#F14-CS

Kingdom of Saudi Arabia
Ministry of Education
Prince Sattam Bin Abdulaziz University
College of Computer
Engineering and sciences
Dep Computer Sciences

المملكة العربية السعودية
وزارة التعليم
جامعة الامير سطام بن عبد العزيز
كلية هندسة وعلوم الحاسب
قسم علوم الحاسب

# SENTIMENT ANALYSIS OF ARABIC TWEETS

**Abdullah Rashed Alqahtani – 439051203**
**Meshary Abdullah Alyam – 439050496**
**Youssef Mohamed Abdelhamid – 439051841**

**To obtain**

## BSc in Computer Science

**Department of Computer Sciences**
**College of Computer Engineering and Sciences**

**Dec & 2021**

**Number: ***/*/****

# SENTIMENT ANALYSIS OF ARABIC TWEETS

| | |
|---|---|
| Abdullah Rashed Alqahtani | – 439051203 |
| Meshary Abdullah Alyami | – 439050496 |
| Youssef Mohamed Abdelhamid | – 439051841 |

**To obtain**

## BSc in Computer Science

**Department of Computer Sciences**
**College of Computer Engineering and Sciences**

**Dec & 2021**

**Declaration Page**صفحة إقرار الملكية

## DECLARATION

We hereby declare that this project report is based on our original work except for citations and quotations which have been duly acknowledged. We also declare that it has not been previously or concurrently submitted for any other degree at PSAU or any other institution.

Date: 4/12/2021

| **Names:** | **ID:** |
|---|---|
| Meshary Alyami | 439050496 |
| Signatures: | |
| Abdullah Alqahtani | 439051203 |
| Signatures: | |
| Yousef Abdelhamid | 439051841 |
| Signatures: | |

## Supervisor Approval Page صفحة موافقة المشرف

### APPROVAL FOR SUBMISSION

I certify that this project report entitled “**SENTIMENT ANALYSIS OF ARABIC TWEETS**” was prepared by **Meshary Alyami, Abdullah Alqahtani and Yousef Abdelhamid** has met the required standard for submission in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.at PSAU.

Approved by

Signature:

Supervisor: Dr. Bandar Almaslukh

Date: 6/12/2021

**Acknowledgements Page صفحة الشكر والعرفان**

## ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to successful completion of this project, in particular Dr. Bandar, for his assistance in many aspects of this project. Thank you to the department for giving them the necessary time to work on the project. We also thank our families for their psychological and moral support to work without tension or pressure. We hope that the project will win everyone's satisfaction.

**Abstract Page(English)** **صفحة المستخلص (إنجليزى)**

## SENTIMENT ANALYSIS OF ARABIC TWEETS

## ABSTRACT

Sentiment analysis is a method for analyzing texts based on their context and meaning. It has gained more interest in many applications, such as tweets, comments classification in social media and products review in marketing. The efficiency of this method increases from time to time with the development of algorithms. Arabic language is one of the world's most famous languages and it is spoken by millions of people, making its use a dominating around the world. Arabic sentiment analysis is also very important in social media and many other text analysis applications. Even though machine and deep learning methods are widely used in several works for classifying texts and sentiment analysis, the number of studies is still limited. Moreover, the complexity of Arabic words in meaning and forms makes a challenge in effectiveness of Arabic sentiment analysis methods. In this project, we develop a methodology for sentiment analysis of Arabic tweets using a different number of machine learning methods such as NB, CNN, and SVM. The methodology consists of a set of steps includes collecting public datasets of Arabic tweets, data cleaning and preprocessing, feature extraction, experimental validation design, ML models training and testing, as well as evaluating the results of ML models.

## Abstract Page(Arabic) صفحة المستخلص (عربى)

### عنوان المشروع

### المستخلص

تحليل المشاعر هي طريقة لتحليل النصوص بناءً على سياقها ومعناها. وقد اكتسب اهتمامًا أكبر في العديد من التطبيقات، مثل تصنيف التغريدات والتعليقات في وسائل التواصل الاجتماعي وتقييم المنتجات في عمليات التسويق. تزداد كفاءة هذه الطريقة من وقت لآخر مع تطوير الخوارزميات. اللغة العربية هي واحدة من أشهر لغات العالم ويتحدث بها الملايين من الناس، مما يجعل استخدامها مهيمنًا في جميع أنحاء العالم. يعد تحليل المشاعر العربية أيضًا مهمًا جدًا في وسائل التواصل الاجتماعي والعديد من تطبيقات تحليل النص الأخرى. على الرغم من استخدام أساليب التعلم الآلي والعميق على نطاق واسع في العديد من الأعمال لتصنيف النصوص وتحليل المشاعر، إلا أن عدد الدراسات لا يزال محدودًا. علاوة على ذلك، يشكل تعقيد الكلمات العربية في المعنى والأشكال تحديًا في فعالية أساليب تحليل المشاعر العربية. في هذا المشروع ، نقوم بتطوير منهجية لتحليل المشاعر للتغريدات العربية باستخدام عدد مختلف من أساليب التعلم الآلي مثل NB و CNN و SVM. تتكون المنهجية من مجموعة من الخطوات تشمل جمع مجموعات البيانات العامة للتغريدات العربية، وتنقية البيانات ومعالجتها مسبقًا، واستخراج الميزات، وتصميم التحقق التجريبي، وتدريب واختبار نماذج التعلم الالي، وكذلك تقييم نتائج نماذج التعلم الالي.

## TABLE OF CONTENTS

**List of Tables Page** صفحة قائمة الجداول

## LIST OF TABLES

**TABLE TITLE PAGE**

List of Figures Page

## LIST OF FIGURES

## List of Abbreviations Page

### LIST OF SYMBOLS/ABBREVIATIONS

| | |
|---|---|
| SA | Sentiment analysis |
| NLP | Neutral language processing |
| ML | Machine learning |
| TF | Term frequency |
| IDF | Inverse Document Frequency |

## List of Abbreviations Page

# LIST OF KEYWORDS

SENTIMENT ANLYSIS, MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, TWITTER DATASET, SUPERVISED LEARNING

# 1. Introduction

## 1.1 Natural Language Processing

The branch of computer science and artificial intelligence that interacts with human language and separates it into parts to analyze and understand the meanings of words. In theory, it deals with techniques that include calculating, analyzing and representing texts to help it understand the nuances of language. The NLB depends on machine learning methods and is used to understand meaning of documents such as twitter tweets. Types of NLP are:

**1- Morphological processing:** is the process of determining the morphemes from which a given word is constructed. It must be able to distinguish between orthographic rules and morphological rules [1].

**2- Syntax analysis:** is the process of analyzing natural language with the rules of a formal grammar. Grammatical rules are applied to categories and groups of words, not individual words [2].

**3- Semantic analysis:** the way we understand what someone has said is an unconscious process relying on our intuition and knowledge about language [3].

**4- Pragmatic analysis:** is the process of interpreting the results of semantic analysis.

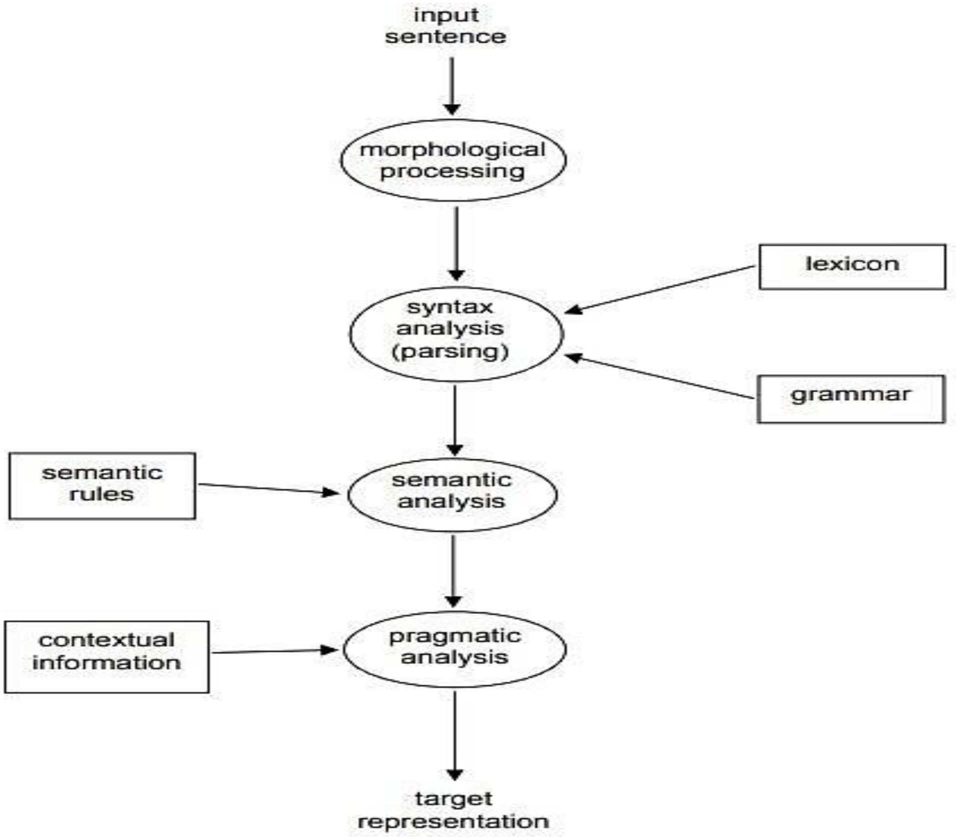


Fig. 1.1: Steps of natural language processing [4].

### 1.2 Sentiment Analysis

The prospecting method used to extract information about people's feelings and opinions about things. SA it is considered one of the important things through which many things are known, such as feedback, and this helps in many works to know the nature of people about anything. For example, comments about a specific product on Twitter or any other site that helps companies improve the product and get an idea of what people think about it. With the tremendous development in social media in terms of the number of people such as Twitter, it is not easy to extract an opinion because of the large number of tweets, so a clear methodology must be found and the machine will help analyze this data accurately.

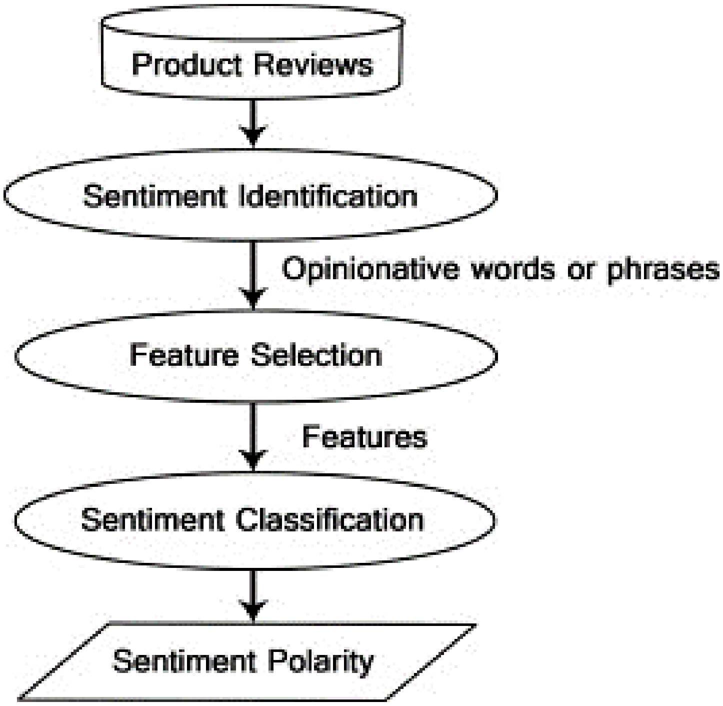


Fig. 1.2: Classification of Sentiment Analysis [5].

### 1.3 Problem Statement

Through a group of tweets containing various opinions and feelings, there is a lot of methods and approaches to classify them into positive or negative. However, it is difficult to express the Arabic words with multiple meanings. A word may have different meanings in different contexts. Moreover, it is difficult to express different grammatical or semantic information. Therefore, the problem statement is with the complexity of Arabic tweets to be classified in present the above limitations.

## 1.4 Motivation

Sentiment analysis has become a must, especially with the huge demand for sites and social media, for example, twitter has hundreds of millions of users and thousands of tweets are posted daily. It is important to know their feelings about something to build on them in several things such as reactions about a restaurant or hotels or products. The user is now able to easily search for something specific that they need, but with so many tweets, some of which have nothing to do with the topic, the topic becomes more difficult to get an adequate response so it would be nice to categorize them.

## 1.5 Objective

In this project, we focus on classifying Arabic tweets into positive or negative using supervised machine learning models. These model use the labeled data to learn how to classify the tweets into positive or negative sentiments.

## 1.6 Document Organization

- **Chapter 1**: In this chapter, we define natural language processing, sentiment analysis, and its terminology. In addition, the motivation, problem statement and objective of this project are given.
- **Chapter 2:** Arabic language challenge, data preprocessing and some algorithms used in sentiment analysis are discussed in this chapter. (background)
- **Chapter 3:** In this chapter we will display the data set and go deeper in algorithms and evaluation**.**
- **Chapter 4:** This chapter introduces the proposed methodology with its main steps.
- **Chapter 5:** This chapter concludes the work done in project 1.
- **Chapter 6:** This chapter will contain the implementation and experimental results and findings in future in project 2.

## 2. Background

Arabic language is popular language and has many challenges, many other languages precede it in sentiment analysis, we will get Sentiment analysis concept and how we prepare data to analysis, and we will get old methods to sentiment analysis and comer it.

### 2.1 Arabic Language Challenges

Arabic language is important for Muslims because there is 1.9 billion Muslims in the world [6], those Muslims have to learn Arabic because it's the language of Quran the holy bible for Islamic religion, so that’s why the Arabic language is important, there is also 466 million Arabic people in the world [7], they speak Arabic as their native language. the Arabic language is the 6th language spoken and also being the hardest in learning [8]. What make Arabic hard is it’s grammar, it’s used to identify the subject and the object [9] Who did that, for example ‘‘ضرب زيدُ عمر ’’ Zaid beat Omar, how do we know who got beaten? We know that by formation in Arabic “ ”التشكيل بـ الحركات ,We form it by formatting signs “الضم, الفتح ,الكسر” the formatting sign is not used just for gramma, It’s also used to give the word other meanings like “رجل” that word can mean feet and can mean man it can have different meanings [10], But we can know which meaning we want by formation, These are real problems caused formation and is removed when we preprocess the data, There is too many word that have many meanings, We can identify meanings by context but not as accurate as formation, So actually we just have context to identify the meaning, and this is the strongest factor that we keep in mind when we come to choose an algorithm, other unique factor of Arabic language when we have a two or more objects in Arabic we deal with it by adding two letters depending on our object[11], We don’t add many numbers of written objects etc... and just write it like in English, Disjunctive and Continuous [12] Characters every character in Arabic has two or more shapes like “ب” it’s like B letter in English, if it comes in the middle of a word or sentence it’s written “ ”ـبـ,You connect it to other letters to make a word like “باب” which means door in English, That word has two different shapes of the letter “ب” some words has more than two shapes, Like “ه” it also comes “ ـه, ـهـ , ه ” that is another reason that made Arabic unique, The most hard challenge in Arabic is the dialect, all Arabic people spoke classical Arabic in the past, with the time the people evolved dialects, they are over 50 dialects around the Arabic world, for example, in KSA Najde, Hejaze, Janoby, Qasime, etc. In every single dialect there is new vocabulary, dialects are around all the Arabic world, North Africa, Egypt, Syria, Iraq, etc. Every region has more than one dialect, Some vocabulary meanings are changed in author dialect, Every dialect has different rules and it may seem similar but it’s different, The dialect has evolved, for example Saudi dialect in 50s are not like now, if you ask random modern Saudis about “حشف" Most of them won't know what it means, it's means fruit kernel in 50s , With the time the dialect has changed, this is another challenge in Arabic you must be up to date with different types of dialects.

## 2.2 Sentiment Analysis Problem Definition

Sentiment analysis is very important because we need to extract the writer’s feelings, and by analogy with the complexities and challenges in the Arabic language, we will face a little difficulty to know the purpose of the written word or sentence. The problem is solved by a set of steps that will explain in next sections.

## 2.3 Data Preprocessing

The preprocessing task includes tokenization, punctuation removal, stop word removal, rooting, and document vector construction. Tokenization is a crucial process for dividing a text into meaningful units called tokens. For every token received, we apply a normalization task; then we consider whether or not to keep the Arabic language marks, remove the stop words or not, and stem the word or not. More specifically, this task includes:

1- Remove punctuation signs like - ؟ ، .

2- Remove faces and emoji out of the sentences.

3- Remove stopping words like (في، عن، على، إلى).

4- Stemming words to its roots to make it easier for the machine to know the meaning of similar words (أحمد، مدرسة).

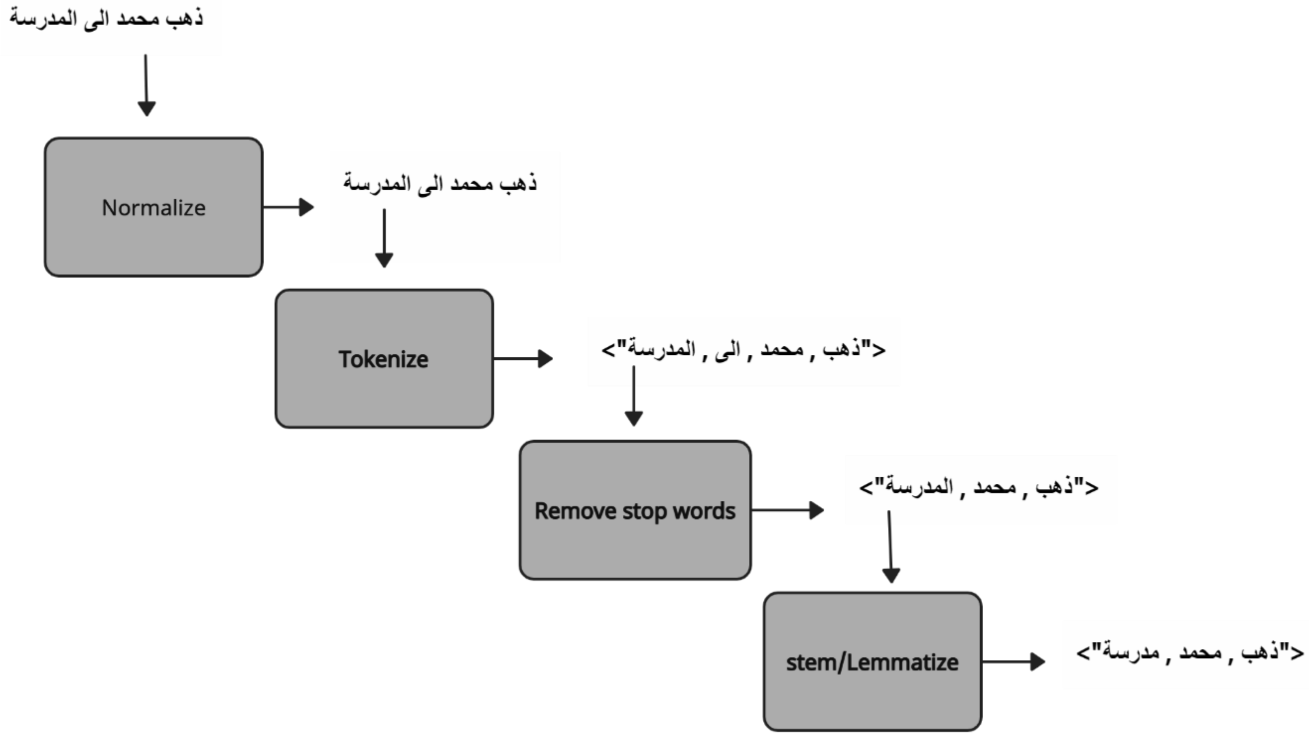


Fig. 2.3: Data preprocessing steps

## 2.4 Feature Extraction

Feature extraction is used to reduce the number of features in a dataset and to build meaningful features for the computer from raw data and columns. Different types of feature extraction models including:

### 2.4.1 Classical Models

### A. Bag of words

Is a method to extract useful features from the data by putting unique words of a set of sentences or paragraphs into an array associated with each word occurrence.

Example of paragraphs in Arabic language:

موقع تويتر هو الموقع الأكثر شهرة في المملكة العربية السعودية

تويتر هو أفضل مواقع التواصل الاجتماعي

هناك الكثير من مواقع التواصل الاجتماعي مثل فيسبوك وتوتير وانستغرام

Example of constructing bag of words in Arabic language from the above paragraphs:

Put all of the unique occurrences of a word in an array like that:

موقع, “"الموقع, “"الأكثر, “"شهرة, “"في, “"المملكة, “"العربية, “"السعودية, “"تويتر, “"هو, “"مواقع, “"التواصل, ""“]

]“"الاجتماعي, “"هناك, “"الكثير, “"من, “"مثل, “"فيس, “"بوك, “"وتوتير, “"وانستغرام , “"أفضل.

To make a paragraph vector associate each word with its number of occurrences in each paragraph:

”موقع تويتر هو الموقع الأكثر شهرة في المملكة العربية السعودية“

= [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0]

”تويتر هو أفضل مواقع التواصل الاجتماعي“

= [0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,1]

”هناك الكثير من مواقع التواصل الاجتماعي مثل فيسبوك وتوتير وانستغرام“

= [0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0]

Bag of words is used to extract important characteristics of data like word frequencies.

**B. TF-IDF**

TF-IDF is an abbreviation for term-Frequency—inverse Document Frequency

A statistic that is used to make a vector of each word in a document to be interpreted by a programming language like R or python.

TF-IDF is used to prepare the document for multiple jobs like clustering or identifying similarities.

TF-IDF consists of:

- **Term Frequency (TF)**

It’s defined as the number of occurrences of a word or term in a document divided by the number of words in the documents.

Usually, words repeated in a document are more important than the others that’s why TF-IDF takes the term frequency into the equation

The count of every word is divided by the number of words in the document to normalize the results in a range to compare to other documents.

So, **term frequency = TF (t, d) = count (t)/number of words (d)**;

- **Document Frequency DF**

It is the number of documents that a specific term or a word appears in them. It is normalized by dividing the number of occurrences by the number of documents that contains the word.

- **Inverse Document Frequency IDF**

A measurement for the in formativeness of a specific word in a document

-getting the inverse of document frequency helps to disqualify stop words by giving them a very low value. So now IDF = N/DF where:

N=number of occurrences

DF=number of documents which N appeared in

When we have many documents, it's better to get the log of N/DF to prevent the value from increasing by a lot unexpectedly.

So IDF=Log (N/DF)

If the term is not found in any document, we add 1 to DF to prevent division by zero.

So IDF=Log (N/DF+1)

So, the final equation for TF-IDF would be: TF-IDF = TF * IDF = TF (t, d) *Log (N/DF+1)

### 2.4.2 Representation Learning

#### A. Continuous Words Representation (Non-Contextual Embedding)

There are two famous model for Continuous Words Representation: **Word2Vec** and **Global vectors (Glove).**

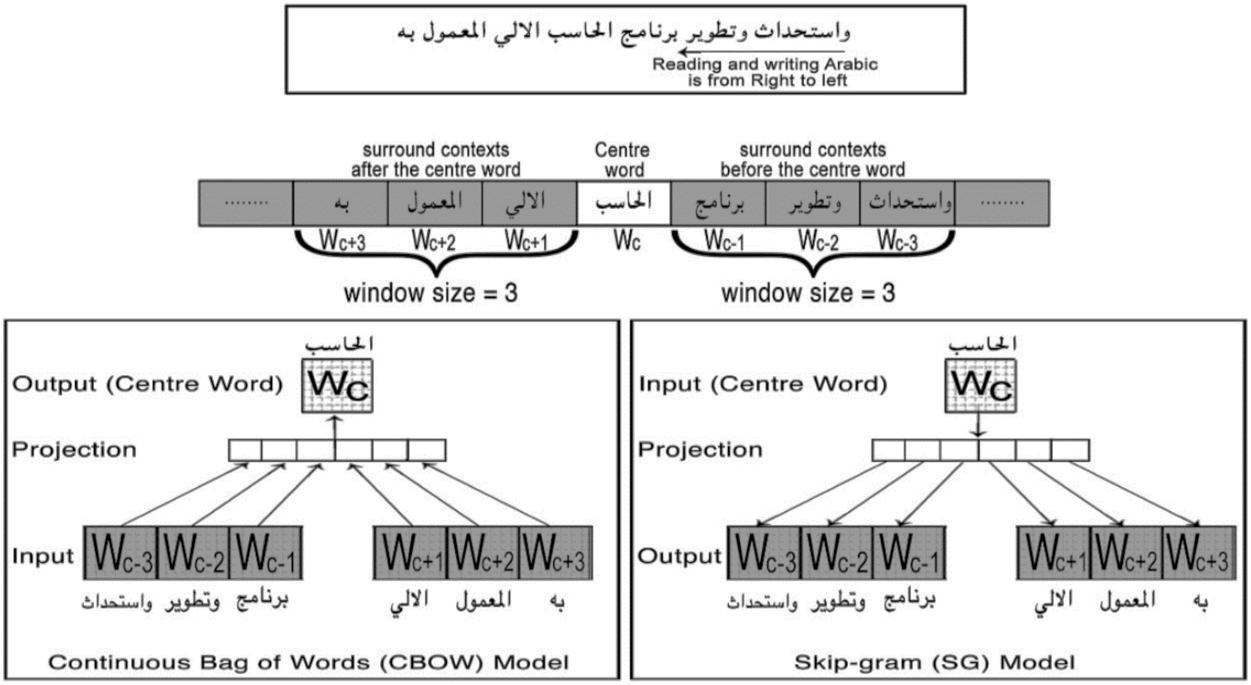
**Word2Vec:** The Word2Vec is an NLP type of models that is used to convert texts into numerical vectors to make it readable for the computers and deep neural networks it reduces the number of dimensions of the vector space to recognize the similarities of words and make very good estimates about the meaning of a word based on past data and occurrences. It is also a useful technique for data internet search preparation. As shown in Fig. 2.4, we can make Word2Vec model using two different architectures:

**1. Continuous Bag of Words (CBOW) Model**

The CBOW is an abbreviation for continuous bag of words. This model is the simplest one from word to Vector models. It is trained by trying to predict a target word giving a specific context

**2. Skip-gram (SG) Model**

It is trained in an opposite manner of CBOW model. The skip-gram model is given a specific word and tries to predict its context. For instance, in the word: windy is usually associated with the sentence (the weather is). A two-layer neural network tries to infer the context (The weather is) out of a given word this model is more accurate that the CBOW model.



(a) (b)

Fig. 2.4: The Word2vec representation model: (a) Continuous Bag of Words (CBOW), and (b) Skip-gram (SG) [15].

**Global vectors (Glove):** It is an **NLP** advanced technique used to convert texts into numerical vectors. Unlike word2vec, which learns the meaning of a word, using a local surrounding context, Glove takes into consideration all of the occurrences of a certain word in the entire corpus so it depends more on the count of the word by building an occurrence matrix.

### B. Contextual Word Representations

In contextual representations, words have various representations based on their contexts, thus capturing the use of words through diverse contexts and encoding the knowledge that can transfer by the languages. If a word has different meanings in two sentences, then we will get two different contextual representations of the word. But, more importantly, we found that the pre-trained context representation performed quite well in most downstream tasks. This context-based word representation has played a cornerstone role in the current NLP system. By using the contextual word representations, large text corpus datasets were used to learn machine learning models that embed the contexts from sentences and target words of the same low dimensions. Then, they can be optimized in order to reflect the interdependencies between their entire sentential context and targets as a whole [16].

### 2.5 Model Training and Evaluation

In this step the data set is divided into two sets that are training set and testing set. The model is trained using training data and evaluated using testing data and it can be divided as follows 80% training data, 20% testing data. The ML approach is based on two main phases training and testing as shown in Fig. 2.5. In the training phase, machine learning methods take training data to train the model, and in testing phase system predict result  of the testing data.

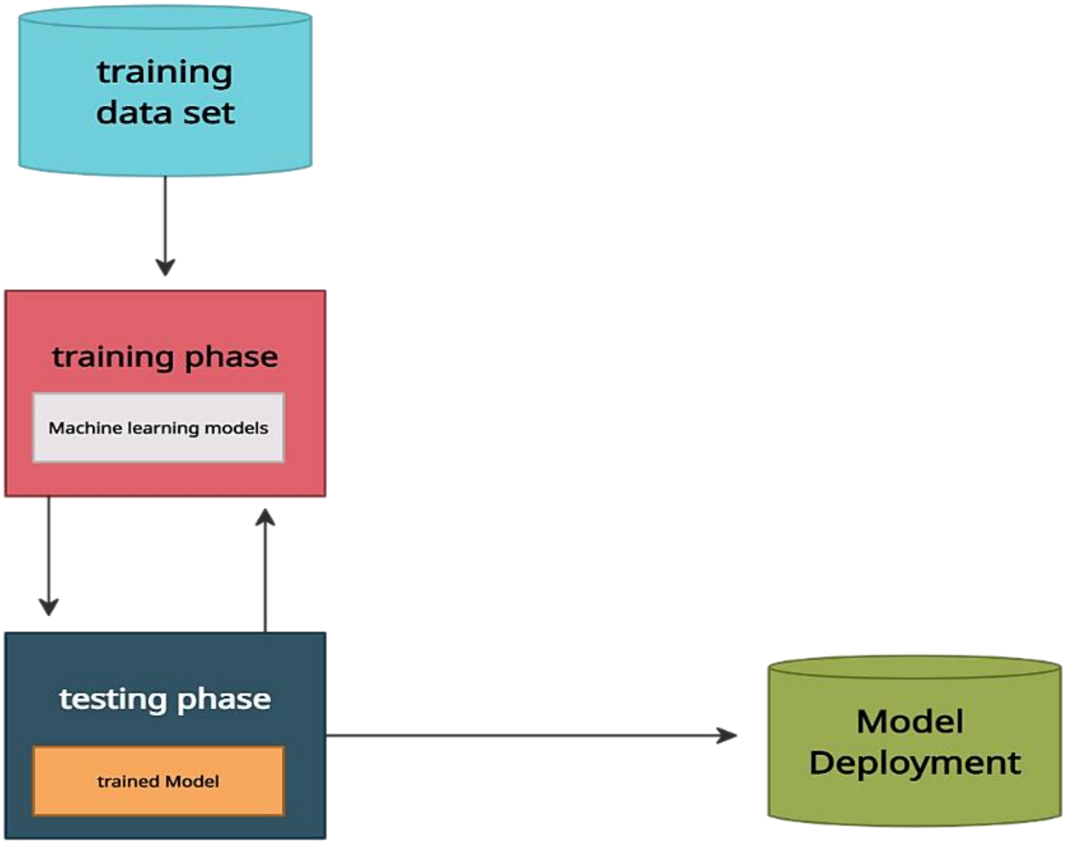


Fig 2.5: Process of Training and Testing.

# 3. Literature Review

## 3.1 Introduction

Sentiment analysis means knowing a person or a group of people emotions through a set of given words. Studies about Simental analysis has increased in the last decade due to the wide availability of World Wide Web and large numbers of internet users, studies have showed that around 88.5% [13] of people in Saudi Arabia is using the internet, at the same time 98% of young people in the kingdom use the internet. The increasing use of social media such as twitter in Saudi Arabia prompted a lot of researchers to perform scientific analyzes of user's behavior in the country through social media. Twitter is one of the most used social networking sites in Saudi Arabia, with 25 million users on the platform. Twitter is not only used in Saudi Arabia, but in the whole world as a major platform to share thoughts in a specific character limit. This part of the research focuses on previous scientific research that was carried out on the Twitter platform to analyze Arabic tweets. Some of the research was done in Saudi Arabia, Egypt and the UAE. Most of the researchers collected data and filtered it manually from impurities and some researchers took advantage of the spread of the corona epidemic and distance education to perform large statistical analyzes on user behavior by using large numbers of tweets (corpus), some of which consisted of several thousand tweets and some of which contained millions of tweets. Researchers have used different classification methods for tweets. Some researchers rated tweets as positive or negative only and filtered the rest, and some researchers added other ratings such as neutral, mixed between the two, or even objective. The researchers also used different methods to analyze the data. Some researchers used Naive Bayes and others used Support vector machine, Random Forest and K-nearest neighbor Classifiers. In the next section a table representing previous research that has been done on Arabic tweets with some related scientific papers.

## 3.2 Arabic Tweets Datasets

Table 3.1 presents the Arabic tweets datasets reviewed throughout this project. We write down in the table the following information: the dataset name, the class name, the number of tweets, the collection date, reference of the dataset, and its availability.

Table 3.1. Arabic tweets datasets.

| Data set name | Class | # of tweets | Collection date | References | Availability |
|---|---|---|---|---|---|
| 1- Arabic Sentiment Twitter Corpus | -Positive<br>-Negative | 47,000 | Collect data from twitter API in 2021 | Kaggle | Public under CC License |
| 2- Twitter Benchmark Dataset for Arabic Sentiment Analysis | -Positive<br>-Negative | 151,548 | Collect data from tweepy in 2018 | https://2u.pw/i12G5 | Data not available |
| 3- AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets | -positive<br>-negative<br>-neutral<br>-mixed | 6,341,135 | Collect data from twitter API in January 2016 | https://cutt.us/SnLVJ | Data not available |
| 4- Arabic Tweets Sentiment Analysis about Online Learning during COVID-19 in Saudi Arabia | -positive<br>-negative<br>-neutral | 10,000 | Collect data from twitter API in March 2020 | https://cutt.us/zWR0J | Data not available |
| 5- ASTD: Arabic Sentiment Tweets Dataset | -positive<br>-negative<br>-mixed<br>-Objective | 10,000 | Collect data from twitter API in Aug 7, 2015 | https://cutt.us/SXE2v | Public under GPL 2.0 License |
| 6- Arabic Sentiment Analysis: Corpus-based and Lexicon-based | -positive<br>-negative | 2000 | Collect data from twitter API in 2014-04-11 | IEEE | Public under CC-By SA |
| 7- Arabic-Sentiment-Analysis | -positive<br>-negative<br>-neural | 16,702 | Collect data from twitter API in Mar 17, 2018 | GitHub | Public ,the published didn't mention the license |

### 3.2.1 Dataset 1

The purpose of this study is to develop a reliable system for classifying Arabic-based tweets on twitter. The analysts employed assessment methods to assess the accuracy, precision, recall, and F1 measure of their model. They used the dataset to train four different algorithms: Naive Bayes Algorithm Ridge Classifier algorithm, passive aggressive classifier algorithm and logistic regression algorithm.

### 3.2.2 Dataset 2

The research outlines how they built the twitter benchmark dataset, which includes a variety of Arabic dialects, as well as the tools and algorithms they utilized. The first section discusses the methods involved in collecting and preparing Arabic tweets, while the second section offers a list of machine learning algorithms that were used to analyze the dataset.

### 3.2.3 Dataset 3

It includes information on how they built the dataset as well as how they cleaned it in preparation for the study, because around 60% of Arabic tweets are written in Saudi dialect, the data set solely covers Saudi Arabic tweets. The tweets were then processed by MADAMIRA for normalization and tokenization, followed by a benchmark experiment in which they identified a variety of issues with tweet annotation task for the Arabic language.

### 3.2.4 Dataset 4

The purpose of the paper was to gain a better understanding of tweeters' thoughts and opinions concerning online learning. They did so by collecting over 10,000 tweets and then used the Text Blob python module to assess emotive traits such as subjectivity and polarity. According to their findings, the majority of tweets represent a neural viewpoint. they also mentioned some difficulties they have had with Arabic language analysis, such as a lack of resources, limited availability of MSA lexicons etc...

### 3.2.5 Dataset 5

Working on the dataset is introduces ASTD, they collect the data from twitter, they collect more 10000 to classified dataset as objective positive, negative, and mixed, they run them experiment by using standard partitioning of the dataset, they proved benchmark by 4-way sentiment classification on the data.

### 3.2.6 Dataset 6

In the paper of this data set they study tow approached of (SA), corpus-based and lexicon-based, they want to apply those approached in Arabic language, they build own dataset and describe steps of building, they goal is increase the accuracy of the system and compare them to corpus-based approach.

### 3.2.7 Dataset 7

In that dataset the polisher builds sentiment analysis model for Arabic tweets and classified it for positive, negative and neutral, he apply three ML algorithms to build the model for SA, he used Gaussian Naive Bayes, multinomial Naive Bayes and Support vector machines, by split dataset for three different datasets for each algorithm then compare between the algorithms, that how he tries to get best model.

### 3.3 The Performance of the Related Studies

Table 3.2 shows the performance of the related sentiment analysis studies. It describes the dataset number, the ML model used, accuracy, precision, recall, and F-measure.

Table 3.2. The performance of the related studies**.**

| Dataset # | ML model used | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| **1** | Naïve-Bayes | 69% | 66% | 60% | 66% |
| | Ridge-Classifier | 71% | 70% | 66% | 70% |
| | Passive-Classifier | 65% | 58% | 47% | 58% |
| | Logistic-Regression | 77% | 76% | 73% | 76% |
| **2** | Naïve-Bayes | 96% | 96% | 96% | 96% |
| | Adaptive Boosting | 73% | 77% | 77% | 71% |
| | Support Vector Machines | 99% | 99% | 99% | 99% |
| | Maximum Entropy | 94% | 94% | 94% | 94% |
| **3** | SVM with linear kernel (Two-way Classification) | - | - | - | 62% |
| | SVM with linear kernel (Three-way Classification) | | | | 58% |
| | SVM with linear kernel (Four-way Classification) | | | | 54% |
| **4** | C-Neural Network | 64% | - | - | - |
| | Support Vector Machines | 84% | | | |
| **5** | Support Vector Machines | 68.7% | - | - | 62.0% |
| | K-nearest-neighbor | 66.57% | | | 55.8% |
| | Logistic-Regression | 68.07% | | | 57.6% |
| **6** | Naïve-Bayes | 80% | - | - | - |
| | Support Vector Machines | 84% | | | |
| | K-nearest-neighbor | 51% | | | |
| **7** | Support Vector Machines | 51% | - | - | - |

Each data set used more than one algorithm and different measure to evaluate the quality and accuracy of the proposed method as shown in the previous table. The accuracy is most popular measure but it has some limitation when there is skewed classes. The F-Measure is considered one of the most accurate measurements because it combines recall and precision to gives the best possible result. However the calculation of these measurements  according to the confusion matrixes (Fig 3.3) as follows:

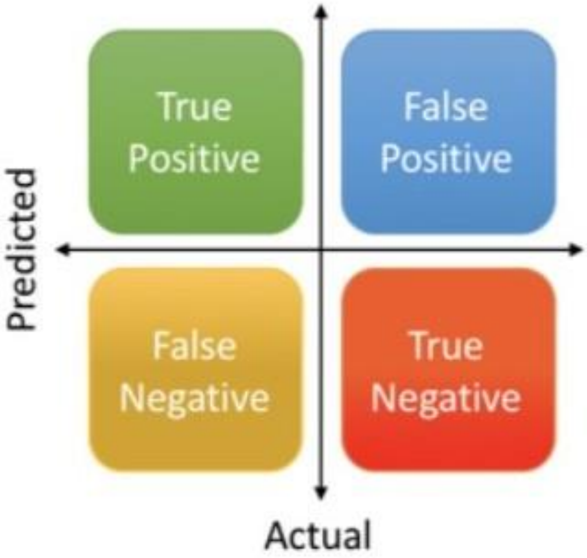


Fig. 3.3: example of measures [14].

- Precision = True Positives / (True Positives + False Positives)
- Recall = True Positives / (True Positives + False Negatives)
- F-Measure = (2 * Precision * Recall) / (Precision + Recall)
- Accuracy = True positive + True negative / Total

# 4. Proposed Methodology

## 4.1 Introduction

In previous chapter, a literature review of methods and datasets used for sentiment analysis is given. As well, the performance of the related studies is presented and summarized in the tables. This chapter introduces the workflow of the proposed methodology to overcome the drawbacks of Arabic sentiment analysis task. This proposed methodology exercised a number of Arabic words representations with machine learning methods to reach the final desired goal. The description of methodology workflow with its steps is explained next subsection.

## 4.2 Workflow of Proposed Methodology

The workflow of proposed methodology is given below in Fig. 4.1. It consists of a number of stages, which are collecting public datasets of Arabic tweets, data cleaning and preprocessing, feature extraction, experimental validation design, ML models training and testing, as well as evaluating the results of ML models.

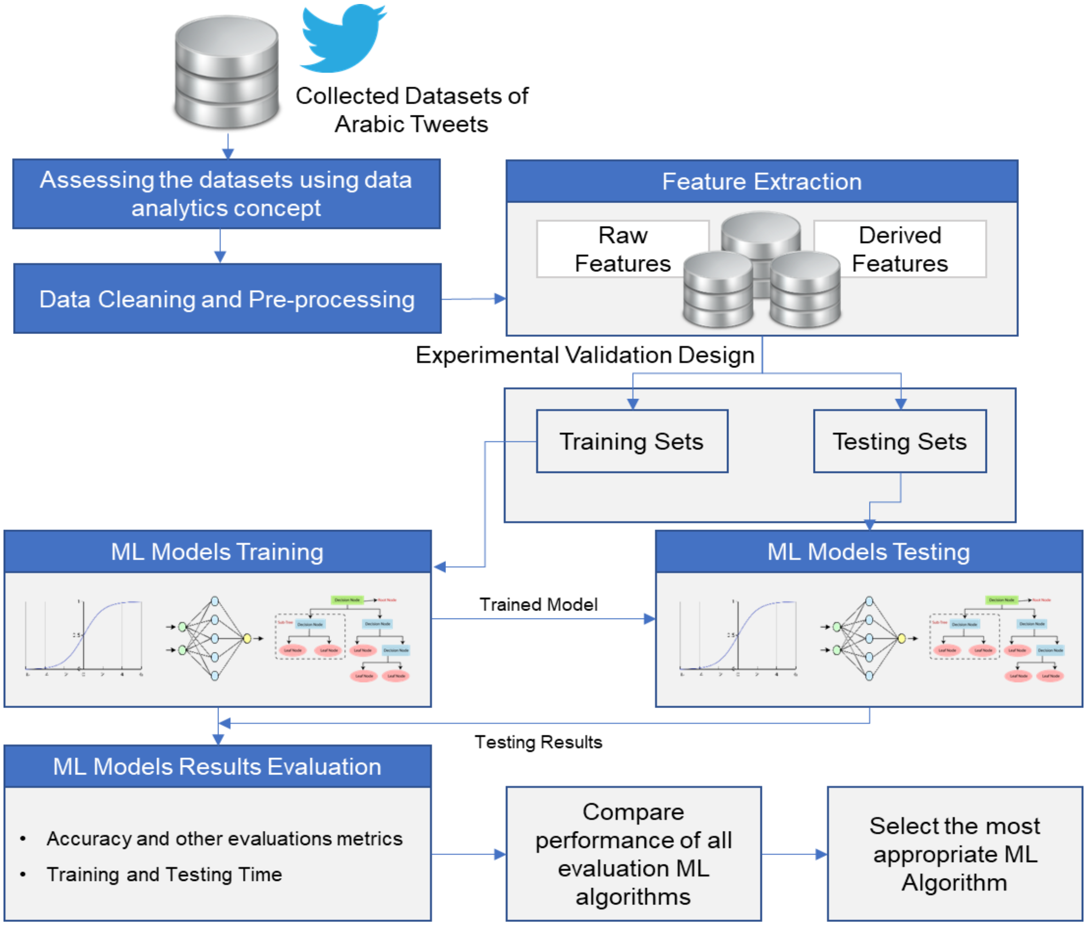

Collected Datasets of Arabic Tweets
Assessing the datasets using data analytics concept
Data Cleaning and Pre-processing
Feature Extraction
Raw Features
Derived Features
Experimental Validation Design
Training Sets
Testing Sets
ML Models Training
ML Models Testing
Trained Model
Testing Results
ML Models Results Evaluation
Accuracy and other evaluations metrics
Training and Testing Time
Compare performance of all evaluation ML algorithms
Select the most appropriate ML Algorithm

Fig. 4.1: The workflow of proposed methodology.

The description of these steps are given as follows:

1. Collecting public datasets of Arabic tweets: In this step, public Twitter datasets will be downloaded and collected from the Internet sources. These datasets were collected using Twitter APIs and based on chosen keywords to the domain of the collection concern, such as product or service reviews.

2. Assessing the datasets using data analytics concept: In this step, the collected datasets will be evaluated in terms of the nature of the tweets, their appropriateness for training, the noise texts, and the number of labels. Data analysis or analytics is a science for analyzing the raw data to draw conclusion about its information. Many of data analysis processes and techniques have been automated into algorithms that can process the raw data for human consumptions.

3. Cleaning and preprocessing the tweets in the datasets: In this step, the datasets are put over a cleaning and preprocessing processes in which we remove identifying information such as timestamps, twitter handles of the embedded links, message and videos. Such irrelevant

information is mainly may cause incorrect results given by the tool or system. Tweets preprocessing is for preparing the text for analyzing and classification. The preprocessing task includes tokenization, punctuation removal, stop word removal, rooting, and document vector construction. Tokenization is a crucial process for dividing a text into meaningful units called tokens. For every token received, we apply a normalization task; then we consider whether or not to keep the Arabic language marks, remove the stop words or not, and stem the word or not.

4. Extracting features from the tweets of selected datasets: Feature extraction of tweets is the process of mapping words onto real-valued vectors. Various techniques can be used for feature extraction, such as a bag of words and TF-IDF. Because the limitations of preprocessing, Arabic NLP tasks such as classifying the sentiment from tweets is difficult to perform. There are different transformers for this purpose, which have been developed in recent years. The language-specific BERT-based approaches are one of them that shown greater effectiveness and accuracy because they are trained on a very large corpus. The extracted features of the datasets will be through the experimental validation design for splitting to training and testing sets.

5. ML models training and classification: After successful feature extraction, we select a machine learning model for training. There are different models of machine learning such as supervised learning models and unsupervised learning models. Models are selected according to the data set. There is no model that we can say is the best model. Different models behave in opposite ways on different datasets. The data set is divided into two sets, namely the training dataset and the test dataset. The selected model is trained on the training data sets and evaluated on the test datasets. Finally, the trained machine learning models then can be used for classifying the tweets of the test datasets. The results of tweets classification by ML models can be evaluated and compared with the results of other models in other works in terms of the accuracy and other evaluation metrics.

## 5. Conclusion

In this project report, we introduce four chapters. Chapter 1 gives an introduction on natural language processing (NLP) field, including its techniques of computation, analyzing and representing texts. The NLB depends on machine learning methods and it is used to understand meaning of documents such as twitter tweets. The types of NLP can be morphological processing: is the process of determining the morphemes from which a given word is constructed. It must be able to distinguish between orthographic rules and morphological rule, syntax analysis, semantic analysis, and pragmatic analysis. It also explains the problem statement, the motivation, and the project objectives for developing a sentiment analysis on Arabic tweets. Then, chapter 2 introduces a background on Arabic language challenges, sentiment analysis problem definition, data preprocessing, text feature extraction, and model training and evaluation. After that, in chapter 3, we give a literature review on Arabic tweets datasets and the performance of the related studies. In addition, in chapter 4, we give an introduction about the proposed methodology, we draw the workflow of the methodology, and we explain the methodology steps. Next, in chapter 5, we conclude the work done in project 1 throughout the chapters of the report. In future, in project 2, chapter 6, we will give the implementation and experimental results of the ML methods used in the proposed methodology and we will give our conclusions and findings.

## REFERENCES

[1] https://www.gabormelli.com/RKB/Morphological_Parsing_Task

11/30/2021

[2] https://builtin.com/data-science/introduction-nlp

11/30/2021

[3] https://builtin.com/data-science/introduction-nlp

11/30/2021

[4]https://www.researchgate.net/figure/Natural-Language-Processing-NLP-steps_fig2_331345674

11/30/2021

[5]https://www.researchgate.net/figure/Sentiment-analysis-process-on-product-reviews_fig3_261875740

11/30/2021

[6] https://worldpopulationreview.com/country-rankings/muslim-majority-countries

11/30/2021

[7] https://www.tarjama.com/how-many-countries-that-speak-arabic-around-the-world/

11/30/2021

[8] https://www.berlitz.com/en-uy/blog/most-spoken-languages-world

11/30/2021

[9]https://sotor.com/%D8%A7%D9%84%D9%81%D8%A7%D8%B9%D9%84_%D9%88%D8%A7%D9%84%D9%85%D9%81%D8%B9%D9%88%D9%84_%D8%A8%D9%87_%D9%81%D9%8A_%D8%A7%D9%84%D9%86%D8%AD%D9%88

11/30/2021

[10]https://www.almaany.com/ar/dict/ar-ar/%D8%B1%D8%AC%D9%84/?c=%D8%A7%D9%84%D9%85%D8%B9%D8%AC%D9%85%20%D8%A7%D9%84%D9%88%D8%B3%D9%8A%D8%B7

11/30/2021

[11]https://www.alukah.net/literature_language/0/122712/

11/30/2021

[12]https://grammar.yourdictionary.com/grammar-rules-and-tips/rules-for-writing-numbers.html

11/30/2021

[13]https://al3arabi.com/%D8%AA%D9%83%D9%86%D9%88%D9%84%D9%88%D8%AC%D9%8A%D8%A7/%D8%A7%D8%B1%D8%AA%D9%81%D8%A7%D8%B9-%D8%B9%D8%AF%D8%AF-%D9%85%D8%B3%D8%AA%D8%AE%D8%AF%D9%85%D9%8A-%D8%A7%D9%84%D8%A7%D9%86%D8%AA%D8%B1%D9%86%D8%AA-%D9%88%D8%A7%D9%86%D8%AE%D9%81%D8%A7%D8%B6-%D8%B9
11/30/2021

[14] https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488
11/30/2021.

[15] Alayba, Abdulaziz M., Vasile Palade, Matthew England, and Rahat Iqbal. "Improving sentiment analysis in Arabic using word representation." In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp. 13-18. IEEE, 2018.

[16] Naseem, Usman, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models." *Transactions on Asian and Low-Resource Language Information Processing* 20, no. 5 (2021): 1-35.