of 88 % on the test data set.

Conflicts of Interest:-

The authors declare that they have no conflicts of interest.

Authors Contributions:-

Mohamed Abdeldaiem Mahboub, Studied and executing the entire research work, using proper research methodology of the research work. Tiruveedula Gopi Krishna, as per the scientific principles. Pyla srinivasa Rao, Overall Review and understanding, programs verifications.

Acknowledgements:-

The work was supported by The University of Tripoli through.The authors gratefully acknowledge the Technicians of Laboratory for field assistance and providing technical supportsfor this Work

Bibliography:-

1. J.Younes et al.”Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web”, pp. 3–14, 2015, doi: 10.1007/978-3-319-24800-4_1.
2. Husien A. Alhammi and Kais Haddar,”Building a Libyan Dialect Lexicon-Based Sentiment Analysis System Using Semantic Orientation of Adjective-Adverb Combinations”, International Journal of Computer Theory and Engineering, Vol,DOI:10.7763/IJCTE.2020.V12.1280.2020.

3. Ashraf Elnajar et al.” Systematic Literature Review of Dialectal Arabic: Identification and Detection”, IEEE Access • February 2021, DOI:10.1109/ACCESS.2021.3059504.
4. Abir Masmoudi et al.” Deep Learning for Sentiment Analysis of Tunisian Dialect” 2021, pp. 129–148, doi: 10.13053/CyS-25-1-3472.
5. Fréha Mezzoudj et.al,” Arabic Algerian Oranee Dialectal Language Modelling Oriented Topic”, International Journal of Informatics and Applied Mathematics, e-ISSN: 2667-6990 Vol. 2, No. 2, 1-14.
6. Ahmed Abdelali et al.” Arabic Dialect Identification in the Wild”, arXiv: 2005.06557v2 [cs.CL] 15 May 2020.
7. Muhammad Abdul-Mageed et al. “A Deep Learning Toolkit for Arabic Social Media”, arXiv: 1912.13072v2 [cs.CL] 11 Apr 2020.
8. Mohamed Osman Hegazi et.al.” Preprocessing Arabic text on social media”, February 2021 https://doi.org/10.1016/j.heliyon.2021.e06191 .
9. Youssef Fares et.al.” Arabic Dialect Identification with Deep Learning and Hybrid Frequency Based Feature August 1, 2019, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 224–228.
10. Abdullatif Ghallab et al, “Arabic Sentiment Analysis: A Systematic Literature Review”, 2020, Article ID 7403128, https://doi.org/10.1155/2020/7403128.
11. Mustafa Jarrar et al,” Lîsan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with morphological Annotations”, 17 Dec 2022, arXiv: 2212.06468v2 [cs.CL] 17 Dec 2022.
12. Diaa Salama Abdelminaam et al.” Arabic Dialects: An Efficient Framework for Arabic Dialects Opinion Mining on Twitter Using Optimized Deep Neural Networks, July 14, 2021, doi:10.1109/ACCESS.2021.3094173.
13. Christoph Tillmann et al. “Improved Sentence-Level Arabic Dialect Classification”, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 110–119, Dublin, Ireland, August 23 2014.
14. Mohamed Elaraby Muhammad Abdul-Mageed “Deep Models for Arabic Dialect Identification on Benchmarked Data”, Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 263–274 Santa Fe, New Mexico, USA, August 20, 2018.
15. Iman S. Alansari “Artificial Intelligence Model to Detect and Classify Arabic Dialects”, journal of Software Engineering and Applications, 2023, 16, 287-300 1.https://www.scirp.org/journal/jse, DOI: 10.4236/jsea.2023.167015.
16. Mohamed Abdeldaiem Abdelhadi “The Effect of Arabic Dialects on Optimization by Information Retrieval Systems based on Arabic Language”, August 12-15, 2015, ISBN: 978-605-4769-90-2.Antalya-Turkry. https://www.researchgate.net/publication/335664999.