****

Sudan University of Science and Technology

College of Computer Science and Information Technology

**Build a Sentiment Analysis Model for the Sudanese Dialect for the Internet Service**

**(The Case Study: Sudani Telecom Company)**

**بناء نموذج لتحليل المشاعر للهجة العامية السودانية عن خدمة الإنترنت (دراسة حالة: شركة سوداني للاتصالات)**

A scientific project submitted as one of the requirements for obtaining a Bachelor’s degree in Software Engineering and Information Technology

**PROPOSED BY:**

* **Athar Abdalgauom Fadlallah**
* **Rabab Khalifa Mohammed**
* **Rayan Jafar Abdullah**

**SUPERVISOR: Mr. Muhammad Al-fatih Osman**

March 2023

Sudan University of Science and Technology

College of Computer Science and Information Technology

Department of Information Technology and Software Engineering

**Build a Sentiment Analysis Model for the Sudanese Dialect for the Internet Service**

**(The Case Study: Sudani Telecom Company)**

**بناء نموذج لتحليل المشاعر للهجة العامية السودانية عن خدمة الإنترنت (دراسة حالة: شركة سوداني للإتصالات)**

A scientific project submitted as one of the requirements for obtaining a Bachelor’s degree in Software Engineering and Information Technology

March 2023

****

# **الايـــــــــــــــــــــة**

**قال عز من قائل:**

**{ ‏‏يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ}**

**…. صدق الله العظيم**

**(المجادلة‏:‏ 11‏)**

# **DEDICATION**

**الى من كلله الله بالهيبة والوقار ... الى من علمني العطاء بدون انتظار.. الى من أحمل اسمه بكل افتخار ارجو من الله ان يرحمك وان يتقبلك مع الصديقين و الشهداء وستبقى كلماتك نجوم أهتدي بها اليوم وفي الغد والى الأبد**

**...والدي العزيز رحمه الله(ريان جعفر)**

**الى من تعبو لاجلنا ولم يبخلو بأي شيء، الى قدوتنا وسندنا الى من نحمل اسمائهم بكل فخر**

**..... آبائنا الاعزاء**

**الى من ساندننا بدعائهن ولم يدخرن جهداً في سبيل نجاحنا، الى نبع الحنان وبسمة الحياة الغاليات أطال الله في عمركن وجزاكن الله خير الجزاء**

**...... أمهاتنا الحبيبات**

**الى هدية الحياة وومضات الفرح، الى من اثرونا على أنفسهم وساندونا طيلة حياتنا ومشوارنا الدراسي**

**.... إخوتنا الاعزاء**

**الى من قضينا معهم أحلى وأصعب الايام وتشاركنا المقاعد والحلم**

**... الى الاصدقاء الاعزاء.**

**الى من حمل أقدس رسالة،الى الذين مهدٌو لنا طريق العلم والمعرفة**

**...الى الدكاترة والاساتذة الكرام**

**إلى من أمدنا بالعلم، والمعرفة، وأسدى لنا النصح والتوجيه**

**مشرف البحث الاستاذ محمد الفاتح عثمان**

**نهدي لكم هذا العمل المتواضع**

# **ABSTRACT**

There is an urgent need to adopt the field of automated processing of natural languages in order to organize and process what is circulated on social networking sites quickly and with minimal effort and cost.

Sentiment analysis can be expressed by extracting valuable patterns from textual data. These functional patterns involve interpreting and categorizing feelings into neutral, positive, or negative from that data using specific analysis techniques.

This research discusses the construction of a model for analyzing the feelings of the Sudanese colloquial dialect on the Internet by taking the Sudanese Telecom Company as a case study.

Where the researchers manually collected more than 6,000 comments from the official Facebook page of the Sudanese Telecom Company, representing the opinions of the company's customers about the services. Classification was done on the data by applying ten different classification algorithms and the model was evaluated using the four most common metrics (retrieval, precision, F1 score, and precision).

The results showed that it is better to perform binary classification on the data after performing all pre-processing operations except (lemmatization), as both SVM and MLPC models achieved the best accuracy when classifying into two classifiers (positive, negative) accurately. with a score of 93%, while the MLPClassifier model achieved the best accuracy when classifying into two classifiers (neutral, non-neutral) with a score of 92%.

# **المستلخص**

**هناك حاجة ملحة لاعتماد مجال المعالجة الآلية للغات الطبيعية من أجل تنظيم ومعالجة ما يتم تداوله على مواقع التواصل الاجتماعي بسرعة وبأقل جهد وتكلفة.**

**يمكن التعبير عن تحليل المشاعر عن طريق استخراج أنماط قيمة من البيانات النصية تتضمن هذه الأنماط الوظيفية تفسير المشاعر وتصنيفها إلى محايدة أو إيجابية أو سلبية باستخدام تقنيات تحليل محددة.**

**يناقش هذا البحث بناء نموذج لتحليل مشاعر اللهجة العامية السودانية على الإنترنت بأخذ شركة الاتصالات السودانية كدراسة حالة.**

**حيث جمع الباحثون يدويًا أكثر من 6000 تعليق من صفحة الفيسبوك الرسمية لشركة الاتصالات السودانية ، تمثل آراء عملاء الشركة حول الخدمات. تم التصنيف على البيانات من خلال تطبيق عشرة خوارزميات تصنيف مختلفة وتم تقييم النموذج باستخدام المقاييس الأربعة الأكثر شيوعًا وهي (Recall, Accuracy, F1-score, Precision) أظهرت النتائج أنه من الأفضل إجراء التصنيف الثنائي على البيانات بعد إجراء جميع عمليات المعالجة المسبقة باستثناء (lemmatization) ، حيث حقق كل من نموذجي SVM و MLPClassifier أفضل دقة عند التصنيف إلى مصنفين (ايجابي ، سلبي) مع درجة 93٪ ، بينما حقق نموذج MLPClassifier أفضل دقة عند التصنيف إلى مصنّفين (محايد ، غير محايد) بنسبة 92٪.**

Table of Contents

[الايـــــــــــــــــــــة i](#_Toc129820751)

[DEDICATION ii](#_Toc129820752)

[ABSTRACT iii](#_Toc129820753)

[المستلخص iv](#_Toc129820754)

[Table of Figures viii](#_Toc129820755)

[Table of Tables x](#_Toc129820756)

[Abbreviations Table xi](#_Toc129820757)

[**CHAPTER ONE 1**](#_Toc129820759)

[INTRODUCTON 1](#_Toc129820760)

[1.1 Overview 2](#_Toc129820761)

[1.2 Problem Statement 3](#_Toc129820762)

[1.3 Research Aim and Objectives 4](#_Toc129820763)

[1.4 Research Motivations 4](#_Toc129820764)

[1.5 Research Scope 5](#_Toc129820765)

[1.6 Methodology 5](#_Toc129820766)

[1.7 Related Works 6](#_Toc129820767)

[1.8 Research Structure 8](#_Toc129820768)

[**CHAPTER TWO 9**](#_Toc129820769)

[THEORETICAL BACKGROUND AND LITERATURE REVIEW 9](#_Toc129820770)

[2.1 Research Background 10](#_Toc129820771)

[2.1.1 Introduction 10](#_Toc129820772)

[2.1.2 Sentiment Analysis and Opinion Mining 10](#_Toc129820773)

[2.1.3 Challenges of Sentiment Analysis 11](#_Toc129820774)

[2.1.4 Application of Sentiment Analysis: 12](#_Toc129820775)

[2.1.5 Social Media 13](#_Toc129820776)

[2.1.6 Social Networks Analysis 13](#_Toc129820777)

[2.1.7 Arabic Language 14](#_Toc129820778)

[2.1.8 Sudanese Delicate 15](#_Toc129820779)

[2.1.9 Textual Features Presentation Techniques 15](#_Toc129820780)

[2.1.9.1 Bag-of-Words (BOW) 16](#_Toc129820781)

[2.1.9.2 Term frequency-inverse Document Frequency (TF IDF) 18](#_Toc129820782)

[2.1.9.3 Word2vec language model 19](#_Toc129820783)

[2.1.10 Classification Techniques 19](#_Toc129820784)

[2.1.11 Sentiment Analysis Levels 19](#_Toc129820785)

[2.1.12 Sentiment Analysis Methodologies 20](#_Toc129820786)

[2.1.13 Machine Learning Classification Models 21](#_Toc129820787)

[2..1.13.1 Unsupervised Approach 21](#_Toc129820788)

[2.1.13.2 Supervised Approach 21](#_Toc129820789)

[2.1.13.3 Semi‑supervised Learning: 22](#_Toc129820790)

[2.1.14 Classification Analysis 22](#_Toc129820791)

[2.1.15 Lexical Based Approach 28](#_Toc129820800)

[2.1.16 Anaconda 28](#_Toc129820801)

[2.1.17 Jupyter Notebook Application 29](#_Toc129820802)

[2.2 Literature Review 30](#_Toc129820803)

[2.2.1 Introduction 30](#_Toc129820804)

[2.2.2 Literature Review 30](#_Toc129820805)

[2.2.3 Table Summarizing Literature Review 33](#_Toc129820806)

[2.2.4 Summery of Literature Review 35](#_Toc129820807)

[**CHAPTER THREE 36**](#_Toc129820808)

[METHODOLOGY 36](#_Toc129820809)

[3.1 Introduction 37](#_Toc129820810)

[3.2 Data Collection and Reparation 37](#_Toc129820811)

[3.3 Data preprocessing 39](#_Toc129820812)

[3.3.1 Data Cleaning 39](#_Toc129820813)

[3.3.2 Removing Duplicated Characters 39](#_Toc129820814)

[3.3.3 Normalization 39](#_Toc129820815)

[3.3.4 Removing Stop Words 40](#_Toc129820816)

[3.3.5 Handling Negation 40](#_Toc129820817)

[3.3.6 Tokenization 41](#_Toc129820818)

[3.3.7 Lemmatization 41](#_Toc129820819)

[3.4 Feature Extraction 42](#_Toc129820820)

[3.4.1 Term Frequency Inverse Document Frequency(TF-IDF) 42](#_Toc129820821)

[3.4.2 Bag of Word(BOW) 43](#_Toc129820822)

[3.5 Data Splitting 43](#_Toc129820823)

[3.6 Classification Model 47](#_Toc129820824)

[3.6.1 Binary Classification Model 47](#_Toc129820825)

[3.7 Multi Classification Model 48](#_Toc129820826)

[3.8 Performance Analysis 48](#_Toc129820827)

[3.8.1 Confusion Matrix 48](#_Toc129820828)

[3.8.2 Accuracy 48](#_Toc129820829)

[3.8.3 PRECISION 48](#_Toc129820830)

[3.8.4 RECALL 49](#_Toc129820831)

[3.8.5 F1 SCORE 49](#_Toc129820832)

[3.9 Model Training 49](#_Toc129820833)

[**CHAPTER FOUR 52**](#_Toc129820834)

[IMPLEMENTATION 52](#_Toc129820835)

[4.1 Introduction 53](#_Toc129820836)

[4.2 Import Libraries 53](#_Toc129820837)

[4.3 Data Preprocessing 54](#_Toc129820838)

[4.3.1 Normalization 54](#_Toc129820839)

[4.3.2 Remove Stop Words 55](#_Toc129820840)

[4.3.3 Handling Negation 55](#_Toc129820841)

[4.3.4 Tokenization: 56](#_Toc129820842)

[4.3.5 Remove Duplication 56](#_Toc129820843)

[4.4 Classification Techniques 58](#_Toc129820844)

[4.4.1 Multiple Classification 58](#_Toc129820845)

[4.4.2 Binary Classification 59](#_Toc129820846)

[4.6 Data Splitting: 60](#_Toc129820847)

[4.7 Feature Extraction 61](#_Toc129820848)

[4.7.1 TFIDF 61](#_Toc129820849)

[4.8 Train model 62](#_Toc129820850)

[4.9 Classification Algorithms 62](#_Toc129820851)

[**CHAPTER FIVE 67**](#_Toc129820852)

[RESULT AND RECOMMENDATION 67](#_Toc129820853)

[5.1 Introduction 68](#_Toc129820854)

[5.2 Results 68](#_Toc129820855)

[5.3 Model Evaluation 86](#_Toc129820856)

[5.5 Recommendations 87](#_Toc129820857)

[References 88](#_Toc129820858)

[Appendices 92](#_Toc129820859)

[Appendix (I) SAMPLE OF DATASET 93](#_Toc129820860)

[Appendix (II) Classification Comments using Google Form 94](#_Toc129820861)

[Appendix (III) Jupyter Notebook 95](#_Toc129820862)

# **Table of Figures**

[Fig 1. 1Methodology 17](#_Toc129675144)

[Fig 2. 1 The processes related to sentiment analysis in social networks 20](#_Toc128614206)

[Fig 2. 2 The challenges in sentiment analysis 21](#_Toc128614207)

[Fig 2. 3 Number of Monthly Active Facebook Users 2008-2021 22](#_Toc128614208)

[Fig 2. 4 Sentiment analysis methodologies 31](#_Toc128614209)

[\_Toc128662845](#_Toc128662845)

[Fig 3. 1 Data classes 48](#_Toc128666296)

[Fig 3. 2 multiple Classification Training Data 54](#_Toc128666297)

[Fig 3. 3 Multiple Classification Testing Data 54](#_Toc128666298)

[Fig 3. 4 Binary Classification Data 55](#_Toc128666299)

[Fig 3. 5 Binary Classification Data 55](#_Toc128666300)

[Fig 3. 6 Binary Classification Testing Data 56](#_Toc128666301)

[Fig 3. 7 Binary Classification Training Data 56](#_Toc128666302)

[Fig 3. 8 Binary Classification Testing Data 57](#_Toc128666303)

[Fig 3. 9 Methodology 47](#_Toc128666304)

[Fig 4. 1 imported libraries 65](#_Toc128614265)

[Fig 4. 2 Normalization 66](#_Toc128614266)

[Fig 4. 3 Remove Stop Words 66](#_Toc128614267)

[Fig 4. 4 Handling Negation 67](#_Toc128614268)

[Fig 4. 5 Tokenization 67](#_Toc128614269)

[Fig 4. 6 Remove Duplication 68](#_Toc128614270)

[Fig 4. 7 Lemmatization 68](#_Toc128614271)

[Fig 4. 8 Read Data File 69](#_Toc128614272)

[Fig 4. 9 Data in Three Classifier Model 70](#_Toc128614273)

[Fig 4. 10 Tow Classifier Model 70](#_Toc128614274)

[Fig 4. 11 Data after exclude data with a neutral classification 70](#_Toc128614275)

[Fig 4. 12 Classifying positive and negative comments as non-neutral 71](#_Toc128614276)

[Fig 4. 13 Data after Classifying positive and negative comments as non-neutral 71](#_Toc128614277)

[Fig 4. 14 Splitting Data 72](#_Toc128614278)

[Fig 4. 15 Data Partition Ratios 72](#_Toc128614279)

[Fig 4. 16 Implement TFIDF 72](#_Toc128614280)

[Fig 4. 17 Implement BOW 73](#_Toc128614281)

[Fig 4. 18 Function that trains the model 74](#_Toc128614282)

[Fig 4. 19 SVC 74](#_Toc128614283)

[Fig 4. 20 MultinomialNB 74](#_Toc128614284)

[Fig 4. 21 DecisionTreeClassifier 75](#_Toc128614285)

[Fig 4. 22 Logistic Regression 75](#_Toc128614286)

[Fig 4. 23 MLPClassifier 75](#_Toc128614287)

[Fig 4. 24 KNN 76](#_Toc128614288)

[Fig 4. 25 RandomForestClassifier 76](#_Toc128614289)

[Fig 4. 26 Gradient Boosting 77](#_Toc128614290)

[Fig 4. 27 AdaBoostClassifier 77](#_Toc128614291)

[Fig 4. 28 Voting Classifier 78](#_Toc128614292)

[Fig 5. 1 Experimental (1.1) 82](#_Toc129675125)

[Fig 5. 2 Experimental (1.2) 83](#_Toc129675126)

[Fig 5. 3 Experimental (1.3) 84](#_Toc129675127)

[Fig 5. 4 Experimental (1.4) 85](#_Toc129675128)

[Fig 5. 5 Experimental (2.1) 86](#_Toc129675129)

[Fig 5. 6 Experimental (2.2) 87](#_Toc129675130)

[Fig 5. 7 Experimental (3.1) 88](#_Toc129675131)

[Fig 5. 8 Experimental (3.2) 89](#_Toc129675132)

[Fig 5. 9 Experimental (3.3) 90](#_Toc129675133)

[Fig 5. 10 Experimental (3.4) 91](#_Toc129675134)

[Fig 5. 11 Experimental (3.5) 92](#_Toc129675135)

[Fig 5. 12 Experimental (3.6) 93](#_Toc129675136)

[Fig 5. 13 Experimental (4.1) 94](#_Toc129675137)

[Fig 5. 14 Experimental (4.2) 95](#_Toc129675138)

[Fig 5. 15 Experimental (4.3) 96](#_Toc129675139)

[Fig 5. 16 Experimental (4.4) 97](#_Toc129675140)

[Fig 5. 17 Experimental (4.5) 98](#_Toc129675141)

[Fig 5. 18 Experimental (4.6) 99](#_Toc129675142)

**Table of Tables**

[Table 2. 1 Representation of documents 18](#_Toc129821063)

[Table 3. 1 Comments Classification 38](#_Toc128674403)

[Table 3. 2 Data Cleaning 38](#_Toc128674404)

[Table 3. 3 Removing Duplication 38](#_Toc128674405)

[Table 3. 4 Normalization 39](#_Toc128674406)

[Table 3. 5 Removing Stop Words 39](#_Toc128674407)

[Table 3. 6 Handling negation 40](#_Toc128674408)

[Table 3. 7 Tokenization 40](#_Toc128674409)

[Table 3. 8 Lemmatization 41](#_Toc128674410)

[Table 3. 9 Binary Classification 46](#_Toc128674411)

[Table 3. 10 multi classification 47](#_Toc128674412)

[Table 5. 1 Experimental (1.1) 70](#_Toc128665754)

[Table 5. 2 Experimental (1.2) 71](#_Toc128665755)

[Table 5. 3 Experimental (1.3) 71](#_Toc128665756)

[Table 5. 4 Experimental (1.4) 72](#_Toc128665757)

[Table 5. 5 Experimental (2.1) 73](#_Toc128665758)

[Table 5. 6 Experimental (2.2) 74](#_Toc128665759)

[Table 5. 7 Experimental (3.1) 76](#_Toc128665760)

[Table 5. 8 Experimental (3.2) 77](#_Toc128665761)

[Table 5. 9 Experimental (3.3) 78](#_Toc128665762)

[Table 5. 10 Experimental (3.4) 79](#_Toc128665763)

[Table 5. 11 Experimental (3.5) 80](#_Toc128665764)

[Table 5. 12 Experimental (3.6) 81](#_Toc128665765)

[Table 5. 13 Experimental (4.1) 82](#_Toc128665766)

[Table 5. 14 Experimental (4.2) 83](#_Toc128665767)

[Table 5. 15 Experimental (4.3) 84](#_Toc128665768)

[Table 5. 16 Experimental (4.4) 85](#_Toc128665769)

[Table 5. 17 Experimental (4.5) 86](#_Toc128665770)

[Table 5. 18 Experimental (4.6) 87](#_Toc128665771)

**Abbreviations Table**

|  |  |
| --- | --- |
| **Terms** | **Abbreviation** |
| **SVM** | **Support Vector Machine** |
| **LR** | **Logistic Regression** |
| **NB** | **Naive Bayes** |
| **DT** | **Decision Tree** |
| **TF-IDF** | **Term Frequency Inverse Document Frequency** |
| **BOW** | **Bag Of Words** |
| **NLP** | **Natural Language Processing** |
| ME | **Maximum Entropy** |
| **TN** | **True Negative** |
| **TP** | **True Positive** |
| **FP** | **False Positive** |
| **FN** | **False Negative** |
| **NLTK** | **Natural Language Toolkit** |
| **MLPClassifier** | **Multi-layer Perceptron Classifier** |
| **NumPy** | **Numerical Python** |
| **SVC** | **Support Vector Classification** |

# **CHAPTER ONE**

# **INTRODUCTON**

# 

**CHAPTER ONE**

**INTRODUCTON**

# **1.1 Overview**

The 21st century has seen many innovations, including the development of social media platforms. These platforms have led to interactions between people and changed how news is reported, as people can now express their opinions, unlike before, when only reporters were speaking. Social media has become the most influential source of freedom of expression and emotions on its platforms. Anyone can express feelings using social media platforms like Facebook, Twitter, Instagram, and YouTube. Raw data is growing daily for every culture and area of life, so there is a need to process this raw data to get helpful information. If any nation or country wants to know the needs of its people, there must be data mined that shows the actual meaning of their emotions.

The sciences of artificial intelligence, machine learning, and deep learning have contributed to many applications, especially in natural language processing, and are of great interest. Analyzing people's comments and tweets on social media is an essential issue with many practical applications. For example, online stores are very interested in analyzing customers' comments on their products to explore customer trends, weaknesses, and strengths in the store. Social media such as Twitter and Facebook have become very important in knowing people's attitudes on a particular topic and exploiting and using this data produced by social media companies and governments to gauge opinions on a particular topic or product.[1]

Sentiment analysis can be expressed by extracting valuable patterns from textual data. These functional patterns involve interpreting and categorizing feelings as neutral, positive, or negative from that data using specific analysis techniques. From a business point of view, companies can, by analyzing users' opinions through opinion polls or their interactions (their posts or comments about a specific product) on social networking sites, respond to their needs and improve their services and products to suit those needs.[2]

Measuring customer satisfaction and their reactions to the services and products has always been an obsession for the companies and institutions that provide these services, intending to help them make better decisions to improve the quality of their products and services and increase their satisfaction. Customers and gain their trust. This process was difficult and expensive many years ago and followed the traditional methods.

Sentiment analysis is a type of textual analysis also known as mining. It relies on statistical analysis, natural languages, and machine language to locate and extract information from text files. For example, the feelings of people who write their opinions, thoughts, judgments, or assessments of a particular topic, event, company, or activity. This analysis is known as "opinion mining" (with focus extraction) or "dynamic classification." Despite its many different names, this analysis aims to learn the opinions of users or the public about a particular topic by analyzing a large body of text from multiple sources.[3]

This analysis can be used at multiple levels, which depend on the analysis's goal. Another example is if we want to determine if our visitors like or dislike a group of clothes and why. Instead, they compare it to another similar product. It must analyze each opinion or statement, focusing on particular aspects and using keywords. Social networking sites are one of the most prominent forms of daily life in our contemporary reality. People find complete freedom to express their negative and positive opinions. Therefore, this vast amount of primary data must be taken advantage of, collected from websites, analyzed, and used in making decisions.

In this research, we show the use of machine learning techniques in analyzing the feelings of texts written in Arabic and the Sudanese dialect. The language used is not necessarily classical Arabic but can include colloquial expressions that users usually use.

# **1.2 Problem Statement**

Sudanese organizations and company have many data about customer feedback, comments, suggestions, problems, and satisfaction with the service. However, these statements are primarily in the colloquial Sudanese dialect. Therefore, it isn't easy to analyze and classify them and then extract information and statistics that help decision-making. Traditional methods can be less efficient in terms of being time-consuming, unreliable data, large workforce, inaccurate and biased Results.

# **1.3 Research Aim and Objectives**

The research aims to study, design and build a system that analyzes customer posts and reviews published in Arabic on social media to determine their general opinion, whether positive, negative or neutral, based on the concepts and techniques of sentiment analysis. Moreover, prospecting for opinions helps the decision-maker in institutions (companies). Contact) to take the necessary measures to improve the quality of services this institution provides.

Improving product or service quality: Manufacturers use opinions about a product or service as a response to decision-making to improve the quality of the product or service. Therefore, opinion mining and sentiment analysis are used to analyze online consumer or customer opinions of Facebook to determine the advantages and disadvantages of online services. This process saves money in getting customer or user reviews in advance.

# **1.4 Research Motivations**

The traditional method can be used to analyze textual data from several customer interviews, a focus group, and even several hundred tweets or journalistic articles. However, when amount of text that must be analyzed increases, and the human effort required must be repeated with each increase while using sentiment analysis.

Sentiment analysis is very effective when it deals with live data generated by users' comments on the new product's Facebook page or following tweets on a specific hashtag. This is only available through sentiment analysis algorithms integrating data mining tools, analyzing feelings manually, or what is known as objective analysis, by two different people may result in two different polarities, in addition to paying attention to the possibility of bias and objectivity for one reason or another. In contrast, automated sentiment analysis provides a systematic, consistent and objective method. most companies use English as their means of communication, but it is only used by some consumers worldwide.

According to recent studies, about 13% of the world's population speaks English. In addition, the British Council states that around 25% of the world's population has a good understanding of the English language. Many consumers interact with each other and the company in a language other than English. If companies' primary goal is to keep their customer base intact and attract new ones, then they need to understand their customers' opinions expressed in their native language. Manually reviewing or translating each comment into English is simple and will yield ineffective results. A sustainable solution is to develop language models for sentiment analysis that detects and analyzes customer opinions, emotions, and suggestions in social media, forums, surveys, and more.

# **1.5 Research Scope**

• Building a model for Sentiment analysis on the Sudanese dialect into positive, negative and neutral.

• Use the model on comments from Sudani Telecom's page.

# **1.6** **Methodology**

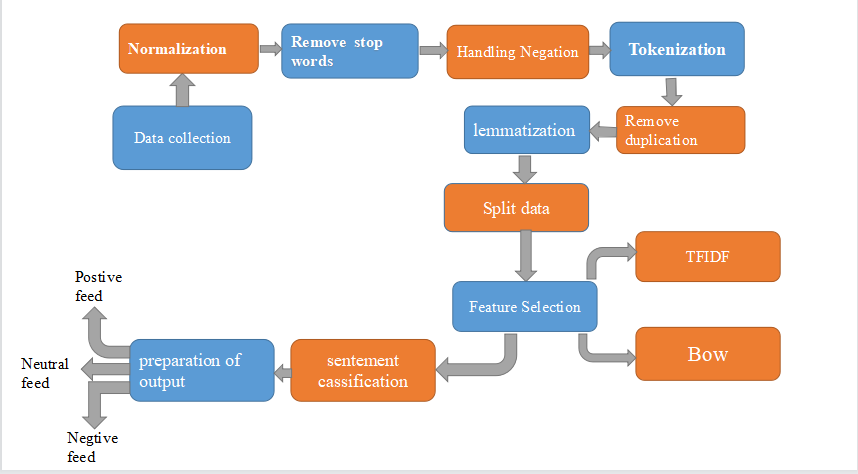


Fig 1.1 Methodology

# **1.7** **Related Works**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Study Objectives** | **Dataset** | **Methodology** | **The Obtained Result** | **Our opinion** |
| 1. Sentiment Analysis Using Systems Based on Machine Learning[26] | analyze the feelings of users using systems based on Machine Learning | 8560 Comments | NB, SVM,  KNN | NAIVE BASE, SVM, and KNN algorithms achieved high accuracy on different data sets, and the model was published as a web application on the Internet. | The researcher made an estimated effort on the Libyan dialect, where he used several classification algorithms on the data set in the Libyan dialect. It must also be done on the rest of the Arabic dialects. |
| 2. Application of sentiment analysis to comments in the Sudanese dialect [27] | Evaluate how processing steps and libraries affect legitimization On the general Sudanese text | 1050 Comments | NB,SVM | Use with libraries lemitimization Improves accuracy of sentiment categorization | The researcher worked on 1050 comments in the Sudanese dialect using two classification algorithms, which opens the way for the use of additional classification algorithms on a larger set of data. |
| 3 Sentiment Analysis of Arabic Tweets Written  by Sudanese Dialect [12] | classifying emotions and comparing current machine learning techniques applied to sentiment analysis for data collected from The social network | 8268 Tweets | NB, SVM, KNN, DT | High accuracy was achieved through the NAIVE BAYES, SVM, and decision tree algorithms during the two experiments. | In his work, the researcher used data in the Sudanese dialect, some of which are general, and some of them are related to the Sudanese government from the Twitter website, creating a space for collecting data from other sites on different affairs. |
| 4. Sentiment Analysis in Arabic Tweets : The Challenges of Language Anatomy [12] | Reviewing Efforts to Build Sentiment Analysis Systems for Arabic | Applications and systems created to analyze Arabic Twitter data | NB, SVM | There is no single system until the date of the study that can deal with all Arabic dialects with high accuracy | The challenges identified by the study must be taken into account in order to build a good model. The researcher did not mention the size of the data set used |
| **Study** | **Study Objectives** | **Dataset** | **Methodology** | **The Obtained Result** | **Our opinion** |
| 5. Analyzing Opinions from Sudanese Comments [32] | Classification of Sudanese comments into religious, entertainment, political, sport and others | Sudanese tweets on Twitter | K means & Hierarchical  R language. | The percentage of frequency of vocabulary in each classification is given | The researchers collected data in the Sudanese dialect from the Twitter website, which were classified into categories based on the subject, unlike what is done in sentiment analysis |
| 6. Twitter data set Moroccan feelings [33] | Measuring the Twitter data set through experiments on four classification methods | Tweets Twitter Moroccan  12k Tweets | SVC, KNN, DT, MNB, BNB,XGB, LR, LinearSVC, RandomForest | accuracy provided by each of the classification algorithms is similar, and the highest score was achieved using a model | The researcher made an estimated effort on the Moroccan dialect in terms of the size of the data set and the number of algorithms used. It must also be done on the rest of the Arabic dialects. |
| 7.Research published in the Journal of King Saud University [34] | Developing an approach to classify comments in the Algerian Arab online press into positive and negative | Comments in the Algerian electronic press | SVM, NB , KNN | NB classifier , the best results were obtained. | The researcher did not explain the size of the data set, but a good number of algorithms were used to work on the Algerian dialect |
| 8. Sentiment Analysis in Arabic and English [35] | Extracting feelings and determining their poles (positive, negative, neutral) | Twitter, Facebook, YouTube  4050 | Social Mention, Twendz | Give the percentages for experimental accuracy to determine the polarity of each suspension | The research dealt with both the Arabic and English languages. There is a need to take into account the dialects spoken in each of the two languages. |
| 9.Analyzing the Sentiments of Twitter Tweets During the US Presidential Elections 2020 [36] | Monitoring and analyzing the feelings of Twitter tweets during the US elections to identify the positive and negative feelings of the tweets | Twitter  425442 | LDA | sentiment analysis can be an accurate and low-cost way to measure public opinion towards candidates and predict election results | The research worked on a good set of data that must be provided in the Arabic dialects |
| 10. A review of feelings research in Arabic [37] | Identify important gaps in reading and writing and suggest future directions | Twitter , Yahoo, News Sites, Facebook | SVM, NB | Experiments revealed that performance is highly dependent on the quality of emotional resources | The conclusion of the study is important to consider when proceeding to build a model |

Table 1. 1 Summery of Literature Review

# **1.8 Research Structure**

The research consists of five chapters:

**Chapter One** represents an introduction to the research, as it discusses the basic concepts such as presenting the problem, the proposed solution, the scope of the research, the importance and objectives of the research, and the research structure.

**Chapter Tow** reviews an overview of the concepts in the field of sentiment analysis used in the research and presents previous studies in the second part of it

**Chapter Three** discusses in detail the methodology used by the researchers in building the model

**Chapter Four** presents the codes used in implementing the methodology to achieve the desired results

**Chapter Five** presents the results, evaluates the model, and discusses the results and recommendations

# **CHAPTER TWO**

# **THEORETICAL BACKGROUND AND LITERATURE REVIEW**

**CHAPTER TWO**

**THEORETICAL BACKGROUND AND LITERATURE REVIEW**

# **2.1 Research Background**

## 2.1.1 Introduction

Sentiment evaluation has emerged as a critical science to reap perception from social networks. The area has reached a stage of maturity that paves the way for its exploitation in many unique fields, such as marketing, health, banking and politics. Modern-day technological advancements, such as deep studying techniques, have solved some of the usual challenges brought about by the shortage of lexical resources. In this Special Issue, unique methods that enhance this self-discipline are presented.

In this chapter, we discussed Sentiment Analysis and Opinion Mining, social network analysis, algorithms for sentiment analysis and classification algorithms.

## 2.1.2 Sentiment Analysis and Opinion Mining

Opinions are central to nearly all human things to do and are key influence of our behaviors. Our beliefs and perceptions of reality, and the alternatives we make are to an extensive degree, conditioned upon how others see and evaluate the world. For this reason, when we want to make a selection we often are searching for out the opinions of others. This is no longer solely genuine for folks but also authentic for organizations.

Opinions and associated principles such as sentiments, evaluations, attitudes and feelings are the topics to learn about in sentiment evaluation and opinion mining [1]. Sentiment Analysis (SA), or Opinion Mining (OM), is the computational study of people's opinions, attitudes and emotions toward an entity. Sentiment analysis (SA) is an ongoing research area in text mining. SA is the computational processing of the text's opinions, feelings, and subjectivity. Sentiments are analyzed through several steps shown in figure (2-1), starting with entering the text and ending with classifying it as appropriate. [2]



Fig 2. 1 The processes related to sentiment analysis in social networks

## 2.1.3 Challenges of Sentiment Analysis

Sentiment evaluation or classification is viewed as a specific case of textual content classification in herbal language processing. Although the variety of lessons in sentiment evaluation is small, the technique of sentiment classification is extra challenging than the typical theme textual content classification [As shown in the figure (2-2), which presents some of the challenges related to the field of sentiment analysis] [5]

In subject textual content classification, classification depends on the usage of keywords. However, this does not typically work correctly in the case of sentiment evaluation [6]—the different difficulties in sentiment evaluation come from the nature of this problem. Sometimes, the poor sentiment would possibly be expressed in a sentence barring the usage of any prominent poor words. Moreover, there is a first-rate line between whether or not a sentence should be labelled goal or subjective. Determining the opinion holder -the one who expresses the sentiment in the text- is one of the most complex duties in sentiment analysis.

The sentiment evaluation pretty relies upon the area of the data. The phrases occasionally have high-quality sentiment in a particular domain, whereas they have another polarity sentiment in a one-of-a-kind domain. Finally, some different writing patterns, such as irony, sarcasm, or negated sentences, may want to deliver extra challenges to sentiment evaluation [5]



Fig 2. 2 The challenges in sentiment analysis

## 2.1.4 Application of Sentiment Analysis:

People share knowledge, experiences and ideas with the world through the usage of Social Media like blogs, forums, wikis, review sites, social networks, tweets, etc. This has modified how humans speak and impacted the social, political and monetary conduct of different humans on the Web Indeed. The Web approves all and sundry having a voice, promising to enhance human collaboration skills worldwide, enabling men and women to share opinions using read-write Web and user’s generated content. According to an opinion, “is certainly a wonderful or negative sentiment, view, attitude, emotion, or appraisal about an entity or an issue of the entity” from an opinion holder at a specific time.

The entity can be a product/service, event, person, organization, or subject matter consisting of aspects (features/attributes) representing each element and attribute of the entity. With the explosion of person-generated opinions, companies, politicians, carrier providers, social psychologists, researchers and different actors need to analyze them to put into effect higher selection choices.

The literature on sentiment evaluation targeted different domains, from administration sciences to pc science, social sciences and commercial enterprise, due to its significance to society as complete and special duties such as subjective expressions, sentiments of words, and subjective sentences [7].

## 2.1.5 Social Media

Social data refers to information gathered via social media networks. It demonstrates how consumers interact with the product by accessing, posting, and exchanging over via platforms as Facebook which is considered one of the applications that attract large numbers of users every year, as shown in the following figure (2-3) Academic study on individual, group and behavior uses social media as a dynamic data source. It refers to Internet apps that are web or mobile-based that enable users to create, access, and trade user generated content. [3]

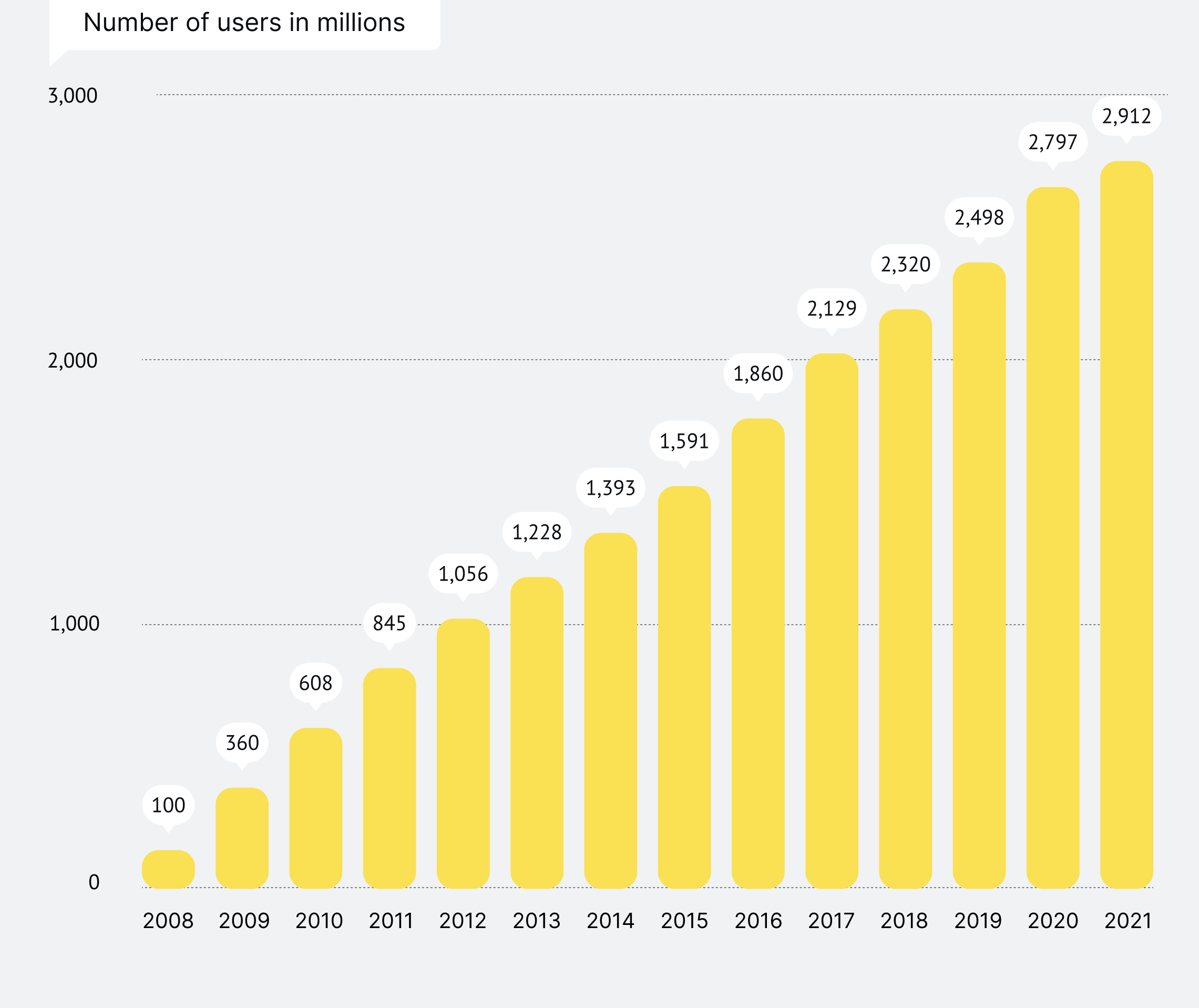


Fig 2.3 Number of Monthly Active Facebook Users 2008-2021

## 2.1.6 Social Networks Analysis

Social media is media used for social interaction. They are enabled with the aid of conversation technologies such as the net and smartphones and they flip conversation into an interactive dialogue.

Interactions on social media being fairly distributed, decentralized and occurring in actual time, they supply the indispensable breadth and immediacy of data required in instances of emergencies[8]**.** Since social media provide a uniquely fast and effective way to disseminate information, correct and inaccurate, top and terrible unfold equally alike as incorrect information can unfold like wild fire. However, there is indication that social networks have a tendency to favor valid statistics over rum ours[9].

Twitter and Facebook are accurate examples of social media beneficial in disaster conditions because they provide vital records as they are happening. Twitter is a micro-blogging service, a structure of lightweight chat permitting customers to put up and change brief 140-character-long messages regarded as tweets. Although most tweets are dialog and chatter, they are additionally used to share applicable information and file information [9]. Twitter is turning into a precious device in catastrophe and emergency situations as there is growing proof that it is no longer simply a social network, it is additionally an information service [10]**.**

Social Network Analysis (SNA) is a sociological method for analyzing patterns of relationships and interactions between social actors in order to find out underlying social shape such as: central nodes that act as hubs, leaders or gatekeepers; tremendously related groups; and patterns of interactions between businesses [11]. SNA has been used to find out about social interplay in a wide range of domains. Examples include: collaboration networks administrators of companies [12], organizational conduct [13], inter-organizational family members **,** computer-mediated communications **,** and many others.[5].

## 2.1.7 Arabic Language

The focal point is in the direction of sentiments expressed in the Arabic language due to the developing populace of Internet customers that use the Arabic language; it is estimated about 5% of global customers. Also, in the ultimate few years, it is regarded one of the most rising languages on the web. The morphological complexity and the lexical ambiguity of the Arabic language can pose a project when working with it. Another trouble is that the Arabic language has three one-of-a-kind varieties: classical Arabic (CA), present day trendy Arabic (MSA) and dialectical Arabic (DA).

Throughout this article The identification of fantastic and terrible opinions is no longer an effortless task; it requires data retrieval, linguistic knowledge, herbal language processing and a profound comprehension of the textual context.

A majority of the SSA research have targeted solely on theEnglish or European languages [14]. Some research analyzed themorphologically-rich languages like Urdu, Arabic, andTurkish [12]. The dialects in the Arabic language are verychallenging. This trouble is extra extraordinary if people, who submittexts on social networking sites, use a casual fashion ofwriting, or if the bi/multi-lingual folks use acombination of extraordinary languages. Any tweet can show the variability and complexity ofthe sentiments. Though the twitter messages are short, theycan be existing a large array of statistics in a compressed format[15].Tweets deliver sarcasm, or an unclear/mixed polarity asfollows: "يعم والشر يخص الخير". high-context cultures characterize achallenge in tweets, now not solely due to sarcasm however additionally due tototally unique meanings of one phrase e.g. the phrase “Eshta”, "إشطة"**l** for Egyptian dialect and the phrase "Zabit" " زابط " in Sudanese delicate [16]

## 2.1.8 Sudanese Delicate

The Sudanese dialect is an Arabic dialect spoken in the northern part of Sudan and in most of the center of the country and to a light extent in the south of the country. The Sudanese dialect is closer to the Hijazi dialect. The letter Qaf is also pronounced like the letter G in the English language. The Sudanese dialect borrows some words from the Nubian languages, and is also used throughout Sudan. It is considered a link between the people and is known among non-Arab groups in the south as the Juba dialect or Juba Arabic. The Sudanese dialect is considered one of the largest Arabic dialects spoken today and is spoken in several neighboring countries. It is spoken in western Eritrea, eastern Chad, Central Africa and southern Sudan.

## 2.1.9 Textual Features Presentation Techniques

**Sentence Tokenization:** Sentence tokenization (also called sentence segmentation) is the problem of dividing a string of written language into its component sentences. The idea here looks very simple. In English and some other languages, we can split apart the sentences whenever we see a punctuation mark.

However, even in English, this problem is not trivial due to the use of full stop character for abbreviations. When processing plain text, tables of abbreviations that contain periods can help us to prevent incorrect assignment of sentence boundaries. In many cases, we use libraries to do that job for us, so don’t worry too much for the details for now.

**Word Tokenization:** Word tokenization (also called word segmentation) is the problem of dividing a string of written language into its component words. In English and many other languages using some form of Latin alphabet, space is a good approximation of a word divider.

However, we still can have problems if we only split by space to achieve the wanted results. Some English compound nouns are variably written and sometimes they contain a space. In most cases, we use a library to achieve the wanted results, so again don’t worry too much for the details.

**Text Lemmatization and Stemming** for grammatical reasons, documents can contain different forms of a word such as drive, drives, driving. Also, sometimes we have related words with a similar meaning, such as nation, national, nationality.

**Word-document Representation**:In order to have a better grasp of vector semantics, let’s assume that we have a set of texts (documents) and we want to find documents which are similar to each other. This task is relevant in information retrieval, for example in search engines, where documents are web pages.As illustration, each column in the table below represents one of 4 documents with the following titles: “As You Like It”, “Twelfth Night”, “Julius Caesar”, and “Henry V”. Words which appear in the documents are represented as rows. These words build our vocabulary. The table tells us that the word “battle” occurs 7 times in the document “Julius Caesar”. This table is also called term-document matrix, where each row represents a word in the vocabulary and each column represents a document, a section, a paragraph, a tweet, a SMS, an email or whatever.

Data vector representation is a powerful tool for analyzing large datasets. It is used to represent data in a way that can be easily understood and analyzed by computers. By representing data as vectors, it becomes easier to identify patterns and correlations between different variables. This makes it an invaluable tool for predictive analytics, machine learning, and other data science applications. Data vector representation also helps to reduce the amount of time and resources needed to process complex datasets. With this technology, businesses can better understand their customer base, optimize marketing campaigns, and improve decision-making processes.

## 2.1.9.1 Bag-of-Words (BOW)

The task of automatically assigning categories to natural language text has become one of the key methods for organizing online information. Since hand-coding such classification rules is costly or even impractical, most modern approaches employ Machine Learning techniques to automatically learn text classifiers from examples. However, it is necessary to transform texts, or documents, into a form appropriate to interpretation by a classifier-building algorithm. These issues include term weighting and dimensionality reduction. The representation of documents has a crucial influence on how well the learning algorithm can generalize A collection of documents D = {d1, d2, . . . dn} and a set of categories C = {c1, c2, . . . cz} associated with the collection of documents D are required in order to perform a Text Mining — TM — task. In this work we shall concentrate in the categorization task, using a simple representation of the documents, which consists of inducting a classifier with the aim of determining if a document di belongs to a category cj with i = 1, 2, . . . n and j = 1, 2, . . . z. Besides the categorization task, other tasks such as document summarization and clustering may be related to the TM process, depending on the application nature and if the set of categories C is known. In any case, some phases are essential to the TM process. These phases are (1) document collection; (2) text preprocessing; (3) knowledge extraction; (4) results evaluation and interpretation. After the document collection phase, the collected documents should be transformed to a format accepted by the knowledge extraction algorithms. This second phase, text preprocessing, outputs a structure, frequently represented as an attribute-value table, that represents the document collection. This phase is computationally intensive and careful text preprocessing is essential to the success of the TM process. More details are presented in Section as soon as the documents are represented in a convenient format, algorithms for extracting knowledge can be applied with the objective of discovering patterns in the documents that are useful and previously unknown. Finally, the evaluation phase verifies if the objectives have been reached or if any of the phases should be carried out again

During this phase, the collected documents should be preprocessed in order to transform their representation. Frequently, each document di is transformed to a vector of terms that occur in the document. The identification of the terms in a document may be based on the words present in the text (bag-of-words), or more sophisticated representations [22]**.** However, results obtained by experimental research works have shown that more sophisticated representations could present inferior performance as compared with representations based on simple words[23]**.** According to, the most probable explanation of these results is that, even if more sophisticated terms present a superior semantical quality, the statistical quality is inferior as compared with simple word terms. In this way, research works related to the use of simple and sophisticated representations of documents are active [22]**.** ´ In the identification of terms as bag-of-words, a term can be represented by simple words (1-gram) or composed words (2, 3, . . ., n-gram) that occur in the document. Each term is used as an attribute of the data set represented in the attribute-value form. The set of documents, after preprocessing, can be represented as shown in Table 2-1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | T1 | T2 | … | Tm | C |
| D1 | A11 | A12 | … | A1m | C1 |
| D2 | A21 | A22 | … | A2m | C2 |
| …. | … | … | … | … | … |
| Dn | Dn1 | Dn2 | … | Anm | Cn |

Table 2. 1 Representation of documents

In Table 1 n documents (examples) are represented, and each document is composed by m terms. Each document di can be understood as a vector di = (ai1, ai2, . . . aim). An aij value refers to the value of the jth term of the ith document, i.e., aij represents the value of the term tj in the document di. If binary values are used, the value 1 means the presence of the term tj in the document di, and the value 0 means the term is absent. However, the binary representation of terms might be inadequate for some applications. Thus, statistical measures that take into account the frequency a term appears in a document and the frequency this term is found in other documents might provide better results. For instance, term frequency — tf — is a measure that counts the number of occurrences of a term ti in a document di, and is defined by tf (ti, di) = # (ti, di), where # (ti, di) is the number of occurrences of ti in di.

## 2.1.9.2 Term frequency-inverse Document Frequency (TF IDF)

Terms with high frequency, i.e., terms that occur in all (or in the majority) of the documents, frequently do not present useful information to discriminate the documents.

The measure inverse document frequency — idf — favors terms that occur in few documents of the collection. idf is inversely proportional to the number of documents having a certain term in a document collection, and is defined by idf = log n #Dtj, where n is the total number of documents in the document collection D, and #Dtj is the number of documents in D where the term tj occurs at least once. Another relevant aspect that should be considered when the measures tf and idf are applied is related to documents that have a large difference in the number of terms. Since larger documents usually have a higher probability of being relevant than smaller documents. However, all relevant documents should be considered as equally important, independently of their size. In this case, a normalization factor should be incorporated with the aim of normalizing the aij values of documents. The tf idf measure is extended to incorporate the normalization factor and is defined by tf idfn (tj, di) = pP tf idf (tj, di) n s=1 (tf idf (ts, di)). The presented measures are commonly used to assign values to terms that occur in the documents.

However, in an attribute-value table, each attribute is a term that occurs in the documents. Thus, some criteria to reduce the dimensionality of the data set should be considered. Several methods can be used in order to reduce the amount of attributes in the data set. The reduction of attributes can result in more representative attributes and in a better performance of the TM process. A widely used method consists of reducing each term to its radical. This transformation can be accomplished by the use of stemming algorithms, which are widely used in text processing. [16]

## 2.1.9.3 Word2vec language model

Representing words or documents by sparse and long vectors is not practically efficient. Those vectors are typically sparse because many positions are filled by zero values. They are long because their dimensionality equals the vocabulary size or the documents collection size.

## 2.1.10 Classification Techniques

Classification techniques are powerful tools used to identify and classify data into different categories. These techniques are used in a variety of applications, such as machine learning, artificial intelligence, computer vision, and natural language processing. Classification techniques use various algorithms and statistical models to determine the most appropriate class for each data point. By leveraging these techniques, businesses can better understand customer behavior and preferences to deliver more personalized experiences. Additionally, classification techniques can be used to detect anomalies in data sets or predict future events.

## 2.1.11 Sentiment Analysis Levels

There are three main classification levels in Sentiment Analysis: document level, sentence-level, and aspect-level.

Document-level sentiment analysis works at document level. It classifies an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document as a basic information unit. Sentence-level sentiment analysis aims to classify sentiment expressed in each sentence. Before analyzing the polarity of sentiments, there is need to identify whether the sentence is subjective or objective. It determines whether the sentence expresses positive or negative opinions. Both the document level and sentence level sentiment analysis cannot interpret the exact sentiment. Aspect or Feature level sentiment classification does finer-grained analysis. It concerns with identifying and extracting product features from the source data [24]

## 2.1.12 Sentiment Analysis Methodologies

Sentiment analysis is an important tool for businesses to understand their customer’s emotions and opinions. It uses natural language processing (NLP) and machine learning algorithms to analyze customer feedback, reviews, and other unstructured text data. By leveraging sentiment analysis methodologies, companies can gain valuable insights into customer sentiment and use this information to improve customer experience and increase customer satisfaction. This article will discuss various sentiment analysis methodologies, such as supervised learning, unsupervised learning, deep learning, rule-based approaches, lexicon-based approaches, hybrid approaches, etc., that can be used to accurately classify customer sentiment.

Numerous methodologies are available for opinion mining, but two main groups are used. The problems of Sentiment Analysis will be solved by the first group using by implementing the machine learning approach. The second group uses lexicon-based method which is a linguistically-inclined method. [25]

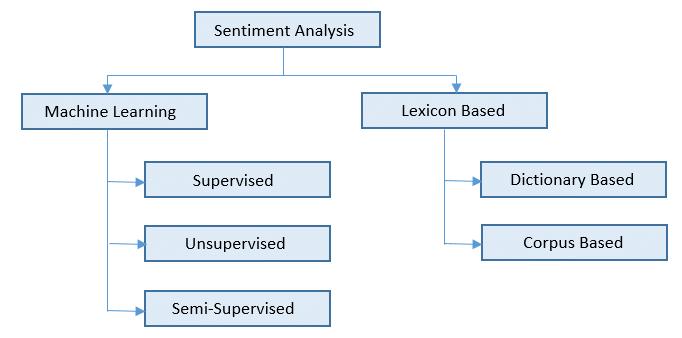


Fig 2. 4 Sentiment analysis methodologies

## 2.1.13 Machine Learning Classification Models

Machine learning strategies work by training an algorithm with a training data set before applying it to the actual data set. Machine learning techniques first trains the algorithm with some particular inputs with known outputs so that later it can work with new unknown data. Some of the most renowned works based on machine learning areas follows [7]

## 2..1.13.1 Unsupervised Approach

Extracting meaningful information from large datasets without any prior knowledge or labels. Unsupervised approaches are used in many areas such as natural language processing, image recognition, and anomaly detection. They are also used to identify patterns and relationships in data that would otherwise be difficult to discover with traditional methods.

## 2.1.13.2 Supervised Approach

Supervised learning is a machine learning approach that uses labeled data to train and evaluate algorithms. It is a type of predictive modeling technique, which can be used to identify patterns in data. The supervised approach involves providing the algorithm with labeled training data, which it then uses to make predictions about unseen data. This approach has been widely used for tasks such as object recognition, text classification, and speech recognition. By using labeled datasets, the supervised approach can help us better understand the underlying relationships between different variables in our dataset and make accurate predictions about future outcomes.

## 2.1.13.3 Semi‑supervised Learning:

In this case, where the training dataset contains both labelled and unlabeled data, semi supervised learning appears to be a viable option. It is motivated that while gathering unlabeled data is relatively easy in many real-world applications, such as collecting articles from various blogs, labelling is expensive or labour-consuming because labelling the training dataset is typically performed by humans [3].

## 2.1.14 Classification Analysis

Classification is regarded as a supervised learning method in machine learning, referring to a problem of predictive modeling as well, where a class label is predicted for a given

example [41]. Mathematically, it maps a function (f) from input variables (X) to output variables (Y) as target, label or categories. To predict the class of given data points, it can be

carried out on structured or unstructured data. For example, spam detection such as “spam” and “not spam” in email service providers can be a classification problem. In the following, we summarize the common classification problems.

**• Binary Classification:** It refers to the classification tasks having two class labels such as “true and false” or “yes and no”. In such binary classification tasks, one class could be the normal state, while the abnormal state could be another class. For instance, “cancer not detected” is the normal state of a task that involves a medical test, and “cancer detected” could be considered

as the abnormal state. Similarly, “spam” and “not spam” in the above example of email service providers are considered as binary classification.

**• Multiclass Classification:** Traditionally, this refers to those classification tasks having more than two class labels. The multiclass classification does not have the principle of normal and abnormal outcomes, unlike binary classification tasks. Instead, within a range of specified classes, examples are classified as belonging to one. For example, it can be a multiclass classification

task to classify various types of network attacks in the NSL-KDD [119] dataset, where the attack categories are classified into four class labels, such as DoS (Denial of Service Attack), U2R (User to Root Attack), R2L (Root to Local Attack), and Probing Attack.

**• Multi-label Classification:** In machine learning, multi-label classification is an important consideration where an example is associated with several classes or labels.

Thus, it is a generalization of multiclass classification, where the classes involved in the problem are hierarchically structured, and each example may simultaneously belong to more than one class in each hierarchical level, e.g., multi-level text classification. For instance, Google news can be presented under the categories of a “city name”, “technology”, or “latest news”, etc. Multi-label classification includes advanced machine learning algorithms that support predicting various mutually non-exclusive classes or labels, unlike traditional classification tasks where class labels are mutually exclusive.

Many classification algorithms have been proposed in the machine learning and data science literature .In the following, we summarize the most common and popular methods that are used widely in various application areas:[26]

## Support Vector Machines (SVMs)

SVMs are a set of related methods for supervised learning, applicable to both classification and regression problems. Since the introduction of the SVM classifier a decade ago [27], SVM gained popularity due to its solid theoretical foundation. The development of efficient implementations led to numerous applications [28]**.** The Support Vector learning machine was developed by Vapnik et al. to constructively implement principles from statistical learning theory [27]**.** In the statistical learning framework, learning means to estimate a function from a set of examples (the training sets). To do this, a learning machine must choose one function from a given set of functions, which minimizes a certain risk (the empirical risk) that the estimated function is different from the actual (yet unknown) function. The risk depends on the complexity of the set of functions chosen as well as on the training set. Thus, a learning machine must find the best set of functions - as determined by its complexity - and the best function in that set. Unfortunately, in practice, a bound on the risk is neither easily computable, nor very helpful for analyzing the quality of the solution.

Support vector algorithm under machine learning, used in classification or regression tasks. However, the support vector machine is mostly used in classification, as a result of the acceptable accuracy it provides the support vector machine is based on the idea of ​​finding a hyper plane, which divides the data into two groups in the best way

## Naïve Bayes (NB)

The naive Bayes [14], is a simple Bayesian net-work classifier that assumes that the predictors or variables are independent given each class value. Despite its simplicity and strong assumptions, the naive Bayes classifier has been proven to work satisfactorily in many domains. Typically, the parameters of the naive Bayes model are found by maximizing the joint likelihood of the model. The naive Bayes model's accuracy, however, declines in the presence of noisy predictors. A noisy predictor can be a predictor that either carries no useful information for the classification (irrelevant) or is strongly dependent on another predictor (redundant). Redundancy is particularly harmful, because the predictor information has double the influence than it should. For variable selection purposes, it is common to use Al-tering approaches, which perform variable selection disregarding the classifier, or (greedy) wrapper algorithms, which simultaneously introduce variables into the model and itera-tively estimate the parameters. We focus on the wrapper par-adigm. The (stepwise) selective naive Bayes is a popular example of greedy wrapper algorithm. Regularization techniques introduce additional information, usually to solve an ill-posed problem or to avoid over-fitting. Also, by imposing certain restrictions, regularization trades off a little bias against a larger reduction in variance. L\-regularization, which imposes an L\-penalty on the parameters, is also useful for variable selection, because it drives some parameters to exactly zero. An example of regularization within the naive Bayes model is the Li//^-regularized naive Bayes, taken by van Gerven and Heskes [9], which applies optimization techniques to minimize the negative log-likelihood function of the data given the model plus an Li/L2-group penalty on the model complexity. This penalty encourages some predict-tors to be discarded. While they apply this idea only to the continuous predictor case, we extend it to deal with discrete predictors. Also, we introduce an adaptive penalty that further improves the method's performance. The main contribution of this paper, however, is a stage-wise version of the selective naive Bayes that is particularly useful when there are predictors that are relevant but, to some extent, redundant. At each iteration, instead of adding an "entire" predictor to the model, the parameters of the selected predictor are updated just a little. This method is inspired by the forward stage wise selection method for linear regression, which is also related to boosting and can be considered a form of regularization. We call this method forward stage wise naive Bayes.

## Decision tree (DT)

It is a machine learning algorithm used in classification and regression tasks, and it is a powerful tool for knowledge representation usually a decision tree is defined as a hierarchical structure used to divide a large group from records into small successive groups by applying simplified decision rules according to a specific sequence, it is a classification technique with a tree-like structure and consists of three types of nodes (Root Node, Internal node, Leaf)

## Logistic regression (LR)

Logistic regression is machine learning method, which is a model used for classification and prediction, the probability of an event occurring by fitting the data on the logistic curve, and logistic regression is used several times Independent variables, which can be relative, categorical, nominal, or ordinal, corresponding to a dependent predictor variable, for example: the probability of the customer paying the loan amount or not paying it, can be predicted by knowing information about the marital status, age, salary, thus it is possible to predict the possibility of repaying the loan or not.

The logistic model analyzes the relationship between a group It is also known OR, by estimating the probability of an event occurring of the independent variables and the dependent variable is a monotonic classification Whether or not by installing a logistic curve.

## The K-Nearest Neighbor (KNN)

This approach finds the K nearest neighbors of a text document among the training documents. The classification is done on the basis of the similarity score of the class to the neighbor document. Winnow is another commonly used approach. The system first predicts a class for a particular document and then receives feedback. In presence of false classification (i.e., error) the system updates its weight vectors accordingly. This process is repeated over a collection of sufficiently large set of training data.[3]

## Maximum Entropy(ME)

In Maximum Entropy Classifier, no assumptions are taken regarding the relationship between features. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label. 3.1.5 Artificial Neural Network A neural network has emerged as an important tool for classification. During past decade neural network classification has established as a promising alternative to various conventional classification methods. The neural network with appropriate network structure can handle the correlation/dependence between input variables. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Second, they are universal functional approximates in that neural networks can approximate any function with arbitrary accuracy. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is doubtlessly important[3]

## Neural Networks

It is a system designed to simulate the way the human brain performs a particular task Parallelism is composed of simple processing units, these units are nothing but computational elements called neurons or nodes that have a property

Neuron, as it stores scientific knowledge and experimental information to make it available to the user, by adjusting it Weights. [25]

Neural networks are a system for processing data in a structural manner similar to natural neural networks and nets artificial ANNs that contain a number of simple processing units called neurons or neurons, each nerve cell (neuron) contains an external function called activation.

The algorithm follows a set of successive and iterative stages, as it is divided into two main stages, the network training stage On one set of data and the second phase testing the network on other data [7]

## Ensemble Learning

An ensemble of classiﬁers is a set of classiﬁers whose individual predictions are combined in some way (typically by voting) to classify new examples. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classiﬁers. The attraction that this topic exerts on machine learning researchers are based on the premise that ensembles are often much more accurate than the individual classiﬁers that make them up. Most of the research on classiﬁer ensembles is concerned with generating ensembles by using a single learning algorithm, such as decision tree learning or different classiﬁers are generated by manipulating the training set neural network training. (as done in boosting or bagging), manipulating the input features, manipulating the output targets or injecting randomness in the learning algorithm. The generated classiﬁers are then typically combined by majority or weighted voting.

Another approach is to generate classiﬁers by applying different learning algorithms (with heterogeneous model representations) to a single dataset. More complicated methods for combining classiﬁers are typically used in this setting. Stacking is often used to learn a combining method in addition to the ensemble of classiﬁers. Voting is then used as a baseline method for combining classiﬁers against which the learned combiners are compared. Typically, much better performance is achieved by stacking as compared to voting Some of the advanced ensemble classifiers are: Stacking, Blending, Bagging and Boosting

Stacking: Stacking is a method where a single training dataset is given to multiple models and trained. The training set is further divided using k-fold validation and the resultant model is formed. Here each model indicates a different algorithm used. The predictions made from these M models are used as predictors for the final model. The variables thus collectively formed are used to predict the final classification with more accuracy than each base model

**Blending:** Blending is a similar technique compared to stacking but the only difference being the dataset is directly divided into training and validation instead of k-fold validation.

**Bagging:** In this method, n samplings of training data are generated by picking various data items from the training data with replacement.in bagging, the items in sampling are chosen randomly as the data is unweighted. for every iteration, a base model is created on each of these samplings. Then the models run in parallel and are independent of each other and the final predictions are determined by combining the predictions from all the models, example of bagging algorithms are Bagging meta-estimator and Random forest

**Boosting:** Boosting is a self-learning technique. It learns by assigning a weight for various items in the data. The boosting technique initially starts with equal weights but after every model, each model is assigned a weight based on its performance. example of boosting algorithms: are AdaBoost, GBM, XGBM, Light GBM and CatBoost.[26]

## 2.1.15 Lexical Based Approach

Lexicon Based techniques work on an assumption that the collective polarity of a sentence or documents is the sum of polarities of the individual phrases or words. This method is based on emotional research for sentiment analysis dictionaries for each domain. Next, each domain dictionary was replenished with appraisal words of appropriate training collection that have the highest weight, calculated by the method of RF (Relevance Frequency). The word-modifier (increases or decreases) the weight of the following appraisal word by a certain percentage. Word-negation shifts the weight of the following appraisal word by a certain offset: for positive words to decrease, for negative to increase. The procedure of the text sentiment classification was carried out as follows. First weights of all training texts the classified text is calculated. All the texts are placed into a one dimensional emotional space. The proportion of deletions was determined by the cross-validation method. Then the average weights of training texts for each sentiment class were found. The classified text was referred to the class which was located closer in the one-dimensional emotional space [3]

## 2.1.16 Anaconda

It is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. As an Anaconda, Inc. product, it is also known as Anaconda Distribution or Anaconda Individual Edition, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, both of which are not free.

Package versions in Anaconda are managed by the package management system conda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for things other than Python. There is also a small, bootstrap version of Anaconda called Miniconda, which includes only conda, Python, the packages they depend on, and a small number of other packages.

## 2.1.17 Jupyter Notebook Application

Jupyter Notebook (formerly IPython Notebook) is a web-based interactive computational environment for creating notebook documents. Jupyter Notebook is built using several open-source libraries, including IPython, ZeroMQ, Tornado, jQuery, Bootstrap, and MathJax. A Jupyter Notebook application is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Github Flavored Markdown), mathematics, plots and rich media.

Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathethe 1980s. Jupyter interest overtook the popularity of the Mathematica notebook interface in early 2018.

# **2.2 Literature Review**

## 2.2.1 Introduction

In the past few years, social media platforms have become the most popular for individuals to share their views and experiences towards different services and products. Therefore, it attracts many researchers to use it as a reference for opinion extraction and sentiment analysis. Most of the previous research studies in this field use machine learning-based and lexical-based approaches to classify the emotional states of English-language tweets. Limited research work has been done to determine opinion trends for tweets in other languages ​​such as Arabic. In addition, machine learning and deep learning approaches have recently achieved remarkable results compared to traditional methods of analyzing a huge amount of data as is the case with social networking data. In the department, we seek to shed light on these researches and studies.

## 2.2.2 Literature Review

1. In 2022, the researcher Abdel Hamid Ahmed from Sebha University conducted a research entitled Sentiment Analysis Using Systems Based on Machine Learning. The study aims to analyze the feelings of users of Libyan telecommunications companies from the social networking site (Twitter) using systems based on Machine Learning, the final results showed that the SVM was the best algorithm on the Libyana data set by percentage accuracy of 81.15% and the Naive Bayes was the best algorithm on a data set of Al-Madar Company with an accuracy of 81.19%, and on the data set of the Libya Telecom and Technology Company, the superiority of the decision tree algorithm was with an accuracy of 79.16%. A sample of Libyana dataset has been published as a web application on the Internet[29]

2. In 2021, PhD student Islam Saif and Dr. Al-Samani built a model for sentiment analysis in the Sudanese dialect based on 1048 comments of the Internet service from the Sudani Telecom Company in Facebook, and a lexicon was created that contains 1000 feelings of positive and negative words. Their study aims to evaluate how pre-processing steps using natural language processing techniques affect model accuracy. They relied on testing the proposed model on two different classifiers, SVM and NAIVE BAYES(NB). The results showed that the pre-processing steps with lemmatization increased the accuracy of sentiment classification, and the SVM classifier excelled as it achieved the best measurement accuracy of 68.6%and NB of 63.1%.[30]

3. In 2019, Huda Gamal presented her PhD research at Sudan University of Science and Technology titled Sentiment Analysis of Arabic Tweets written in Sudanese dialect to classify sentiment and compare current machine learning techniques applied to sentiment analysis to data collected from Twitter's social network. A dictionary of the Sudanese dialect has been built

From 2500 feelings. The results of the first experiment showed that the automated support algorithm achieved the best accuracy, retrieval, and measurement accuracy of the test, which were 76.5%, 95.1%, and 84.4%, respectively. While the NAIVE BAYES algorithm achieved the best accuracy with 85.1%, As the results showed, the second experiment showed that the support algorithm achieved the best accuracy and test measurement accuracy of 75.2% and 83.9%, respectively. Meanwhile, the NAIVE BAYES algorithm achieved the best accuracy of 88.41%. The best recovery achieved through the decision tree algorithm is 99.9%. In addition, the percentage of positive and negative opinions about the Sudanese government services was calculated showing that 9.4% represent positive opinions towards government services, while 90.6% represent negative opinions.[14]

4. In 2017, Malak Talib and Mirsad Hadzkadic presented a study on sentiment analysis in Arabic tweets. This study aims to review the challenges of building a sentiment analysis models for the Arabic language. The results showed that there is still a great need to gain an extensive understanding of Arabic dialects, and the study indicated the need to build and disseminate Additional dictionaries in Arabic with different types and different dialects [31]

5. In 2016, Afrah Mutawakil and Malak Adel conducted research to obtain a bachelor's degree. Their study aims to retrieve Sudanese tweets from the social networking site (Twitter) to extract the most frequently discussed words in tweets and the most discussed topics and classify them into (political, religious, sports, entertainment, etc.) using Assembly algorithms in artificial intelligence achieved the best measurement accuracy of 36.41% for religious views, 27.20% for political views, 13.03% for entertainment views, 11.51% for sports opinions, and 11.85% for others.[32]

6. In 2020, research was presented at Al Hassan University and Suleiman University on sentiment analysis of Moroccan tweets based on the social networking site (Twitter) through experiments conducted on four classification methods, as well as polarity classification into positive and negative through different machine learning algorithms to highlight extraction techniques to reach optimal settings. Different machine learning results showed similarity in the accuracy provided by both classification algorithms, and the highest accuracy was achieved using the NAIVE BAYES model. [33]

7. In 2019, Rehab Hisham, Judy Mohieddine, and Zitouni Abdel Hafeez published a study on the analysis of opinions on newspaper comments in Algeria. The NAIVE BAYES, SVM, and KNN algorithms were used, which explained the effect of stemming on the results. The results also showed that the KNN algorithm gives high accuracy, while SVM is the most dominant[34]

8. In 2013, the student Izzat Muhammad Al-Samadi from the International Islamic University for Science and Education presented research to sentiment analysis in both Arabic and English, based on the available data from social networking sites, which amounted to 4050 comments in both Arabic and English, where three dictionaries were created from Arabic and English and symbols of feelings in order to extract Emotions and determining their polarity, the results showed that the NAIVE BAYS algorithm gives the best results, and that the social sentiment analysis tool is more effective than Twendz. [35]

.9In 2020, Dr. Salwa Ahmed Mohamed presented a study entitled Analyzing the Sentiments of Twitter Tweets During the US Presidential Elections 2020 Using the framework of big data to detect positive and negative sentiments of users' tweets using new analytical tools for analyzing big data, the results showed that sentiment analysis using Twitter data can be an accurate and low-cost way to measure public opinion towards candidates and predict election results[36]

10. A review in 2019 by Moez Bin Haj Hamida and Prof. Omaima Oueslati at the University of Tunis Al-Manar presented a review of Arabic sentiment research to detail problems related to feelings and study the challenges of the Arabic language scale. They applied supervised machine learning algorithms to the user's social network. [37] Table (2-2) presented the summarize of the above studies:

## 2.2.3 Table Summarizing Literature Review

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Study Objectives** | **Dataset** | **Methodology** | **The Obtained Result** | **Our opinion** |
| 1. Sentiment Analysis Using Systems Based on Machine Learning[26] | analyze the feelings of users using systems based on Machine Learning | 8560 Comments | NB, SVM,  KNN | NAIVE BASE, SVM, and KNN algorithms achieved high accuracy on different data sets, and the model was published as a web application on the Internet. | The researcher made an estimated effort on the Libyan dialect, where he used several classification algorithms on the data set in the Libyan dialect. It must also be done on the rest of the Arabic dialects. |
| 2. Application of sentiment analysis to comments in the Sudanese dialect [27] | Evaluate how processing steps and libraries affect legitimization On the general Sudanese text | 1050 Comments | NB,SVM | Use with libraries lemitimization Improves accuracy of sentiment categorization | The researcher worked on 1050 comments in the Sudanese dialect using two classification algorithms, which opens the way for the use of additional classification algorithms on a larger set of data. |
| 3 Sentiment Analysis of Arabic Tweets Written  by Sudanese Dialect [12] | classifying emotions and comparing current machine learning techniques applied to sentiment analysis for data collected from The social network | 8268 Tweets | NB, SVM, KNN, DT | High accuracy was achieved through the NAIVE BAYES, SVM, and decision tree algorithms during the two experiments. | In his work, the researcher used data in the Sudanese dialect, some of which are general, and some of them are related to the Sudanese government from the Twitter website, creating a space for collecting data from other sites on different affairs. |
| 4. Sentiment Analysis in Arabic Tweets : The Challenges of Language Anatomy [12] | Reviewing Efforts to Build Sentiment Analysis Systems for Arabic | Applications and systems created to analyze Arabic Twitter data | NB, SVM | There is no single system until the date of the study that can deal with all Arabic dialects with high accuracy | The challenges identified by the study must be taken into account in order to build a good model. The researcher did not mention the size of the data set used |
| **Study** | **Study Objectives** | **Dataset** | **Methodology** | **The Obtained Result** | **Our opinion** |
| 5. Analyzing Opinions from Sudanese Comments [32] | Classification of Sudanese comments into religious, entertainment, political, sport and others | Sudanese tweets on Twitter | K means & Hierarchical  R language. | The percentage of frequency of vocabulary in each classification is given | The researchers collected data in the Sudanese dialect from the Twitter website, which were classified into categories based on the subject, unlike what is done in sentiment analysis |
| 6. Twitter data set Moroccan feelings [33] | Measuring the Twitter data set through experiments on four classification methods | Tweets Twitter Moroccan  12k Tweets | SVC, KNN, DT, MNB, BNB,XGB, LR, LinearSVC, RandomForest | accuracy provided by each of the classification algorithms is similar, and the highest score was achieved using a model | The researcher made an estimated effort on the Moroccan dialect in terms of the size of the data set and the number of algorithms used. It must also be done on the rest of the Arabic dialects. |
| 7.Research published in the Journal of King Saud University [34] | Developing an approach to classify comments in the Algerian Arab online press into positive and negative | Comments in the Algerian electronic press | SVM, NB , KNN | NB classifier , the best results were obtained. | The researcher did not explain the size of the data set, but a good number of algorithms were used to work on the Algerian dialect |
| 8. Sentiment Analysis in Arabic and English [35] | Extracting feelings and determining their poles (positive, negative, neutral) | Twitter, Facebook, YouTube  4050 | Social Mention, Twendz | Give the percentages for experimental accuracy to determine the polarity of each suspension | The research dealt with both the Arabic and English languages. There is a need to take into account the dialects spoken in each of the two languages. |
| 9.Analyzing the Sentiments of Twitter Tweets During the US Presidential Elections 2020 [36] | Monitoring and analyzing the feelings of Twitter tweets during the US elections to identify the positive and negative feelings of the tweets | Twitter  425442 | LDA | sentiment analysis can be an accurate and low-cost way to measure public opinion towards candidates and predict election results | The research worked on a good set of data that must be provided in the Arabic dialects |
| 10. A review of feelings research in Arabic [37] | Identify important gaps in reading and writing and suggest future directions | Twitter , Yahoo, News Sites, Facebook | SVM, NB | Experiments revealed that performance is highly dependent on the quality of emotional resources | The conclusion of the study is important to consider when proceeding to build a model |

Table 2.2 Summery of Literature Review

## 2.2.4 Summery of Literature Review

Based on previous studies, we find that despite the availability of some research on the Arabic language in general and colloquial dialects in particular, there is still a need for more research and study on analyzing feelings on the Arabic language and dialects, and due to the availability of studies from different countries, good attention is noted in this In particular, we find that the algorithms of SVM, NIVE BAYES,KNN and sometimes the Decision Tree are the most common and best algorithms for achieving results in the presented studies. The researchers indicated that more experiments should be conducted with different classification algorithms and to test the accuracy of their classification.

# **CHAPTER THREE**

# **METHODOLOGY**

**CHAPTER THREE**

**METHODOLOGY**

# **3.1 Introduction**

In this chapter, we will explain the methodology we used in this research to build a model for sentiment analysis of the Sudanese colloquial dialect on the Sudani Telecom Company's Internet service application, where we collected comments from the company's official page on Facebook and then implemented the pre-processing steps on it to be ready to extract features using (TFIDF) and (BOW) techniques, then classification was done and evaluative analysis was done.

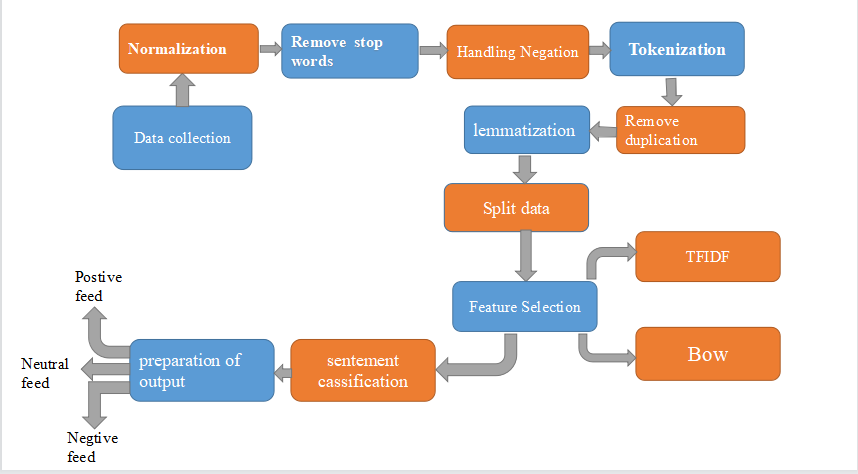


Fig 3. 1 Methodology

# **3.2 Data Collection and Reparation**

We collected 6,161 comments from the official page of the Sudanese Telecom Company on Facebook, representing the opinions of the company's customers about the various services expressed in the Sudanese colloquial dialect. The classification was done by creating Google forms and distributing them to a number of Sudanese colloquial speakers to help in the correct classification of the comments, as each form contains 50 comments, and one form is sent to two or three people, and dealing with the difference of opinions in the classification category is done by taken by the majority.

Using this method, all comments were classified based on their polarity, as in the figure (3-1) which displays examples of comments that have been classified. The data contains 1756 positive comments, 1357 neutral comments, and 3048 negative comments.

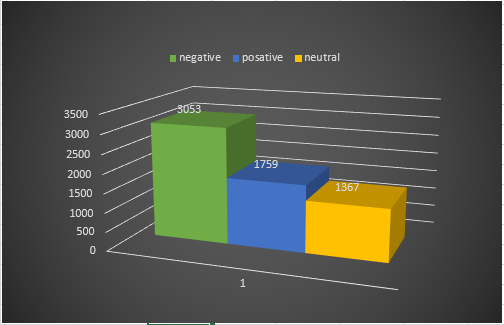
****

Fig 3. 2 Data Classes

As shown in table (3-1), most comments contain emoji, links, punctuation, repetition of letters or words, English words, numbers, diacritics, and character extensions that must be removed before the data can be used to train the model.

To deal with data use two types of files where comments are placed in a text file with the extension (txt) to facilitate handling, and apply preprocessing operations by performing one operation after another by reading from the previous file, applying the related operation, and then writing to the new file and another excel file to make it easier to deal with data as a data frame. Here are the pre-processing steps that have been implemented.

|  |  |
| --- | --- |
| **Comment** | **Label** |
| **سوداني تعبانة والله حتي لو الخدمه مجان** | **سلبي** |
| **رمز الاشتراك شنو** | **محايد** |
| **كل الجمال ومتعت الكلام مع سوداني** | **ايجابي** |

Table 3. 1 Comments Classification

# **3.3 Data preprocessing**

This step includes cleaning and normalizing the data, tokenization, removing stop words, lemmatization and removing duplicated letters and duplicated words for the purpose of preparing and unifying the data, as the different forms of one word and the presence of words that do not affect the classification of the comment in addition to extra characters and signs have a negative impact on the efficiency of the model in finding patterns.

## 3.3.1 Data Cleaning

Clean the data by removing English letters, numbers, punctuation marks, links, and emojis using the Aranorm library available on GitHub. As shown below in table (3-2):

|  |  |
| --- | --- |
| **Before** | **After** |
| **١٥٠٠جنيه ؟!!! والدولار ينزل واسعاركم تاني ما بتنزل تبا لكم اينما حللتم 😡** | **جنيه والدولار ينزل واسعاركم تاني ما بتنزل تبا لكم اينما حللتم** |

Table 3. 2 Data Cleaning

## 3.3.2 Removing Duplicated Characters

We frequently find letters and words repeated to express an exaggeration of feeling or by mistake, resulting in misleading, as the word is treated when one of its letters is duplicated as another word, such as "شديد" and "شديييييييد" and repeating the same word is useless, such as"حلو حلو “As shown below in table (3-3) So we removed the redundancy using the re library in Python.

|  |  |
| --- | --- |
| **Before** | **After** |
| **المشكلة الباقة بتخلص سرررريع بالذات الأسبوعية** | **المشكلة الباقة بتخلص سريع بالذات الأسبوعية** |

Table 3. 3 Removing Duplication

## 3.3.3 Normalization

When writing in the colloquial dialect, people do not abide by the rules of spelling, so they write the same word in a different way. The aim of this step is to unify the forms of words as much as possible using the Aranorm library available. which gives good results compared to the rest of the available libraries, as it is dedicated to working on the Arabic language, as it deals with types of pre-processing resulting from the characteristics of the language itself, such as removing tashkeel “سَيئةٌ " became “سيئة ", tatweel "شديـــــــــد" became "شديد", unifying hamzat “ أ, ئ, ؤ, إ " to "ء", correcting some spelling errors, removing underscores, punctuation marks, extra spaces, and non-Arabic characters. As shown below in table (3-4):

|  |  |
| --- | --- |
| **Before** | **After** |
| **الشركة الرائدة فى السودان سُــــــــــــودانى** | **الشركة الراءدة فى السودان سودانى** |

Table 3. 4 Normalization

## 3.3.4 Removing Stop Words

Stop words are frequently used words that do not affect positively or negatively the meaning of the sentence, but they constitute an obstacle to classification, as their presence misleads the model about words affecting the meaning of the sentence. We used the NLTK library, which provides lists of stop words for many languages, including Arabic, to remove stop words such as " من, على, في " from the comments, word ""ما was excluded because it expresses negation in the Sudanese dialect and thus affects the meaning of the sentence. As shown below in table (3-5):

|  |  |
| --- | --- |
| **Before** | **After** |
| **طيب ما تشرحو طريقه الاشتراك في الباقه** | **طيب ما تشرحو طريقه الاشتراك الباقه** |

Table 3. 5 Removing Stop Words

## 3.3.5 Handling Negation

Negation affects the polarity of sentences, as the presence of one of the negotiation letters in a negative phrase may completely change its classification into a positive word, and vice versa. There are several words used to express negation in the Arabic language, and sometimes negation is done in a scientific way. We handle negation of word “ma” Which is commonly used for negation in the Sudanese dialect by merging it with the next word so that the model can recognize it as a completely different word, which improves the accuracy of the model in recognizing polarity. The bad side of this method is that sometimes the word “ma” exists without meaning negation but it is treated in the same way. As shown below in table (3-6):

|  |  |
| --- | --- |
| **Before** | **After** |
| **الشبكة ما سيئة خاصة الصباح بدري** | **الشبكة ماسيئة خاصة الصباح بدري** |
| **طيب ما تشرحو طريقه الاشتراك الباقه** | **طيب ماتشرحو طريقه الاشتراك الباقه** |

Table 3. 6 Handling negation

## 3.3.6 Tokenization

In order for model to deal with texts, the text must be divided into units, or what is known as token. The unit can represent one word (unigram), two words (bigram), or three words (trigram), where the white space is considered as the end of a word as is customary in the Arabic language. The steps to train Moodle are based on this step that was implemented using the NLTK library in Python. As shown below in table (3-7):

|  |  |  |
| --- | --- | --- |
| **Token** | **Before** | **After** |
| **unigram** | **شكرا التوضيح مفيد اكرر الشكر سوداني الابداع والتميز** | **شكرا , التوضيح , مفيد , اكرر , الشكر , سوداني , الابداع , والتميز** |
| **bigram** | **شكرا التوضيح مفيد اكرر الشكر سوداني الابداع والتميز** | **شكرا التوضيح , التوضيح مفيد , مفيد اكرر , اكرر الشكر , الشكر سوداني , سوداني الابداع , الابداع والتميز** |
| **trigram** | **شكرا التوضيح مفيد اكرر الشكر سوداني الابداع والتميز** | **شكرا التوضيح مفيد , التوضيح مفيد اكرر, مفيد اكرر الشكر, اكرر الشكر سوداني , الشكر سوداني الابداع , سوداني الابداع والتميز** |

Table 3. 7 Tokenization

## 3.3.7 Lemmatization

The words were returned to their similar origins verbally, where words such as

"يرجع ,يسترجع ,استرجع ,استرجاع ,مرجع", became "رجع" through this step, the different forms of the word are dealt with as one word, as this represents one of the challenges specific to the Arabic language. Among the number of libraries available through which lemmatization can be performed, there are a few that deal with the Arabic language. We used the Stanza library, which provided good results, it is good with words from the classical Arabic language compared to the nltk library, but it gave bad results with the Sudanese dialect, as the same word returns as it is or often gives a wrong result, which negatively affected the accuracy of the model, in addition to the need to repeat the normalization after this step to remove the tashkeel resulting from the use of the library as shown below in table (3-8):

|  |  |  |
| --- | --- | --- |
| **Before** | **After lemmatization** | **After normalization** |
| **شكرا التوضيح مفيد اكرر الشكر سوداني الابداع والتميز** | **شُكر تَوضِيح مُفِيد أَكرَر شُكر سُودَانِيّ إِبدَاع وتميز** | **شكر توضيح مفيد اكرر شكر سوداني ءبداع وتميز** |

Table 3. 8 Lemmatization

## 3.4 Feature Extraction

In this research, we implemented feature extraction using TF\_IDF and BOW and compared the effect of each of them on classification quality

## 3.4.1 Term Frequency Inverse Document Frequency(TF-IDF)

We used TFIDF vectorizer from the Sklearn library to extract the features, TFIDF was extracted for any feature, where TF represents the frequency of the feature at the level of one comment and IDF represents the frequency of the feature at the level of all comments. As the importance of the feature increases, its frequency in one comment increases, as but doesn’t its frequency in all comments, so the weight for each feature was calculated by multiplying TF by IDF.

TFIDF was calculated once using a unigram, where the feature stands for one-word ending in a white space, and again using a bigram, where the feature stands for two words, and a third time using a trigram, where it stands for three words. Figure (3-2) shows the tfidf value for each word in a list containing three comments



Fig 3. 3 TFIDF

## 3.4.2 Bag of Word(BOW)

The features were extracted by count vectorizer from the Sklearn library, and then the BOW value was calculated for any feature, which represents the number of times the feature is repeated in the comment and shows the importance and impact of the feature on the comment classification with a high BOW value for it.

Figure (3-3) shows the BOW value for each word in a list containing three comments



Fig 3. 4 BOW

# **3.5 Data Splitting**

To learn the model, we need to divide the data into training data that constitutes the largest proportion of the data and has been categorized into (positive, negative, neutral) by which the model will identify the patterns that define feedback classification and test data that is used to test the accuracy of the model's learned classification of the data.

Since the classification itself was based on two types of classification, binary classification and multiple classification, we find that the division of data differs according to the type of classification used, the division ratio is fixed and is 80% for training and 20% for testing, in the triple classification in which the division is done into three poles ( positive, negative, neutral) Based on the percentages shown, the data was divided into a first group for training representing 80%, which is equivalent to 4928 comments out of the total comments, including 1405 positive comments, 1085 neutral comments, and 2438 negative comments. As show below in Figure (3-4)

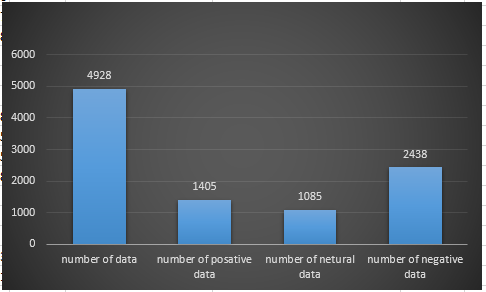


Fig 3.5 Multiple Classification Training Data

And a second group for the test represents 20%, which is equivalent to 1233 comments out of the total comments, including 351 positive comments, 272 neutral comments, and 610 negative comments. As show below in Figure (3-5)

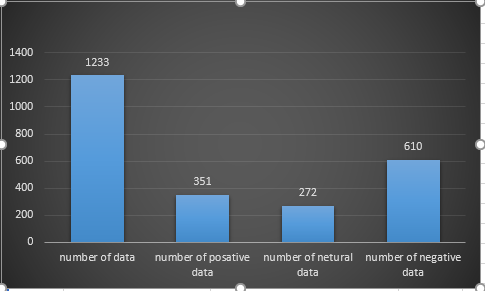


Fig 3. 6 Multiple Classification Testing Data

This is with regard to the multiple classification. As for the binary classification, which was carried out in two stages, and in which the division took place in the same proportions, in the first time all neutral comments were excluded, so the data became equal to 4804 comments, which is the sum of positive and negative comments, of which 1756 are positive comments and 3084 are negative comments. As show below in Figure (3-6)



Fig 3. 7 Binary Classification Data

It was divided into training data representing 80%, which is equivalent to 3843 comments, including 1405 positive comments and 2438 negative comments. As show below in Figure (3-7)



Fig 3. 8 Binary Classification Data

And test data represent 20%, which is equivalent to 961 comments, including 351 positive comments and 610 negative comments. As show below in Figure (3-8)



Fig 3. 9 Binary Classification Testing Data

And in the second stage, in the binary classifier, all the positive and negative comments were classified as non-neutral, and the classification was done on the two poles (neutral, not neutral), which means that the classification was done on all 6161 data, which was divided into 80% training data, which is what Equivalent to 4928 comments, including 1085 neutral comments and 3843 non-neutral. As show below in Figure (3-9)

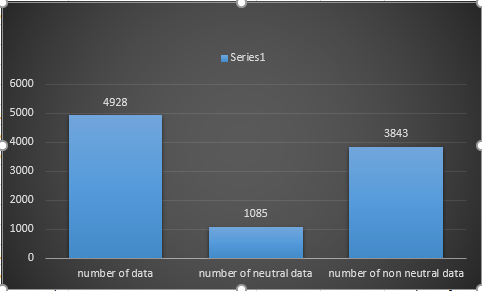


Fig 3. 10 Binary Classification Training Data

comments and test data represent 20%, which is equivalent to 1233 comments, of which 272 were neutral comments and 961 not neutral. As show below in Figure (3-10)

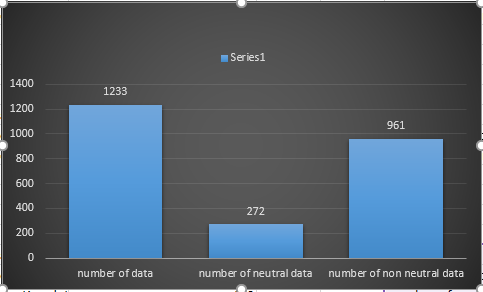


Fig 3. 11 Binary Classification Testing Data

# **3.6 Classification Model**

# **3.6.1 Binary Classification Model**

In this model, we adopted the classification into two classifiers only each time

In the first step, the classification was performed on the positive and negative classifiers, where we excluded the neutral comments by deleting the comments with a neutral rating from the data frame and marking the model on the data with a positive or negative classification only, and measuring the accuracy on that.

In the second step, the classification was conducted on the classifiers, neutral and non-neutral, where both positive and negative comments were classified as non-neutral, and the model was taught and accuracy was measured on that. shown below in table (3-9):

|  |  |  |
| --- | --- | --- |
| **Classification Categories** | | |
| **Step 1** | **Positive** | **Negative** |
| **Step 2** | **Neutral** | **Non-neutral** |

Table 3. 9 Binary Classification

# **3.7 Multi Classification Model**

In this model, different classification algorithms are applied to the data classified into (positive, negative and neutral) and the model is trained to classify the comment into one of the three categories, and then test the accuracy of its classification on the test data. shown below in table (3-10):

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification Categories** | **Positive** | **Negative** | **Neutral** |

Table 3. 10 multi classification

# **3.8 Performance Analysis**

At this point the researchers measured the accuracy of the model with the help of the most common accuracy measures

## 3.8.1 ****Confusion Matrix****

The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.

## 3.8.2 Accuracy

Which is measured by the following equation:

## 3.8.3 ****PRECISION****

Which is measured by the following equation:

## 3.8.4 ****RECALL****

Which is measured by the following equation:

## 3.8.5 F1 SCORE

Which is measured by the following equation:

F1-Score

**3.9 Model Training**

Generally, SVMs and neural networks tend to perform much better when dealing with multidimensions and continuous features. On the other hand, logic-based systems tend to perform better when dealing with discrete/categorical features. For neural network models and SVMs, a large sample size is required in order to achieve its maximum prediction accuracy whereas NB may need a relatively small dataset. There is general agreement that k-NN is very sensitive to irrelevant features: this characteristic can be explained by the way the algorithm works. Moreover, the presence of irrelevant features can make neural network training very inefficient, even impractical. Most decision tree algorithms cannot perform well with problems that require diagonal

partitioning. The division of the instance space is orthogonal to the axis of one variable and parallel to all other axes. Therefore, the resulting regions after partitioning are all hyper rectangles. The SVMs perform well when multi-collinearity is present and a nonlinear relationship exists between the input and output features. Naive Bayes (NB) requires little storage space during both the training and classification stages: the strict minimum is the memory needed to store the prior and conditional probabilities. The basic kNN algorithm uses a great deal of storage space for the training phase, and its execution space is at least as big as its training space. On the contrary, for all non-lazy learners, execution space is usually much smaller than training space, since the resulting classifier is usually a highly condensed summary of the data. Moreover, Naive Bayes and the kNN can be easily used as incremental learners whereas rule algorithms cannot. Naive Bayes is naturally robust to missing values since these are simply ignored in computing probabilities and hence have no impact on the final decision. On the contrary, kNN and neural networks require complete records to do their work. Finally, Decision Trees and NB generally have different operational profiles, when one is very accurate the other is not and vice versa. On the contrary, decision trees and rule classifiers have a similar operational profile. SVM have also a similar operational profile. No single learning algorithm can uniformly outperform other algorithms over all datasets. Different data sets with different kind of variables and the number of instances determine the type of algorithm that will perform well. There is no single learning algorithm that will outperform other algorithms based on all data sets according to no free lunch theorem.

At this stage, different machine learning techniques were used in several aspects. We chose the SVC, GaussianNB, DecisionTreeClassifier, logistic regression, MLP classifier, Multinomial, AdaBoostClassifier, GradientBoostingClassfier, Voting Classifier, and random forest classifier to work on as they represent the most common classification algorithms in previous studies. As for how to deal with them and use them in the best way with the best justification, this was done in several stages as we first divided the previously classified data into two groups. The first is the training group, which represents 80%; the second is the testing group, which represents 20%. Then the ten classifiers were used as triple classifiers from the Sklearn library by entering the training set for learning after extracting the features using the unigram TFIDF once, using the bigram TFIDF a second time, using the trigram TFIDF a third time, and using the BOW a fourth and final time., then testing its accuracy on the test data using four measures (accuracy, precision, F1 Score, recall) and recording the results. These steps were done on the data directly before any of the pre-processing steps were performed To make comparisons and seek the effect of pre-processing on the accuracy of the model, the whole process was repeated after performing the pre-processing steps.

Then we calculated the results of the binary classification model of the algorithms by excluding the comments with a neutral label and applying all the classification steps on the positive and negative classification and recording the results as a first stage, and in the second stage, we modified the database for each of the positive and negative comments by classifying them as non-neutral and keeping the neutral comments. And conduct classification steps on the neutral and non-neutral classification and record the results.

# **CHAPTER FOUR**

# **IMPLEMENTATION**

**CHAPTER FOUR**

**IMPLEMENTATION**

# **4.1 Introduction**

In this chapter, we will present the methods that we followed in implementing the previously described methodology step by step, as we relied on the use of the Python programming language, as it is rich in useful libraries in the field of data analysis in general and in the field of sentiment analysis in particular, and because of the simplicity of the language and ease of implementation, which was attractive to many researchers Which constituted a reliable resource for us in previous research and studies.

# **4.2 Import Libraries**

In this step, all libraries that were used to build the model were imported from Python programming language is very rich in libraries specially built to deal with machine learning issues or useful multipurpose libraries. At this stage, all the libraries that will be used in the different coding stages have been called, including “Pandas” library, which provides functions for reading data from an Excel file and a wide range of functions to deal with data in the form of a data frame, which guarantees greater flexibility and control, we also find that the “NumPy” library has been called, which offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. Then we import the “re” library it is used to work with Regular Expressions It is often needed when preprocessing text comments, also we have “sklearn”, It is an important library that has many applications, as it is used in classification, clustering, regression, dimensionality reduction, preprocessing and model selection

NLTK is one of the libraries that must be used as well, because it is specialized in natural language processing for a number of languages, including the Arabic language, so it provides some resources such as lists of stop words and some functions such as tokenization and as an alternative in some cases, such as lemmas, we used the “Stanza” library, which provides almost the same functionality with better performance with the Arabic language.

Also, we call “normalize\_arabic\_text” function from “aranorm” The accompanying file, which represents a library provided by “Faris Alasmary” on GitHub, although the library is not available as Python Module yet, but it is useful for pre-processing Arabic texts.



Fig 4. 1 imported libraries

# **Data Preprocessing**

In order to prepare the data for classification, the pre-processing process was carried out in several steps, including normalization, then removing stop words, then dealing with negation, then tokenization, duplication, and lemmatization the following is a detailed explanation of each of these steps.

## 4.3.1 Normalization

By reading the comments one by one and applying the functions inside “normalize\_arabic\_text” function, each of which receives text as input and returns a modified text, these functions start with a function that deletes all non-Arabic letters, then treats some spelling errors, then replaces all hamzat with hamzat on the line, and finally two functions to remove “tashkeel” and “tatweel”.

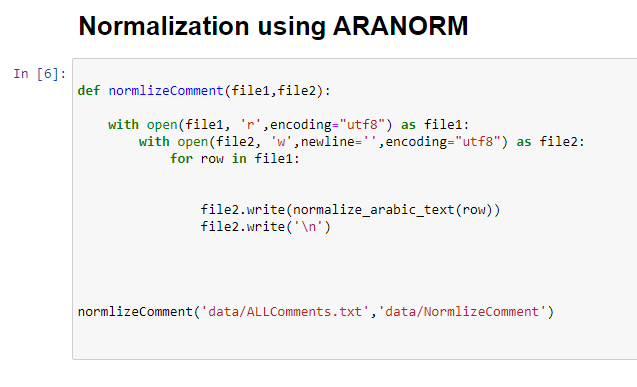


Fig 4. 2 Normalization

## 4.3.2 Remove Stop Words

NLTK Library includes in its list of stopping words in the Arabic language 754 words, for example ['it', 'me', 'me', 'ah', 'aha', 'or', 'awla', 'those', 'oh', 'i'] which means ['إنه', 'أنى', 'أنى', 'آه', 'آها', 'أو', 'أولاء', 'أولئك', 'أوه', 'آي'] These words have been removed from all comments that imply any of them, except for the word "ma", which expresses negation.



Fig 4. 3 Remove Stop Words

4.3.3 Handling Negation

To deal with negation in the Sudanese dialect, the most common word “ma” was adopted by merging the word “ma” with what follows it so that their presence together is recognized as a

new word, with the addition of a special condition for the presence of the word “ma” as the last word in the sentence to avoid error.



Fig 4. 4 Handling Negation

## 4.3.4 Tokenization:

At this stage, Word Tokenize from NLTK library was used to divide the single comment into several units called tokens. The unit represents one word in uni tokenization, two words in binary tokenization, and three words in triple tokenization, where the white space expresses the end of the word. The output of this step is A text array for each comment in which the element represents a token. It is saved as text in a text file for display and preview.



Fig 4. 5 Tokenization

## 4.3.5 Remove Duplication

In this step, the consecutive duplicate characters were removed using Re and Itertools libraries, which are useful for iterating over all elements through two different functions, one for

characters and the second for words. They were applied to the comments one by one and placed in a new file for review and preview. It is worth noting here that Each of the two functions was transferred from one of the appendices attached to a study from the previous studies mentioned above.



Fig 4. 6 Remove Duplication

**4.3.6 Lemmatization**

Using the stanza library to implement lemmatization for each word in each comment, where we found that using the nltk library is useless, as the lemmatization returns the words as they are without change, but the stanza library and while it gives good results for the Arabic language, it cannot find the root of the words from The Sudanese dialect, and it was necessary to remove the "Tashkeel" resulting from lemmatization using the aranorm library, as the "Tashkeel" loses this process, its main purpose, which is to standardize the shape of the data



Fig 4. 7 Lemmatization

# **4.4 Classification Techniques**

To classify the data, multiple classification can be used in which classification algorithms are used on the data and it is classified into more than two different classifications, three in our case. Or a binary classifier can be used in which classification algorithms are used on the data and it is classified into two classifications only

## 4.4.1 Multiple Classification

In this method, the data is read and classified into three classifiers (positive, negative, and neutral) and then different classification algorithms are applied to it.

Read Data File: convert the excel file that contains the data to a data frame using the Pandas library

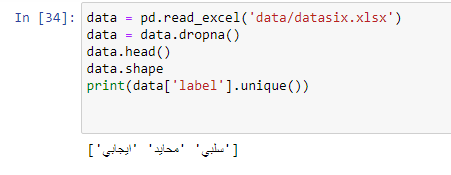


Fig 4. 8 Read Data File

* show the data frame that represents the data

****

Fig 4. 9 Data in Three Classifier Model

## 4.4.2 Binary Classification

In the first stage, data with a neutral classification are excluded and work is done on the classifiers (positive, negative).

****

Fig 4. 10 Tow Classifier Model

Here we present the data after excluding the data with a neutral classification



Fig 4. 11 Data after exclude data with a neutral classification

In the second stage, the positive and negative comments were classified as non-neutral, and the neutral comments were kept as they are, where the classification will be based on two categories (neutral, non-neutral).

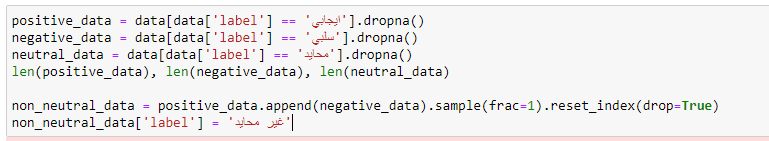
****

Fig 4. 12 Classifying positive and negative comments as non-neutral

Reviewing the data after classifying the positive and negative comments as non-neutral and keeping the neutral comments as they are

****

Fig 4. 13 Data after Classifying positive and negative comments as non-neutral

# **4.6 Data Splitting:**

Prepare the data by dividing it into training data and test data, each of which is data expressed in the form of features, and each of these data corresponds to the appropriate classification

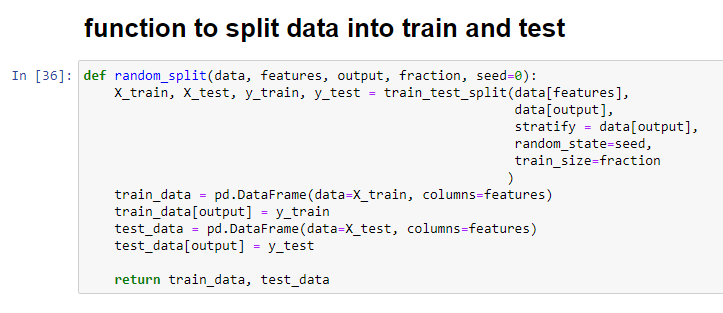


Fig 4. 14 Splitting Data

Choose the proportions of the data sections, which determines the ratio of the training data from the data set, and what remains is considered test data. The column containing the comments is also indicated, from which the features will be extracted, and the column containing the classification is indicated.

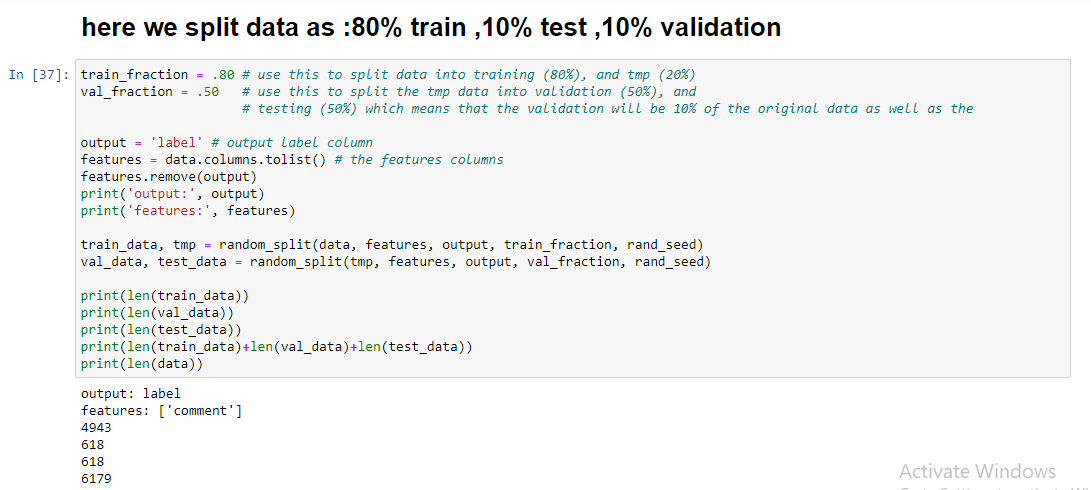


Fig 4. 15 Data Partition Ratios

# **4.7 Feature Extraction**

Since the data cannot be processed in its text form, it is expressed by extracting features using two methods

## 4.7.1 TFIDF

Extract features as tfidf using TfidfVectorizer from the sklearn library

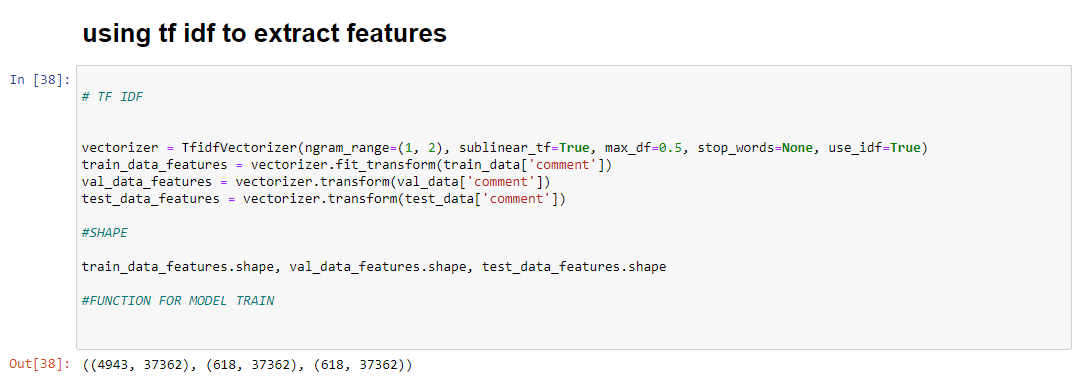


Fig 4. 16 Implement TFIDF

**4.7.2 BOW**

Extract features as bow using CountVectorizer from the sklearn library

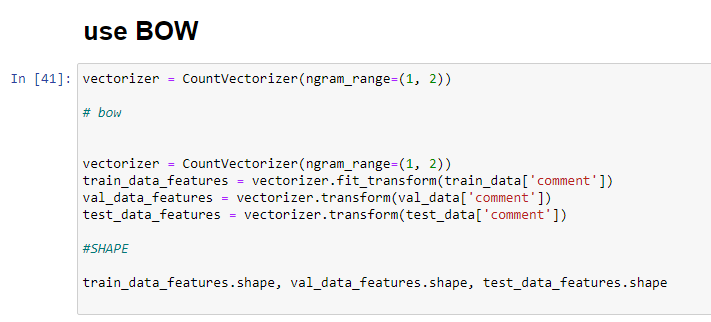


Fig 4. 17 Implement BOW

# **4.8 Train model**

This function shown in figure () represents the basis for the training process, where the classifier and data are passed to the function, which performs the whole process. This abstract method was adopted, taking into account the great similarity between the different classifiers due to its simplicity and ease of implementation, where the researchers were able to experiment with a larger number of classifiers on the data as it is. However, this method is flawed because it does not give the classifiers their full right of customization, as this may be attributed to the low accuracy of the classifier. In general, the function receives, in addition to the classifier, each of the training data in the form of features and their classifications, and test data in the form of features as well, their classifications and data The test is complete, then you train the classifier on the data using the fit function, the training data is passed to it in the form of features and their classifications, and then the model is tested via the predict function, test data is passed to it in the form of features, and the output is compared with the test data classifications to measure accuracy via classification\_report

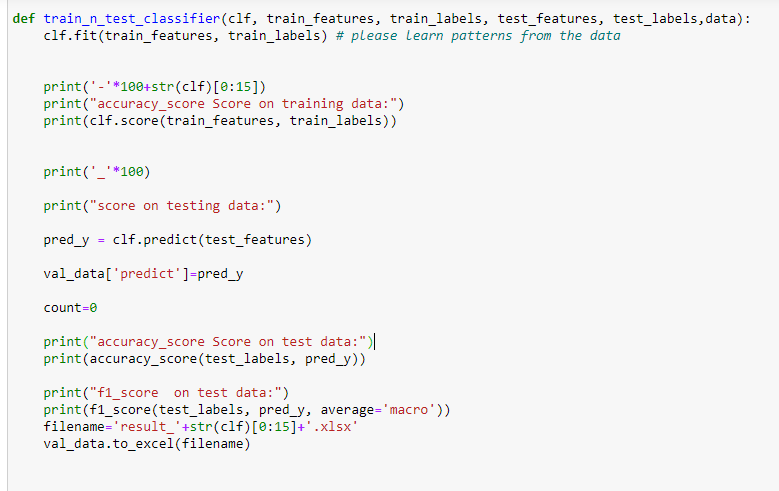


Fig 4. 18 Function that trains the model

# **4.9 Classification Algorithms**

Below we explain the implementation of the different algorithms used for classification

Implementation of training function on svc from the svm model using the SKlearn library.

Implementation of training function on svc from svm model model using SKlearn library by creating instance of the model and pass it with the train data features and label and by passing argument kernel as the linear kernel to “train\_n\_classifier” method above

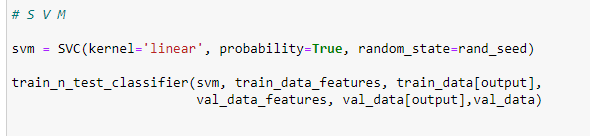


Fig 4. 19 SVC

Implementation of training function on MultinomialNB from the naïve\_bayes model using the SKlearn library by creating instance of the model and pass it with the train data features and label to “train\_n\_classifier” method above.

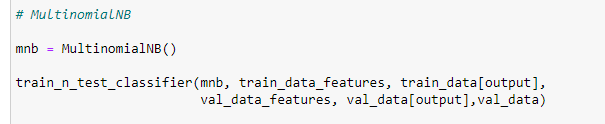


Fig 4. 20 MultinomialNB

Implementation of training function on DecisionTreeClassifier from the tree model using SKlearn library by creating instance of the model and by set random\_state parameter to 0 so that the train-test splits are always deterministic and pass it with the train data features and label to “train\_n\_classifier” method above.

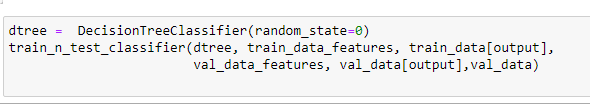


Fig 4. 21 DecisionTreeClassifier

Implementation of training function on logistic regression from the linear\_model using SKlearn library by creating instance of the model and by set random\_state parameter to 0 so that the train-test splits are always deterministic and pass it with the train data features and label to “train\_n\_classifier” method above.

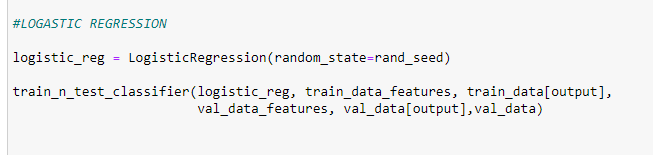


Fig 4. 22 Logistic Regression

Implementation of training function on MLPclassifier from the neural\_network model using SKlearn library by creating instance of the model and set the number of layers to 4 layers with determine 20 nodes as the number of nodes in each layer. And set verbose parameter to true to be aware of the implementation details and by set random\_state parameter to 0 then pass it with the train data features and label to “train\_n\_classifier” method above.

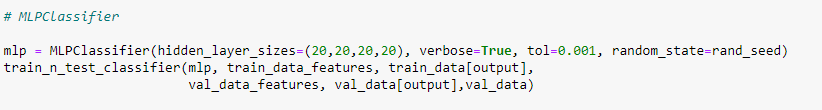


Fig 4. 23 MLPClassifier

Implementation of training function on KNeighborsClassifier from neighbors model using SKlearn library by creating instance of the model and by set n-neighbors parameter which means the tuning parameter/hyper parameter (k) to 3 and pass it with the train data features and label to “train\_n\_classifier” method above.

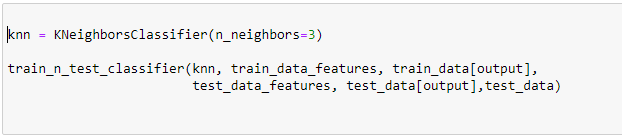


Fig 4. 24 KNN

Ensemble algorithms were also used, which improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. The following algorithms mentioned fall under this type.

Implementation of training function RandomForestClassifier from the ensemble model using the SKlearn library by creating instance of the model and by set random\_state parameter to 0 so that the train-test splits are always deterministic and set the parameter n\_estimators which express the number of decision trees to 100 and pass it with the train data features and label to “train\_n\_classifier” method above

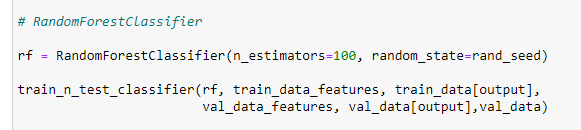


Fig 4. 25 RandomForestClassifier

Implementation of training function GradientBoosting from the ensemble model using the SKlearn library by creating instance of the model and by set random\_state parameter to 0 so that the train-test splits are always deterministic and set the parameter n\_estimators which express the number of decision trees to 100 and set max\_depth parameter which express the maximum depth of regression tree to 1, and pass it with the train data features and label to “train\_n\_classifier” method above

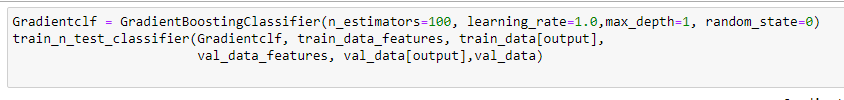


Fig 4. 26 Gradient Boosting

Implementation of training function AdaBoostClassifier from the ensemble model using the SKlearn library by creating instance of the model and by set random\_state parameter to 0 so that the train-test splits are always deterministic and set the parameter n\_estimators which express the number of decision trees to 100 and pass it with the train data features and label to “train\_n\_classifier” method above

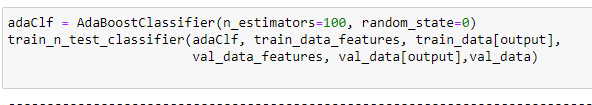


Fig 4. 27 AdaBoostClassifier

Befor Implementation of training function on votingClassifier from the ensemble model using the SKlearn library we first need to put the above six classfiers on a list and pass it as parameter to “train\_n\_classifier” method along with the train data features and label and by siting the the voting type as soft sience soft voting entails combining the probabilities of each prediction in each model and picking the prediction with the highest total probability



Fig 4. 28 Voting Classifier

# **CHAPTER FIVE**

# **RESULT AND RECOMMENDATION**

**CHAPTER FIVE**

**RESULT AND RECOMMENDATION**

# **5.1 Introduction**

In this last chapter of the research, we will review and discuss the results and identify the cases that recorded the best accuracy

# **5.2 Results**

To choose the best model, we implemented a number of machine learning algorithms from the Sklearn library on the data in all possible cases based on the choices that could be made in the previous stages as shown below attached are tables and pictures explaining these experiments, numbered by the trial number, followed by the case number in the relevant experiment

**Experimental 1: Study the effect of using binary classification on the performance of the classification model**

This experiment was conducted on the data to be classified into two categories at a time

Where the data was first entered into the binary classification model (positive, negative) by excluding the neutral comments after performing the complete pre-processing and feature extraction using TFIDF and recording the accuracy of the model in the classification of the test data as shown in the table (**5-1**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **93%** |
| **MultinomialNB** | **89%** |
| **KNN** | **87%** |
| **DecisionTreeClassifier** | **83%** |
| **logistic Regression** | **88%** |
| **MLP classifier** | **93%** |
| **Random forest classifier** | **89%** |
| **Voting Classifier** | **91%** |
| **AdaBoostClassifier** | **86%** |
| **GradientBoostingClassfier** | **86%** |

Table 5. 1 Experimental (1.1)

When using the binary classification model with two classifiers (positive, negative) on the data with pre-processing steps, SVC achieved the best result with an accuracy scale of (93%) when extracting features using TFIDF

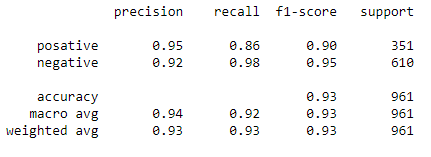


Fig 5. 1 Experimental (1.1)

Then, secondly, the data was entered on the binary classification model (positive, negative) by excluding neutral comments after performing the complete pre-treatment operations, but the features were extracted using BOW and recording the accuracy of the model in the classification of the test data to monitor the effect of using BOW on the accuracy of the model. as shown in the table (**5-2**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **91%** |
| **MultinomialNB** | **91%** |
| **KNN** | **77%** |
| **DecisionTreeClassifier** | **86%** |
| **logistic Regression** | **91%** |
| **MLP classifier** | **92%** |
| **Random forest classifier** | **89%** |
| **Voting Classifier** | **91%** |
| **AdaBoostClassifier** | **87%** |
| **GradientBoostingClassfier** | **86%** |

Table 5. 2 Experimental (1.2)

and when extracting features with BOW the MLPClassifier achieved the best result with an accuracy scale of (92%)

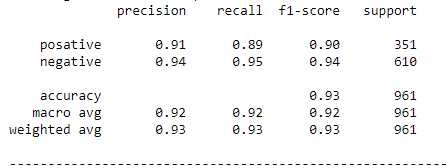


Fig 5. 2 Experimental (1.2)

In the second stage, the data were first entered into the binary classification model (neutral, non-neutral) by classifying the positive and negative comments as non-neutral after performing complete pre-processing and feature extraction using TFIDF and recording the accuracy of the model in classifying the test data as shown in the table (**5-3**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **83%** |
| **MultinomialNB** | **79%** |
| **KNN** | **80%** |
| **DecisionTreeClassifier** | **77%** |
| **logistic Regression** | **80%** |
| **MLP classifier** | **83%** |
| **Random forest classifier** | **82%** |
| **Voting Classifier** | **82%** |
| **AdaBoostClassifier** | **79%** |
| **GradientBoostingClassfier** | **78%** |

Table 5. 3 Experimental (1.3)

When using the binary classification model with two classifiers (neutral, non-neutral) on the data with the pre-processing procedure, MLPClassifier achieved the best result with an accuracy scale (83%) when extracting features using TFIDF,

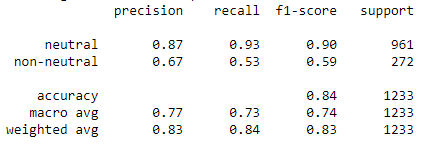


Fig 5. 3 Experimental (1.3)

Then, secondly, entering the data into the binary classification model (neutral, non-neutral) by classifying the positive and negative comments as non-neutral after performing the complete pre-processing operations, but extracting the features was done using BOW and recording the accuracy of the model in the classification of the test data to monitor the effect of using BOW on the accuracy of the model as shown in the table (**5-4**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **81%** |
| **MultinomialNB** | **81%** |
| **KNN** | **77%** |
| **DecisionTreeClassifier** | **79%** |
| **logistic Regression** | **82%** |
| **MLP classifier** | **81%** |
| **Random forest classifier** | **81%** |
| **Voting Classifier** | **83%** |
| **AdaBoostClassifier** | **80%** |
| **GradientBoostingClassfier** | **80%** |

Table 5. 4 Experimental (1.4)

while voting classifier achieved the best accuracy when extracting features by BOW with an accuracy scale (83%)

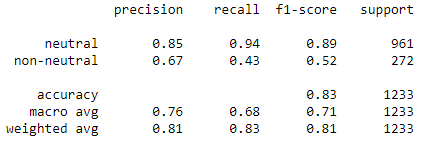


Fig 5. 4 Experimental (1.4)

**Experimental 2: Study the effect of using multi classification on the performance of the classification model**

This experiment was conducted using a three-way classification model, where the data were first entered into the multiple classification model (positive, negative, neutral) after performing complete pre-processing and feature extraction using TFIDF and recording the accuracy of the model to classify the test data as shown in the table (**5-5**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **77%** |
| **MultinomialNB** | **72%** |
| **KNN** | **68%** |
| **DecisionTreeClassifier** | **69%** |
| **logistic Regression** | **76%** |
| **MLP classifier** | **77%** |
| **Random forest classifier** | **74%** |
| **Voting Classifier** | **76%** |
| **AdaBoostClassifier** | **69%** |
| **GradientBoostingClassfier** | **70%** |

Table 5. 5 Experimental (2.1)

When using the ternary classification model on the data directly after performing the complete pre-processing, MLPClassifier achieved the best result with an accuracy scale of (77%) when extracting features using TFIDF

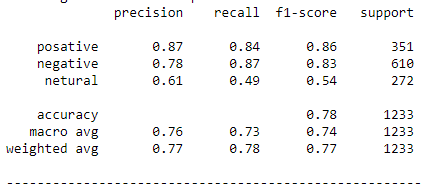


Fig 5. 5 Experimental (2.1)

Then, secondly, enter the data into the multiple classification model (positive, negative, neutral) after performing complete pre-processing and extracting features using BOW and recording the accuracy of the model to classify the test data to monitor the effect of using BOW on the accuracy of the modelas shown in the table (**5-6**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **75%** |
| **MultinomialNB** | **76%** |
| **KNN** | **52%** |
| **DecisionTreeClassifier** | **68%** |
| **logistic Regression** | **75%** |
| **MLP classifier** | **76%** |
| **Random forest classifier** | **73%** |
| **Voting Classifier** | **76%** |
| **AdaBoostClassifier** | **67%** |
| **GradientBoostingClassfier** | **71%** |

Table 5. 6 Experimental (2.2)

and when extracting features by BOW Voting Classifier achieved the best result with an accuracy scale of (76%).

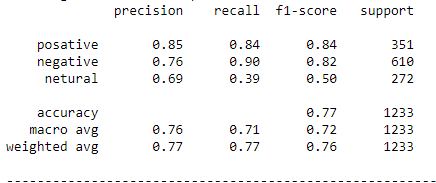


Fig 5. 6 Experimental (2.2)

**Experimental 3: Study the effect of preprocessing on the performance of the classification model**

This experiment was done by conducting the classification directly on the data without performing the pre-processing steps to give figures showing the effect of the pre-processing on the binary classification and the multiple classification.

Where the multiple classification model was first used by entering data on the multiple classification model (positive, negative, neutral) directly without prior processing and extracting features using TFIDF The first time and record the accuracy of the model to classify the test data as shown in the table (**5-7**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **76%** |
| **MultinomialNB** | **73%** |
| **KNN** | **70%** |
| **DecisionTreeClassifier** | **66%** |
| **logistic Regression** | **75%** |
| **MLP classifier** | **75%** |
| **Random forest classifier** | **74%** |
| **Voting Classifier** | **76%** |
| **AdaBoostClassifier** | **67%** |
| **GradientBoostingClassfier** | **68%** |

Table 5. 7 Experimental (3.1)

When using the ternary classification model on the data directly without pre-processing, SVC achieved the best result with an accuracy scale of (76%) when extracting features using TFIDF,

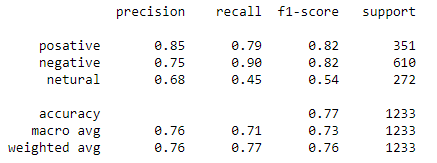


Fig 5. 7 Experimental (3.1)

Then input the data on the ternary classification model (positive, negative, neutral) directly without pre-processing and extracting the features but using BOW this time and recording the accuracy of the model to classify the test data in order to observe the effect of using BOW on the accuracy of the model as shown in the table (**5-8**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **74%** |
| **MultinomialNB** | **75%** |
| **KNN** | **55%** |
| **DecisionTreeClassifier** | **68%** |
| **logistic Regression** | **76%** |
| **MLP classifier** | **76%** |
| **Random forest classifier** | **73%** |
| **Voting Classifier** | **75%** |
| **AdaBoostClassifier** | **68%** |
| **GradientBoostingClassfier** | **71%** |

Table 5. 8 Experimental (3.2)

while MLP Classifier represented the best model when extracting features by BOW with an accuracy scale of (76%).

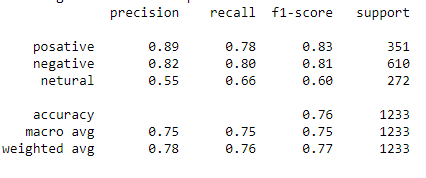


Fig 5. 8 Experimental (3.2)

Then the binary classification model was used secondly by entering data on the binary classification model (positive, negative) in the first stage by excluding neutral comments without performing pre-processing operations and extracting features using TFIDF in the first time and recording the accuracy of the model in classifying the test data as shown in the table (**5-9**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **84%** |
| **MultinomialNB** | **79%** |
| **KNN** | **81%** |
| **DecisionTreeClassifier** | **76%** |
| **logistic Regression** | **81%** |
| **MLP classifier** | **84%** |
| **Random forest classifier** | **82%** |
| **Voting Classifier** | **82%** |
| **AdaBoostClassifier** | **81%** |
| **GradientBoostingClassfier** | **79%** |

Table 5. 9 Table Experimental (3.3)

When using the binary classification model with two classifiers (positive, negative) on the data directly without pre-processing, the MLPClassifier achieved the best result with an accuracy scale of (84%) when extracting features using TFIDF

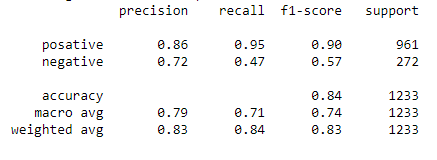


Fig 5. 9 Experimental (3.3)

Then input the data on the binary classification model (positive, negative) by excluding the neutral feedback without performing the pre-processing and feature extraction operations but using BOW this time and recording the accuracy of the model to classify the test data in order to note the effect of using BOW on the accuracy of the model as shown in the table (**5-10**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **89%** |
| **MultinomialNB** | **90%** |
| **KNN** | **62%** |
| **DecisionTreeClassifier** | **84%** |
| **logistic Regression** | **88%** |
| **MLP classifier** | **90%** |
| **Random forest classifier** | **87%** |
| **Voting Classifier** | **90%** |
| **AdaBoostClassifier** | **84%** |
| **GradientBoostingClassfier** | **83%** |

Table 5. 10 Experimental (3.4)

and also MLPClassifier achieved the best result when extracting features by BOW with an accuracy scale of (90%)

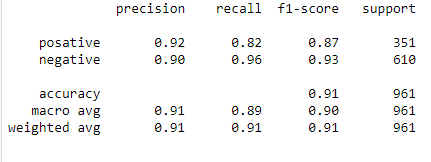


Fig 5. 10 Experimental (3.4)

And in the second stage, the data were entered into the binary classification model (neutral, non-neutral) by classifying each of the negative and positive comments as non-neutral without pre-processing and extracting features using TFIDF and recording the accuracy of the model in classifying the test data as shown in the table (**5-11**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **84%** |
| **MultinomialNB** | **79%** |
| **KNN** | **80%** |
| **DecisionTreeClassifier** | **79%** |
| **logistic Regression** | **81%** |
| **MLP classifier** | **84%** |
| **Random forest classifier** | **83%** |
| **Voting Classifier** | **84%** |
| **AdaBoostClassifier** | **81%** |
| **GradientBoostingClassfier** | **78%** |

Table 5. 11 Experimental (3.5)

When using the binary classification model with two classifiers (neutral, non-neutral) on the data without pre-processing, MLPClassifier achieved the best result with an accuracy scale of (84%) when extracting features using TFIDF

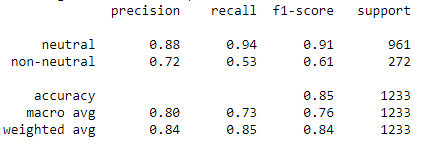


Fig 5. 11 Experimental (3.5)

Then the data were entered into the binary classification model (neutral, not neutral) by classifying both positive and negative comments as non-neutral without pre-processing and feature extraction but using BOW this time and recording the accuracy of the model to classify the test data in order to note the effect of using BOW on the accuracy Sample as shown in the table (**5-12**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **82%** |
| **MultinomialNB** | **83%** |
| **KNN** | **78%** |
| **DecisionTreeClassifier** | **80%** |
| **logistic Regression** | **83%** |
| **MLP classifier** | **81%** |
| **Random forest classifier** | **83%** |
| **Voting Classifier** | **84%** |
| **AdaBoostClassifier** | **82%** |
| **GradientBoostingClassfier** | **81%** |

Table 5. 12 Experimental (3.6)

while the Voting Classifier achieved the best when extracting features by means of BOW with an accuracy scale of (84%)

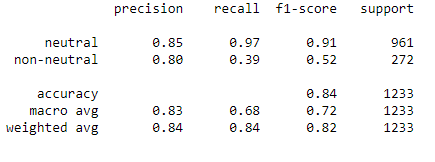


Fig 5. 12 Experimental (3.6)

**Experimental 4: Study the effect of lemmatization on the performance of the classification model**

This experiment was conducted by conducting pre-processing operations on the data, except for the lemmatization, to clarify its effect on the accuracy of the model for both multiple classification and binary classification.

Where the multiple classification model was first used by entering the data on the multiple classification model (positive, negative, neutral) after performing the pre-processing operations, except for lemmatization, extracting the features using TFIDF in the first time and recording the accuracy of the model to classify the test data as shown in the table (**5-13**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **77%** |
| **MultinomialNB** | **72%** |
| **KNN** | **62%** |
| **DecisionTreeClassifier** | **68%** |
| **logistic Regression** | **75%** |
| **MLP classifier** | **74%** |
| **Random forest classifier** | **74%** |
| **Voting Classifier** | **77%** |
| **AdaBoostClassifier** | **67%** |
| **GradientBoostingClassfier** | **70%** |

Table 5. 13 Experimental (4.1)

When using the ternary classification model on the data directly after the pre-processing, except for lemmatization Voting Classifier achieved the best result with an accuracy scale (77%) when extracting features using TFIDF

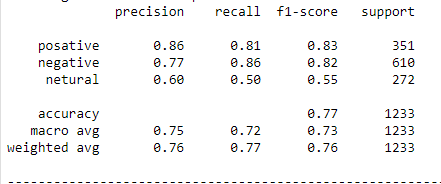


Fig 5. 13 Experimental (4.1)

Then enter the data on the ternary classification model (positive, negative, neutral) after performing the pre-processing operations, excluding lemmatization, extracting the features but using BOW this time and recording the accuracy of the model to classify the test data in order to note the effect of using curvature on the accuracy of the model as shown in the table (**5-14**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **75%** |
| **MultinomialNB** | **76%** |
| **KNN** | **52%** |
| **DecisionTreeClassifier** | **68%** |
| **logistic Regression** | **75%** |
| **MLP classifier** | **76%** |
| **Random forest classifier** | **73%** |
| **Voting Classifier** | **77%** |
| **AdaBoostClassifier** | **67%** |
| **GradientBoostingClassfier** | **71%** |

Table 5. 14 Experimental (4.2)

and also Voting Classifier achieved the best result when extracting features by BOW with an accuracy scale (77%).

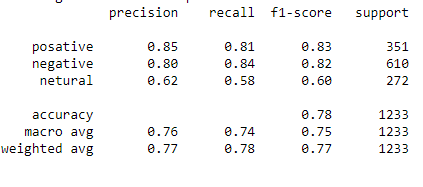


Fig 5. 14 Experimental (4.2)

Then the binary classification model was secondly used by inputting the data on the binary classification model (positive, negative) in the first stage by excluding the neutral comment after performing the pre-processing operations, excluding the lemmatization extracting the features using TFIDF in the first time and recording the accuracy of the model in the classification data the test as shown in the table (**5-15**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **91%** |
| **MultinomialNB** | **90%** |
| **KNN** | **87%** |
| **DecisionTreeClassifier** | **82%** |
| **logistic Regression** | **87%** |
| **MLP classifier** | **91%** |
| **Random forest classifier** | **88%** |
| **Voting Classifier** | **90%** |
| **AdaBoostClassifier** | **85%** |
| **GradientBoostingClassfier** | **83%** |

Table 5. 15 Experimental (4.3)

When using the binary classification model with two classifiers (positive, negative) on the data with pre-processing steps except for lemmatization, MLPClassifier achieved the best result with an accuracy scale (91%) when extracting features using TFIDF,

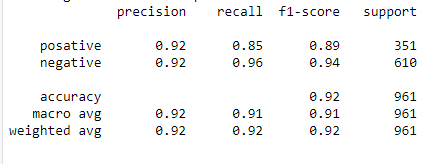


Fig 5. 15 Experimental (4.3)

Then enter the data on the binary classification model (positive, negative) by excluding the neutral comments after performing the preprocessing operations, excluding lemmatization, extract the features but using BOW this time and record the accuracy of the model to classify the test data in order to note the effect of using BOW on the accuracy of the model as shown in the table (**5-16**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **91%** |
| **MultinomialNB** | **92%** |
| **KNN** | **65%** |
| **DecisionTreeClassifier** | **87%** |
| **logistic Regression** | **90%** |
| **MLP classifier** | **92%** |
| **Random forest classifier** | **88%** |
| **Voting Classifier** | **90%** |
| **AdaBoostClassifier** | **85%** |
| **GradientBoostingClassfier** | **85%** |

Table 5. 16 Experimental (4.4)

And also achieved the best model when extracting features by BOW with an accuracy scale (92%)

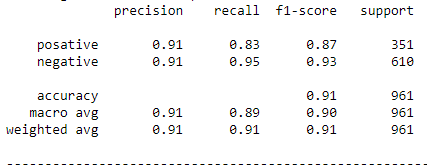


Fig 5. 16 Experimental (4.4)

In the second stage, the data were entered into the binary classification model (neutral, non-neutral by classifying the positive and negative comments as non-neutral after performing pre-processing operations, excluding lemmatization, extracting features using TFIDF and recording the accuracy of the model. in the classification of the test data. as shown in the table (**5-17**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **83%** |
| **MultinomialNB** | **79%** |
| **KNN** | **73%** |
| **DecisionTreeClassifier** | **78%** |
| **logistic Regression** | **79%** |
| **MLP classifier** | **81%** |
| **Random forest classifier** | **80%** |
| **Voting Classifier** | **83%** |
| **AdaBoostClassifier** | **79%** |
| **GradientBoostingClassfier** | **78%** |

Table 5. 17 Experimental (4.5)

When using the binary classification model with two classifiers (neutral, non-neutral) on the data with the pre-processing procedure except for Lemmatization, voting classifier achieved the best result with an accuracy scale (83%) when extracting features using TFIDF

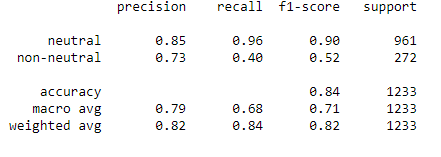


Fig 5. 17 Experimental (4.5)

The data was then entered into the binary rating model (neutral, not neutral) by classifying the positive and negative comments as non-neutral after performing preprocessing operations, excluding lemmatization, extracting the features but using BOW this time and recording the accuracy of the model for the test classification in order to note the effect of using BOW on sample accuracy as shown in the table (**5-18**) below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **SVC** | **82%** |
| **MultinomialNB** | **81%** |
| **KNN** | **57%** |
| **DecisionTreeClassifier** | **77%** |
| **logistic Regression** | **82%** |
| **MLP classifier** | **77%** |
| **Random forest classifier** | **79%** |
| **Voting Classifier** | **82%** |
| **AdaBoostClassifier** | **80%** |
| **GradientBoostingClassfier** | **79%** |

Table 5. 18 Experimental (4.6)

and also when extracting features by BOW Voting classifier achieved the best result with an accuracy scale (82%)

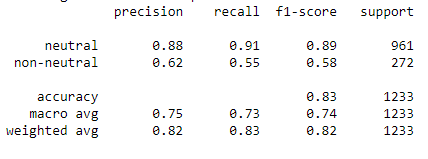


Fig 5. 18 Experimental (4.6)

# **5.3 Model Evaluation**

By observing the above experimental, we found that the most classifications that achieve high accuracy are MLPClassifier and Voting classifier. Based on the experiments shown, we find that it is better to build a binary model, as it works on two classifiers each time, in order to clarify and identify weaknesses and work more on areas Shortcomings We also find that the pre-processing improves the accuracy of the model, except for the lemmatization that affect it negatively. where both the SVM and MLPClassifier models achieved the best accuracy when extract features using TF IDF and classifying into two classifiers (positive and negative) with a score of 93%, while the MLPClassifier model had the best accuracy when extract features using BOW and classifying into two classifiers (neutral and non-neutral) of 92%.

**5.4 Conclusion**

The field of sentiment analysis arose from the need of companies and institutions to analyze customer opinions, which are often not written in an official and unified way, but rather by using the dialect, which sometimes makes it impossible to conduct statistical analyzes using traditional methods. There is still a great need to apply the field to languages ​​other than English, especially on local dialects.

The researchers focused on finding the best model for applying machine learning algorithms to the data. Among the available options, the data is represented in more than 6,000 comments in the Sudanese dialect that were collected from the official page of the Sudanese Telecom Company on Facebook, expressing customer opinions about the services. They were collected manually and classified according to polarity to (positive, negative, neutral) It was necessary first to carry out pre-processing steps for the data to remove excesses that could negatively affect the accuracy of the model. Within this step, lemmatization was carried out using the stanza library, whose negative effect was observed against the expected, and then this was followed by the feature extraction process using TF IDF and BOW techniques.

To teach the model, the researchers selected ten commonly used classifiers from the sklearn library (SVC, MultinomialNB, GaussianNB, DecisionTreeClassifier, logistic Regression, MLP classifier, AdaBoostClassifier, Voting Classifier, Random forest classifier**).**

They were trained once as a three-class model on the data, which are classified for three classifiers (positive, negative, and neutral), and the second time as a two-class model in the first stage by excluding neutral comments and keeping the classifiers (positive, negative) and in the second stage by classifying each of the positive and negative comments as non-neutral and the classification on two poles (neutral, non-neutral) then record and compare all results.

# **5.5 Recommendations**

The following are the researchers' suggestions for developing this work:

• Implementation of the model as a website that allows instant classification of data

• Adopting the classification for more multiple poles.

• Building a pre-processing library on the Sudanese dialect

• Deal better with negation and sarcasm.

# **References**

[1] D. Wilding, P. Fray, S. Molitorisz, and E. Mckewon, “The Impact of Digital Platforms on News and Journalistic Content,” *Cent. Media Transit.*, p. 175, 2018, [Online]. Available: https://www.accc.gov.au/system/files/ACCC commissioned report - The impact of digital platforms on news and journalistic content%2C Centre for Media Transition %282%29.pdf

[2] R. I. Mukhamediev *et al.*, “Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges,” *Mathematics*, vol. 10, no. 15, pp. 1–25, 2022, doi: 10.3390/math10152552.

[3] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.

[4] S. Bhatia, M. Sharma, and K. K. Bhatia, “Sentiment Analysis and Mining of Opinions,” *Stud. Big Data*, vol. 30, no. May, pp. 503–523, 2018, doi: 10.1007/978-3-319-60435-0\_20.

[5] “Liu et al., 2009 1,” pp. 1–3, 2009.

[6] Peter D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 417–424, Morristown, NJ, USA, 2001. Ass,” *Society*, vol. 12, no. 3, pp. 417–424, 2002, [Online]. Available: http://www.google.com

[7] M. D. Devika, C. Sunitha, and A. Ganesh, “Sentiment Analysis: A Comparative Study on Different Approaches,” *Procedia Comput. Sci.*, vol. 87, no. June, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[8] L. Palen, S. Vieweg, S. B. Liu, and A. L. Hughes, “Crisis in a networked world: Features of computer-mediated communication in the April 16, 2007, Virginia Tech event,” *Soc. Sci. Comput. Rev.*, vol. 27, no. 4, pp. 467–480, 2009, doi: 10.1177/0894439309332302.

[9] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on Twitter,” *Proc. 20th Int. Conf. Companion World Wide Web, WWW 2011*, pp. 675–684, 2011, doi: 10.1145/1963405.1963500.

[10] D. Yates and S. Paquette, “Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake,” *Int. J. Inf. Manage.*, vol. 31, no. 1, pp. 6–13, 2011, doi: 10.1016/j.ijinfomgt.2010.10.001.

[11] “Wasserman\_Faust\_1994.pdf.”

[12] F. Cheong and C. Cheong, “Social media data mining: A social network analysis of tweets during the Australian 2010-2011 floods,” *PACIS 2011 - 15th Pacific Asia Conf. Inf. Syst. Qual. Res. Pacific*, no. December 2013, 2011.

[13] S. P. Borgatti and P. C. Foster, “The network paradigm in organizational research: A review and typology,” *J. Manage.*, vol. 29, no. 6, pp. 991–1013, 2003, doi: 10.1016/S0149-2063(03)00087-4.

[14] H. J. Abdelhameed and S. Muñoz- Hernández, “Sentiment Analysis of Arabic Tweets in Sudanese Dialect,” *Int. J. New Technol. Res.*, vol. 5, no. 6, pp. 17–22, 2019, doi: 10.31871/ijntr.5.6.20.

[15] Z. H. Kilimci, “Evaluation of Customer Satisfaction about Telecom Operators in Turkey by Analyzing Sentiments of Customers through Twitter,” *Data Sci. Appl.*, vol. 3, no. 2, pp. 15–20, 2020, [Online]. Available: https://www.researchgate.net/publication/348098304

[16] M. E. M. Abo, N. A. K. Shah, V. Balakrishnan, M. Kamal, A. Abdelaziz, and K. Haruna, “SSA-SDA: Subjectivity and sentiment analysis of sudanese dialect Arabic,” *2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019*, no. April, 2019, doi: 10.1109/ICCISci.2019.8716466.

[17] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, “Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts,” *J. Biomed. Inform.*, vol. 62, no. January 2017, pp. 148–158, 2016, doi: 10.1016/j.jbi.2016.06.007.

[18] M. Annett and G. Kondrak, “A comparison of sentiment analysis techniques: Polarizing movie blogs,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5032 LNAI, no. June, pp. 25–35, 2008, doi: 10.1007/978-3-540-68825-9\_3.

[19] S. Kiritchenko, X. Zhu, and S. M. Mohammad, “Sentiment analysis of short informal texts,” *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, 2014, doi: 10.1613/jair.4272.

[20] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[21] P. Borele and D. A. Borikar, “An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques,” *IOSR J. Comput. Eng.*, vol. 18, no. 2, pp. 64–69, 2016, doi: 10.9790/0661-1802056469.

[22] M. Montes-y-Gómez, A. Gelbukh, and A. López-López, “Detecting deviations in text collections: An approach using conceptual graphs,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2313, pp. 176–184, 2002, doi: 10.1007/3-540-46016-0\_19.

[23] “12.words (Apte, Damerau, & ´ Weiss, 1994; Dumais, Platt, Heckerman, & Sahami, 1998; Lewis, 1992.pdf.”

[24] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, “SemEval-2016 task 4: Sentiment analysis in twitter,” *SemEval 2016 - 10th Int. Work. Semant. Eval. Proc.*, pp. 1–18, 2016, doi: 10.18653/v1/s16-1001.

[25] P. Mehta and S. Pandya, “A review on sentiment analysis methodologies, practices and applications,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 601–609, 2020.

[26] C. Giraud-carrier, R. Vilalta, and P. Brazdil, “Is Combining Classifiers with Stacking Better than Selecting the Best One ?,” pp. 255–273, 2004.

[27] “Vapnik, 1998.pdf.”

[28] B. Mahony, “Reasoning about semantic Web in Isabelle / HOL Reasoning about Semantic Web in Isabelle / HOL Organization of Australia,” no. January 2004, 2014, doi: 10.1109/APSEC.2004.82.

[29] S. Analysis, U. Machine, S. Libyan, and T. Companies, “ايبيل ةلود,” 2022.

[30] I. Saif *et al.*, “Applying Sentiment Analysis on Arabic comments in Sudanese Dialect,” vol. 8, no. 3, pp. 69–73, 2020.

[31] C. Stephanidis, “Foreword,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10280 LNCS, p. VII, 2017, doi: 10.1007/978-3-319-58562-8.

[32] O. Mining, F. The, and S. Tweets, “خ ّٞ اد ٘ ظىا Opinion Mining For The,” 2016.

[33] S. Mihi, B. A. BEN Ali, I. EL Bazi, S. Arezki, and N. Laachfoubi, “MSTD: Moroccan sentiment twitter dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 363–372, 2020, doi: 10.14569/IJACSA.2020.0111045.

[34] H. Rahab, A. Zitouni, and M. Djoudi, “SANA: Sentiment analysis on newspapers comments in Algeria,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 7, pp. 899–907, 2021, doi: 10.1016/j.jksuci.2019.04.012.

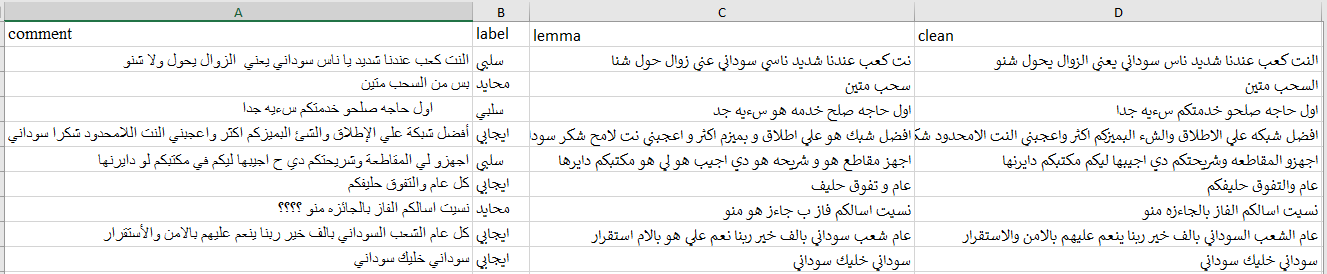
[35] I. M. Alsmadi, M. Al-kabi, M. Al-kabi, T.-A.- Jordan, and M. Dabour, “Arabic / English Sentiment Analysis : An Empirical Study Arabic / English Sentiment Analysis : An Empirical Study”.

[36] S. Aboeleila, “0202 ةيكيرملأا ة ي سائرلا تاباختنلا ا ءاهثأ تريوت تاديرغت رعاشم ليلتح ةمخضلا تانايبلا راطإ مادختساب,” no. July, 2022.

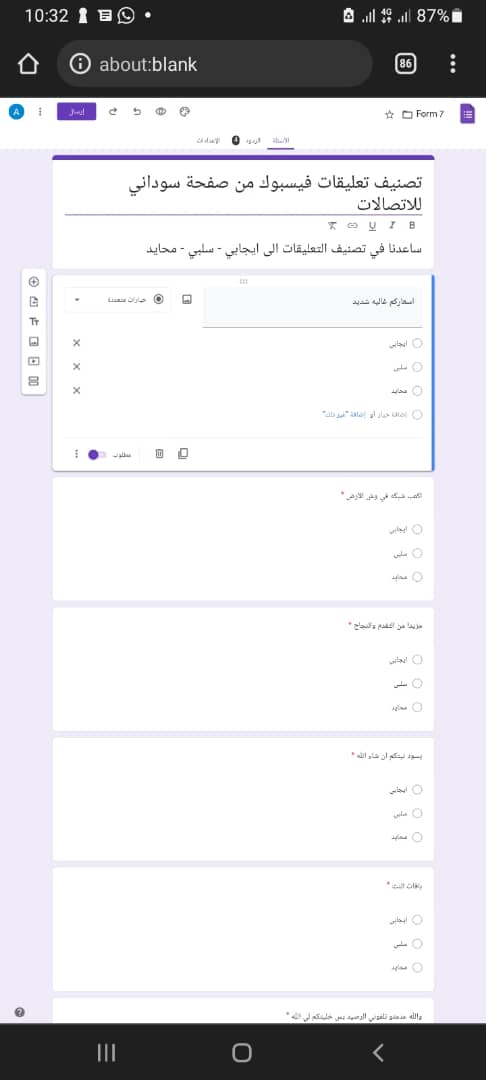
[37] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, “A review of sentiment analysis research in Arabic language,” *Futur. Gener. Comput. Syst.*, vol. 112, pp. 408–430, 2020, doi: 10.1016/j.future.2020.05.034.

# **Appendices**

## Appendix (I) SAMPLE OF DATASET



## Appendix (II) Classification Comments using Google Form



## Appendix (III) Jupyter Notebook

