Categorizing Egyptian Arabic Text into Topics of Interest for Supporting E-services Quality

By

Ahmed Tarek Youssef Ahmed

Bachelor Thesis

Submitted to the Department of Business Informatics

At the Faculty of Management Technology

German University in Cairo

Student registration number/ ID: 49-4345

Date: 6th of June 2023

Supervisor: Dr. Ayman Alserafi

Table of Contents

Abbreviation List

AI	Artificial Intelligence
ANLP	Arabic Natural Language Processing
ANN	Artificial Neural Network
CHAID	Chi-square Automatic Interaction Detector
CRISP-DM	Cross-industry standard process for data mining
IoT	Internet of Things
KNN	K-nearest Neighbour Algorithm
LSTM	Long Short-term Memory
MSA	Modern Standard Arabic
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Tool-kit
NLU	Natural Language Understanding
NN	Neural Networks
UGC	User Generated Content
WEKA	Waikato Environment for Knowledge Analysis

List of Figures

List of Tables

1 Introduction

Nowadays, the categorization of Egyptian Arabic text into Topics of Interest for Supporting E-services Quality has gained very high importance. With the noticeable growth of online content and the increasing reliance on e-services worldwide, understanding the sentiments expressed in Egyptian Arabic language has become prominent for many reasons, like improving customer satisfaction and relationship management and improving recommendation systems based on customer sentiment, and improving marketing campaigns by understanding the exact customers' needs and wants. The motivation behind this thesis is the need of the Arabic world to have an accurate sentiment analysis model that is tailored specifically for Egyptian Arabic text. To clarify, most of the research and sentiment analysis models built are built for other languages (mainly English). This thesis aims to develop an accurate sentiment analysis model capable of accurately classifying 100,000 Arabic movie and book reviews into positive and negative sentiments. This thesis's outline is as follows, firstly we have some theoretical considerations in which we will discuss what is artificial intelligence as well as its different types and formats, then we are going to discuss what is data mining and some of its steps like data pre-processing as well as multiple modelling techniques like linear regression, Naive Bayes, decision trees, clustering, and association rules. After that, we are going to discuss what is Natural Language Processing and its different goals like sentiment analysis, named entity recognition, and text summarization. Also, we are going to discuss what is Arabic Natural Language Processing and its state of the art. Then, we are going to discuss different NLP vendors worldwide, and different real-life applications of NLP like voice applications, voice bots, understanding customer feedback, and speech-to-text applications. After that, we are going to state what is the research gap that our thesis aims to fill, and we are going to discuss how did we use the CRISP-DM model to fill the research gap by building an efficient and effective Arabic sentiment analysis model, then we are going to discuss the results of our model.

2 Theoretical Considerations

Firstly, we are going to have an overview of the previously published works of the topics of our interests, there are 4 main topics which will be discussed and studied. 1) Artificial intelligence. 2) Data mining. 3) Natural language processing. 4) General natural language processing applications.

2.1 Artificial Intelligence Definition

According to Kaplan and Haenlein (2020: 39).Artificial intelligence is defined in this study as the ability to understand and comprehend external data accurately, also the ability to learn from these data and how to utilize these data and use them to attain predefined goals flexibly. This study also states that it can be quite challenging to find an accurate definition of what artificial intelligence is for 3 main reasons. 1) It is not an easy task to clearly define what is human intelligence, so it is more challenging to define the intelligence of computers. 2) Once a computer achieves a given task, we stop defining this ability as being an intelligent one, hence the term computer/artificial intelligence is always changing and moving from complex tasks to more complex tasks and so on. 3) Artificial intelligence has several branches and categories like analytical, human-inspired, and humanized AI these classifications depend on the AI's abilities whether it is cognitive, emotional, or social competencies so with these several classifications, it can be challenging to define what artificial intelligence actually is.

The same authors of the previous paragraph, Kaplan and Haenlein (2019: 17-18) have a similar definition in another research paper which states that an AI program would be successful if it achieves a task that a human would do he/she would be called intelligent. In other words, artificial intelligence is the science of programming computers to do tasks that needs human intelligence to be achieved. AI could use information/data that can be an output of IoT or big data models as an input and use it in its complex algorithms to produce useful outputs. This study classifies AI into 3 main categories. 1) Analytical AI which mainly deals with cognitive abilities. These AI approaches create a cognitive representation of the inputs and make future judgments by learning from past experiences (like fraud detection in financial services). 2) Human-Inspired: both emotional and cognitive intelligence are present in the AI model. These systems can understand human

sentiments and cognitive elements and take them into account in the decision-making process. 3). Humanized AI can deal with most types of competencies (i.e., cognitive, emotional, and social intelligence).

Turner (2018: 1-2) defines AI in a different way which is that artificial intelligence is one of a kind technology because it can execute actions and take decisions solely without being programmed in detail to do these specific actions. In other words, artificial intelligence is a non-natural entity that has the ability to decide and choose what should be executed autonomously, they are complex algorithms that can comprehend inputs and choose the best possible outcome by themselves without being programmed in detail how it should react to all situations and inputs.

Jarek and Mazurek (2019: 47) explains that is important for AI to mimic the human's intelligence. In more details, they stated that AI is a technology that is derived from information technology algorithms, it is also related to other concepts like automation and robotization. Some people may confuse artificial intelligence with machine learning but they are quite different, Oxford dictionary defines artificial intelligence as the creation of computer algorithms and programs that can carry out tasks that would typically need human intelligence, like speech recognition, visual perception, decision-making, and language translation. The main point of using artificial intelligence is to mimic human thinking and brain functions to make logical problem solving with higher accuracy and less time.

2.1.1 Types of Artificial Intelligence

There are three types of artificial intelligence that will be discussed in detail. 1) Machine learning: is applied to make the model forecast and estimate outputs accurately without being explicitly programmed to produce these outputs. 2) Deep learning: is used to produce logical outputs in a way similar to the way of learning/thinking of the human brain. . 3) Neural networks: these are inter joined hierarchical and layered nodes that are used by computers to learn from their past mistakes, this will allow the model to increase its accuracy and produce more reliable outputs, more details will be discussed below.

2.1.1.1 Machine Learning

Ray (2019: 35) states that machine learning is when the model can gain experience from its past mistakes. To clarify, a machine learning model can have higher accuracy each time more, compared to the last time when it is used in attaining tasks. In other words, machine learning can make forecasts and estimations based on historical data without being programmed in detail on how to do so. A real-life example of how machine learning is used is using it to investigate and study cancer possibilities for a given client, based on his medical records and reports. The more the machine studies these reports, the more accurate it will be in future investigations.

Ray (2019: 36-37) also states multiple techniques of how machine learning is applied, some of these techniques are linear regression: which is a supervised algorithm that is used to forecast missing continuous and numerical dependent variables. For example, predicting the future sales value of a certain product. Linear regression is used when the model has one dependent variable and one independent variable, which in real life is not practical as most of the time the dependent variable depends on multiple independent variables, so this is when Multivariable regression is used. Another technique is logistic regression, which is used to classify outputs in a binary format. For example, using logistic regression to predict if this given email is spam or not, so it is a binary output. Lastly, decision tree can also be used to classify data depending on its features.

2.1.1.2 Deep Learning

Chen and Ran (2019: 2) as well as Ranganathan (2021: 66-68) defined deep learning as a multi-layered model that is used to mimic the human brain behavior of how a brain can receive inputs, and process them to produce logically accepted outputs, in a deep learning model, input is going through several layers in a predefined order, each layer process its input which it took from the previous layers, and passes its output to the following layer until the last layer produce the wanted final output of the whole model. Using deep learning and its way of simulating human brains helps in achieving higher accuracy in less operational time. Any given deep learning model requires an input just like any human brain, these inputs can be numerical data, voice, digital signals, sensor outputs, etc. Deep learning can be used in classification, prediction, and estimation models, so it

is a very important algorithm to be used as it can achieve multiple types of goals and needs.

2.1.1.3 Neural Networks

According to Walczak (2019: 632-633) neural network is a technique that consisted of layers of imaginary circles called neuron that is connected with each other. This network tries to mimic the biological neural network in the human system. However, it uses less number of materials to do so. ANN is a computerized model of how the human brain and nervous system operate in an electrical manner. It mimics this biological behavior in a way that an ANN model has processing elements (circles) that are connected to each other in a hierarchical form where the output of a layer is given as an input to the next layer. ANN has two major types based on the type of input given, either binary or continuous input. These two types can also be classified into another two types which are supervised and unsupervised learning. Some examples of supervised learning algorithms are Hopfield network (which takes binary input) and back propagation (which takes continuous input). A common misconception about neural network applications is that they magically identify complicated patterns in data. While the truth is that the phases involved in constructing artificial neural networks must be comprehensively understood and implemented for successful development.

2.1.2 Text Mining

Ferreira-Mello et al. (2019: 2) as well as Gunawan and Aldridge (2018: 7) have as similar definition to text mining, which is defined as the process of retrieval of useful information and knowledge from unstructured text blocks. Text mining can be used to begin a process of an automated investigation and studying of large and complicated text files in less time and with higher accuracy compared to manual investigation. There are several Text mining techniques that can be used in order to receive the desired output, like natural language processing, text classification and clustering, information retrieval, and text summarization. Text mining has many alternative names and terminologies like Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text. Having said that, text mining can be used in a wide spectrum of applications across most of the business fields, it can be used in medical, scientific, business, finance, etc.

The steps for conducting a proper text mining model according to Gunawan and Aldridge, (2018: 9-10) are 1) gathering data, there are many sources for gathering data, some of these sources are websites, emails, and text documents. 2) Pre-process: it is one of the most important steps as it deals with data cleaning, which is formatting the data in a way that is easier to deal with while applying text mining techniques, an example of that is deleting unwanted text in the input text to make the model more efficient and effective. 3) Indexing: which is the process of indexing the text and sorting it to make it easier to deal with. 4) Mining: which is the main step that deals with the actual knowledge extraction. 5) Analysis which is the assessment of the work done to know if the whole process succeeded to produce the desired output or not. These techniques are applied to produce one of many desired outputs. Some of these outputs are, 1) producing and mining valuable text/information. 2) Studying and analyzing text blocks automatically. 3) Investigating sentiments and opinions in a given text block automatically.

2.2 Data Mining Definition

According to Yang et al. (2020: 57-66), data mining is defined as a procedure that is followed in order to extract useful data/information from gigantic and unorganized databases. In other words, it is an efficient and effective procedure that is used to help in information processing. There are several types, algorithms, and forms of data mining, Association analysis: is used to detect frequent patterns and activities that can be frequently happening or associated together in many scenarios or if there is a causal relationship between them, it helps in detecting hidden connections and relationships between data from large and complicated databases, a common practical example of Association analysis is market basket analysis. Another example is Traditional regression which is a data mining technique that uses ordinary linear regression techniques to detect and predict the numeric relations between two or more variables it can be a linear regression or multiple linear regression according to how many independent variables we have. A linear regression model contains only one independent variable and only one dependent variable, while multiple linear regression is the opposite.

Manjarres et al. (2018: 236- 237) have a similar definition to data mining, and they state that data mining is a group of processes that are used on large data sources like data

warehouses and relational databases in order to foresee hidden patterns and forecast numerical or categorical measures. Many people believe that data mining is an evolution and expansion of information technology techniques and algorithms, and it became very important and prominent in the last decade due to the emergence of the internet. This study classifies data mining paradigms into two main classes verification and discovery, they also classified discovery as prediction and description, prediction can be classified into classification and regression and finally, classification can be classified into neuronal networks, Bayesian models, decision trees, and instance-based. The general data mining steps that are stated in this study are "Filtering data, Selection of variables. Extracting knowledge, interpretation and evaluation.

Teng and Gong (2018: 1-4) explained the difference between multiple data mining techniques, like Classification which is the use of a training dataset in order to produce a logical and accurate classification model that can be used on other data sets. The classification model is used to split data into several classes according to some of its characteristics. Classification algorithms have several forms and types like KNN classification algorithm, naive Bayes classification algorithm, and decision tree, and support vector machine. Another data mining task is regression which is applying numeric and statistical approaches and performing mathematical calculations between variables to produce equations and laws that are used to forecast and predict unknown variables accurately. Regression has several forms like regression tree, linear regression, and logic regression. Another task is clustering, which is grouping data with similar characteristics into a group called "cluster", so that all the data in a given cluster have common characteristics, the most popular clustering algorithm is K-means clustering.

Rygielski et al. (2020: 485-489) clarified why data mining is called with that term, as the term mining is an analogy to gold and coal mining which is extracting gold and coal from underground, similarly, data mining is extracting useful data from large data warehouses. It is divided into discovery, which is observing patterns in a given database and using these patterns to predict future values. Forensic Analysis: it is using the patterns observed from the previous step to predict and estimate future values. Forensic Analysis is applying discovered patterns on unknown data elements. There are numerous applications of data mining that are used every day in many different business and work fields like market

basket analysis, which indicates the products that clients frequently buy together. Data mining also can be used to predict the future sales value of a given product/service. It can be used in fraud detection and to prevent as many fraud risks as possible, and also it can be used in customer segmentation which helps in creating efficient and effective marketing campaigns.

2.2.1 Data Pre-processing

Sapountzi and Psannis (2020: 53-54) as well as Alasadi and Bhaya (2017: 4102) both researches summarized what is data pre-processing, which is the process of applying certain procedures to raw data to make them ready for the following stage which is producing useful outputs. Data preprocessing has a number of steps like data cleansing, integration, transformation, and dimensionality reduction. It is used to reduce random and inaccurate variations in data sets to improve the accuracy of the output of the model. Some of the main problems that can be in raw data are that there may be missing values or inaccurate values, having these problems decrease the accuracy of the model, so this is why data preprocessing is important. Data cleaning is used to deal with these missing values and to reduce the problem of duplicates and inconsistencies. Noisy data is an expression that is used when there are inaccurate data entries made by humans. To summarize, data pre processing's main goal is to try as much as possible to decrease the data set size by removing excess and unwanted data, as well as eliminating outliers so that the given data set after the preprocessing process in ready for the following data mining tasks.

Phan et al. (2021: 121) as well as Alasadi and Bhaya (2017: 4102-4104) explained how to deal with missing values, errors and outliers, sometimes there are tuples that contain missing attributes, so these missing attributes need to be dealt with to improve the model accuracy, there are multiple ways to deal with this. For example, we can ignore it if the percentage of missing values is very small, or we can fill them manually but that can be inefficient in time and effort wise, we can use a predefined variable to fill the missing values, but this decreases the model accuracy, we can use the average of all the values to fill the missing values. Also, we can use more complicated methods like linear regression which uses line of best fit to estimate the unknown variables, or clustering which each group of similar attributes is grouped together in a group called a cluster, clustering

techniques are mainly used to detect outliers, there are other steps than data cleaning, like data integration which is merging multiple databases into one main database. Data selection: this is selecting the most relevant data needed in the data mining model. Lastly, data reduction which is reducing the amount of data entering the model as much as possible by removing excess and unwanted data to make the model more efficient and effective.

2.2.2 Supervised Learning

Supervised learning data mining techniques require labeled inputs and outputs in order to perform their job. In other words, the relationship between the dependent and independent variables has to be shown clearly, we will discuss 4 types of supervised data mining techniques. Which are 1) Linear regression 2) Naive Bayes 3) Decision Tree.

2.2.2.1 Linear Regression

Maulud and Abdulazeez, (2020: 140-141) and Mostafa, (2019: 182-183) both defined linear regression as a technique that uses mathematics and with the help of relationships between variables, predicts and forecasts the missing ones. Linear regression is used to initiate a mathematical relationship between dependent and independent variables, thus by creating this relationship the estimation of values can be done. Getting into the technical part of linear regression, the independent variables are variables that are used to estimate the missing figures in the dependent variables. In the equation, the independent variable is denoted by "X" and the dependent variable is denoted by "Y". Regression models can be either linear or multiple regression depending on the number of independent variables in a given equation. Regression techniques use the least square model which is plotting the best fit line for given data points on a graph as well as trying as much as possible to minimize the total number of squares while plotting to make the model more accurate. The least square method's equation is as follows $\hat{y} = \beta 0 + \beta 1 xi$. Y is the dependent variable that needs to be forecasted, $\beta 0$ is the y-intercept (the value of Y if x is 0), and $\beta 1$ is the slope of the line. Using regression techniques makes the data mining model more accurate as it is better to impute the missing values rather than simply deleting them as if these records had never happened.

2.2.2.2 Naïve Bayes

Harahap et al. (2018: 1-3) as well as Parlina et al. (2019: 1-3) both stated that the Naive Bayes method is a technique that relies on the science of probability to impute values and produce outputs. In more detail, Naive Bayes is a method that calculates groups of probabilities by adding a variety of frequencies that is in the dataset. Naive Bayes is a supervised model that considers all variables whether they are dependent or independent and provides value to the class variable. There are many advantages of using Naive Bayes, some of these advantages are that the model does not need large data volumes in order to do accurate training, hence the estimation needed can be done accurately using small amounts of training data. In other words, the Naive Bayes Classifier algorithm is used for classifying data and that is done by forecasting the likelihood that each instance of the target attribute will occur in each given variable class. In other words, the main advantages of using Naive Bayes are 1) it is an uncomplicated model and can be easily made. 2) Can be used with enormous datasets easily.

2.2.2.3 Decision Tree

According to Charbuty and Abdulazeez (2021: 21) decision trees are supervised learning technique that depends on the shape of a tree (roots & leaves) in order to classify data and generate useful outputs, it has a predefined route that the data given goes through in order to be classified. The route begins from the root of the tree, then a number of Boolean operations happen to classify the data until we reach the desired output which is in the node. For each given node, there is a characteristic that this node hold, and there is a minimum of 2 sub-nodes that answer the Boolean question of whether this characteristic happens to appear in the data given or not, if yes then it goes to the yes sub-node, if no then the opposite, then this process is repeated until we have our desired output.

Charbuty and Abdulazeez, (2021: 22) as well as Patel and Prajapati (2018: 75) explained some of the main advantages of using decision trees, as they can be used as a classification algorithm for more than 1 data type (it can be used for categorical and numerical attributes), 2) They can be easily understood and visualized compared to other techniques. Some of its drawbacks are 1) If one early Boolean decision was falsely taken, then the final output will be so different than the desired output, and 2) it has so many layers that can make it complicated calculation-wise. There are several Decision tree algorithms,

some of them are 1) CHAID which deals with categorical data types. 2) ID3 also deals with categorical data types and uses WEKA as a tool. 3) C4.5 deals with both categorical and numerical data types and also use WEKA as a tool.

2.2.3 Unsupervised Learning

Unsupervised learning is the other type of data mining techniques, that is mainly concerned about using unlabeled data sets (there is no clear identification of dependent and independent variables) in order to uncover and hidden relationships and patterns between data in a given dataset, we will be discussing two main unsupervised learning techniques which are clustering and association rules.

2.2.3.1 Clustering

Ahmad and Khan (2019: 31883-31888) stated that clustering techniques are unsupervised learning data mining algorithms. It is used to pack unlabeled data (because it is unsupervised) together in a virtual pack called a cluster. All data in a given cluster has to have high similarity in terms of their features so that every cluster represents a certain amount of features and attributes. In addition, there has to be no similarity in attributes/features between each cluster so that each cluster alone represents a different set of features and attributes. In other words, there should be a high similarity between attributes of the same cluster, and there should be no or minimal similarity between each cluster, this will insure that the classification process is done efficiently and effectively. One of the main advantages of using clustering is that it can deal with both numeric (attributes like height and weight) as well as categorical attributes (attributes like blood group, race, and gender). Most of the clustering techniques used nowadays have to have a predefined and pre-known number of clusters needed in the algorithm before starting, the number of clusters needed in each process varies according to several aspects, and there are several algorithms and calculation that is used to estimate the number of clusters needed in each operation.

Ariza Colpas et al. (2020: 1-2) explained the 4 steps that should be carefully conducted in order to have an efficient and effective clustering model. Step 1: Select the data on which we need to apply the clustering model on, we should carefully know the attributes and features we are applying the model before starting it in order to know at the end

whether the execution was successful or not. Step 2: choosing the most suitable clustering algorithm, there are many types and forms of clustering algorithms, so we should carefully choose the best suitable one according to our data set and data types to produce the most accurate results. Step 3: Results validation, after the selected algorithm is executed, it is important to review and validate its results using the cluster quality calculations and matrix and to know the degree of cohesion and separation in the model. Step 4: Result Interpretation, after making sure that the results were relevant and accurate, it is important to visualize and interpret these findings, and that is the main concern of this step.

2.2.3.2 Association Rules

According to Yoshimura et al. (2018: 371- 372), association rules are used to discover and know hidden relationships between different data in a given data frame. One of its applications is that it is used to know products and their relationships by using market basket analysis. For example, it can show products that are frequently bought together. This is done by studying are reviewing past transactions and buying history. To sum it all up, association rules are used to uncover hidden patterns of data that are in a given transaction, the appearance of an item implies that there is a very large probability that another item will appear with it. Going to the technical part of the association rules, association rules are calculated by calculating confidence, lift, and support ratios. Support ratio: the ratio of occurrence of a given item in the whole transactional log. Confidence ratio: the number of occurrences of a second item right combined with the occurrence of the first item. In other words, a ratio that calculates the following statement, if the first item appeared, then the second item will appear with it in the same transaction. Lastly, the lift is comparing our expected value of the confidence ratio with the calculated confidence ratio.

Nandy (2018: 34-35) defines market basket analysis in more details, as it is the selected combination of products that a customer chose in a single buying transaction, most of the time, the combination of these products is not random and there is a reason why these products were bought together at the same time. For example, spreadable jars like butter or jam are frequently bought with bread as well as the famous beer and diaper combination. Going to the technical part, if A implies B with a confidence ratio of 60%,

this means that given all transactions that contain product A, 60% of these transactions will also contain product B, and this implies that there is a specific relationship and a pattern between products A and B. Looking at the other side of the coin, there is the negative association which is the usage or buying of one product implies that there will be a decrease in the buying of another product. For example, milk and milk powder. Most of the time, if people buy one of them, this implies that they will not buy the other one, using both techniques can help us in discovering hidden patterns that will allow us to discover things and act upon them.

2.3 Natural Language Processing Definition

Rezaii et al. (2022: 251) had a brief definition of NLP which is a computer science field that focuses on acquiring, understanding, and making use of common human languages. Nowadays, new NLP models can study specific syntax, semantic, and practical information from massive volumes of text blocks. NLP models can apply their algorithm even on a large number as billions of words, then it can save it securely in multiple layers/levels of artificial neural grids. The output of an NLP algorithm can be used to solve and end many problems/challenges and also extend the power of using natural language in computer science to solve everyday problems.

Quarteroni (2018: 105) defines NLP differently, they stated that NLP is a subdivision of artificial intelligence that operates using machine learning techniques, to explore a given text/language. NLP is not new as there were NLP studies that were conducted in the 1960s, but in the last few decades, it has become a much more essential and important topic. Natural language understanding (NLU) is a term that defines and explains the scope of activities that are conducted using NLP algorithms. NLP is used in a wide variety of computer activities such as virtual assistants which is a technology that is used to extract and classify information automatically.

Agarwal (2019: 2811) had a similar definition, in which they stated that natural language processing is defined as the unit of machine learning that is concerned with languages, text, and speech. Computers use NLP to explore and comprehend useful information in a given human language efficiently and effectively, and that is made by a span of calculative algorithms that study and describe the natural languages with multiple levels

of a linguistic breakdown so that a similar human-like language processing can be achieved but in a much more efficient way. NLP is used in many technologies like machine translation, email spam detection, information extraction, summarization, etc. As it can mimic the way humans comprehend, manipulate and use languages to achieve many complex goals.

In George et al.'s (2021: 2) study, Natural Language Processing is defined as a technological method that is used in order to make systems apprehend human languages effectively. NLP is made to manage and manipulate human communication information, without depending on the input of the model to be in a specific programming language. Unlike other computer science branches that use conventional programming techniques, NLP uses complex machine learning algorithms to attain tasks like classification, text/speech recognition, and prediction.

2.3.1 Natural Language Processing Goals

With the use of natural language processing, artificial intelligence, and machine learning, there are some goals that programmers seek to attain goals that will help in information extraction and finding hidden patterns in huge textual files. Some of these goals are 1)Sentiment analysis which is the practice of computationally finding and categorizing opinions expressed in a text, particularly to identify whether the writer has a positive, negative, or neutral opinion on a given subject, item, etc. 2)Named entity recognition: is a subtask of information extraction that searches for and organizes named entities stated in unstructured text into predefined categories, such as names of people, places, organizations, animals, medical terms, expressions, amounts, numbers, money, percentages, etc. it can be anything predefined to the machine learning model. 3) Text summarization: the process of reducing lengthy texts to only the key ideas from the document's main points to be included in the summary. The technical details of the above-stated three main goals will be explained below.

2.3.1.1 Sentiment Analysis

According to Xu et al. (2019: 51522) sentiment analysis is a section of artificial intelligence/natural language processing, and it is concerned with understanding and acquiring the sensed sentiment in a given text automatically. This can be attained by using

text classification algorithms. Sentiment analysis can be used to analyze the emotional trends and comments of the general public about a given aspect/topic with the help of artificial intelligence, machine learning, data mining, and natural language processing technologies. As stated above, Sentiment analysis is used to study and examine the emotions/opinions of people so that with the help of this technology we can know the user's positive, negative or neutral opinions and emotions about a given aspect in an efficient way, this aspect can be news, product or a film, etc. Businesses can make useful use of Sentiment analysis techniques in order to increase their customer satisfaction levels.

Puschmann and Powell (2018:1-5) had a similar definition, in which they states that sentiment analysis is a process of opinion mining and these opinion mining techniques are frequently used in social sciences, media, and communication fields. It is also defined as the field of computer science that is mainly concerned with market research, public relations, and political forecasting, also Sentiment analysis is used in social sciences and social media studies. It is mainly concerned with expressing emotions and opinions that are most of the time generated from user-generated content (UGC), Sentiment analysis is used to study and examine ("datafication") of emotions and opinions. For a proper Sentiment analysis model, the given text to be analyzed has to have a big volume and that the text has emotional words that can be analyzed by the algorithm (like opinions, emotions, and sentiment) not only facts that do not show any emotions so that these opinions can be measured and studied.

Haque et al. (2020: 403-404) stated that, in the ERA of the internet, 2.5 quintillions of data are generated on a daily basis, so the use of an algorithm like sentiment analysis became crucial and essential in order to analyze and study words that express emotions and opinions in a given text. In other words Sentiment analysis is a branch of natural language processing that helps in analyzing opinions and emotions in a given text. Sentiment analysis has major applications and importance in the business and academic fields, as it focuses on anticipating polarities (positive or negative emotions).

Natural Language Processing has acquired noticeable attention as Mishev et al. (2020: 131662) stated because of how efficiently it can be used in modeling, and how it can be used in multiple applications and fields as it is a reliable, efficient, and effective text-

oriented algorithm. One of the most important Natural Language Processing algorithms and goals is Sentiment analysis. Oxford University defines Sentiment analysis as an operation of computationally specifying, studying, and organizing opinions and emotions in a given text block to differentiate between people's behaviors and opinions whether it is positive, negative, or neutral. Sentiment analysis studies these behaviors and helps in performing an actionable task.

2.3.1.2 Named Entity Recognition

According to Yadav and Bethard (2019: 1), Named Entity Recognition is defined as an important element in any natural language processing system that helps in answering multiple questions and acquisition of information. There were multiple studies on NER in the past few decades, but recently with the help of deep neural networks (NN), NER systems became more professional and accurate in the last few years. In other words, Named entity recognition is concerned with specifying entities like a person, location, organization, drug, time, clinical procedure, biological protein, etc. in a given text block. Commonly, it is used as a first step when any type of information extraction is needed and to give answers to multiple questions. The first versions of early NER models were established and made by handcrafted protocols and algorithms, not like today where most of the NER models are based on complex computational algorithms.

Li et al. (2020: 1) had a more detailed definition of what NER is as it defines named entity recognition as a procedure that is followed to identify and specify pre-known discrete states like a person, location, organization, etc. in any given text that uses a predefined and known language. Using NER can help in answering multiple questions, summarize texts, and help in automated translations. Nowadays, NER became so much more powerful as deep learning is now used to help in the process. Named entity recognition is not only used for querying and searching for information but it can also be used in many NLP models and algorithms like knowledge base construction. NER has mainly 4 techniques in order to be executed 1) Rule-based approaches, that is mainly based on hand-crafted rules; 2) Unsupervised learning approaches, which rely on unsupervised algorithms where the variables are not well known which is the dependent and which are the independent, 3) supervised learning algorithms were the variables are known which

is the dependent and the independent 4) Deep-learning based approaches, that locates entities needed for classification and categorizing from the raw input data automatically.

Yu et al. (2020: 1) stated a detailed example of what NER is, but firstly they defined NER as an indispensable step in natural language processing algorithms. Named Entity Recognition is concerned with pointing out some special keywords that are related to predefined entities (people, places, etc), in this study NER is classified into two types: 1) Nested Entities are entities that contain a smaller entity which is referenced in a bigger entity phrase/keyword. For example, [Bank of [China]], in which [China] is an independent entity (country) that is being used in a larger entity [Bank of China]. The other type is Flat Named Entity Recognition, it is more commonly used compared to nested entities, and it is based on simple and sequential labeling.

Singh et al. (2018: 27) explained different concepts regarding NER, as they stated that it is an important task in Natural Language Processing, NER is considered a subtask in the Information Extraction process. For an efficient and effective NER process, there should be sufficient entities and information available in the given text. Entity extraction has received a lot of attention. When informal languages and code-mixed (using more than one language in the same sentence) text is used, the NER process becomes a more difficult job, because of how unstructured and vague the information can be.

2.3.1.3 Text summarization

According to Abualigah et al. (2020: 1) text summarization is defined as a procedure that is followed in order to develop a summary of a given text so that this summary has indispensable and prominent information and does not leave any important details which are in the given text. Multidocument abstractive summarization seeks to provide a compressed version of the document while preserving the crucial details. Text summarization became much more important in the last decade due to the huge amount of data that is produced on daily bases on the Internet. In other words, text Summarization helps in making the information extraction process much more efficient and effective with less time and effort, due to the elimination of irrelevant sentences/phrases which makes the information extraction easier.

Talukder et al. (2019: 1) aimed to construct an efficient and effective text summarization algorithm that deals with the Bengali language so that the output of the algorithm is comprehensible and understandable, they used text inputs from newspapers and Facebook posts, and their algorithm works with LSTM their model is sequential, and they faced difficulties in preprocessing and finding out unknown words. The researchers state that there are two ways of Text summarization 1) extractive: which needs to highlight and extract the key phrases or sentences in the passage, 2) abstractive: which creates a bottom-up summary of the given text, keeping in mind that not all of the words will be in from the main passage. In other words, an abstract technique can produce a summary of a particular text from within itself.

Due to the enormous amount of textual data that is produced on daily basis continuously example: (news articles, scientific papers, and legal documents). Summarizing these blocks of text manually is an impossible job due to the amount of time effort and money needed for this process as stated by El-Kassas et al. (2021: 1-2), so since 1950, many researchers were trying to invent algorithms that can summarize texts automatically. There are three main types of Text summarization extractive: selects the important keywords from the input document(s), then unites them to generate the summary, abstractive: constructs the summary with sentences that are different from the actual sentences using a bottom-up approach, or hybrid which is a mix between the two previous methods as stated by Kieuvongngam et al. (2020: 1-2). The steps which were conducted in order to implement a Text summarization are as follows: 1) Pre-Processing making the original input text in a structural form, 2) actual summarization of the output of the first step by using 1 of the 3 techniques stated above, 3) Post-Processing fixing and handling any errors and inconsistencies which were falsely made during the summarization step before producing the final output of the algorithm.

2.3.2 Arabic Natural Language Processing and its state of art

Obeid et al. (2020: 7022) shows the challenges that NLP programmers will face especially while dealing with ANLP (Arabic Natural language processing). Some of these challenges are 1) Orthographic Ambiguity: Arabic uses the Abjad script that has some complementary diacritical marks which are called in Arabic (التشكيل). The absence of these marks in any given word most of the time is not a major problem to Arabic native speakers

as they can understand the word in its context, but the absence of these marks in an NLP model can cause inconsistencies and inaccuracies. 2) Dialectal Variation, even though the official Arabic dialect is the MSA (Modern Standard Arabic) translates to: (العربية اللغه الفصحي), it is not the practical mother language/delicate to most Arabic speakers in their everyday conversations, and there are multiple of very different dialects of the same language (such as Egyptian, Levantine, and Gulf) these dialects differ from the MSA and from each other. So using MSA as the dialect from NLP models can be inefficient and ineffective, so the NLP should use the dialect that is targeted by the model for accurate outputs. 3) Orthographic Inconsistency: Arabic speakers online have a lot of spelling inconsistencies as again their dilacte differs from the MSA, an example this study showed for that challenge is how different the people write an Arabic word like "مبيقولهاش) "translates to "he does not say it"), it has some spelling variations like "مبيقلهاش "and "مبؤلهاش "so this adds more challenges to ANLP models.

Shaalan et al. (2019: 62-73) stated different challenges that ANLP programmers may face, some of these challenges are: 1) shape of the letters, given an Arabic letter, unlike English, it can have more than one shape, the shape of the letter is determined based on the position of the letter in the word it can be connected to the previous or latter letter or just stand alone without any attachments, an example of that is the letter "ف "it can have multiple formats like "ف"/"ـفـ"/"فـ, so this is a challenge that can face ANLP programmers. Another challenge is the nonappearance of capital letters, this makes the named entity recognition much harder as compared to English where the appearance of a capital letter in the first letter of a word helps in the recognition of entities in a given text block, a given example is that in English it is easier to recognize the name "Adam" in a given text because of the capitalization of the first letter, so it is much easier to recognize this word as a name, while in Arabic it is written as "ادم", so no capitalization that makes this word stand out, hence it is harder to recognize. Another challenge is Multi word expressions this means that combing two Arabic words together can give a third meaning which is completely different than the meaning of the two words used in the combination, an example can be "فقر "which means "poor" and "دم "which means blood when combined "دم فقر "it will not mean "poor blood", but it will mean "Anemia", so the challenge of Multi word expressions can make ANLP much more challenging.

Farghaly and Shaalan, (2009: 2-11) emphasized on how (ANLP) had an increase in prominence and attention. Recently, ANLP programmers developed a wide range of ANLP applications, like machine translation, information retrieval and extraction, speech synthesis and recognition, localization, and multilingual information retrieval systems. While programming these applications, ANLP developers faced a number of challenges, such as spelling inconsistencies and the absence of capitalization of the first letter, unlike the English language. The main target for designing the ANLP applications is to allow non-Arabic speakers to understand and interpret Arabic texts. There are multiple solutions to the challenges of ANLP. One of these solutions is while dealing with the challenge of the absence of capitalization if we are searching for Arabic names in Arabic texts using named entity recognition techniques, a large percentage of Arabic names especially in the gulf region contain the word "بن "meaning "son of" in the middle of their names, this trick can help in recognizing a large percentage of Arabic names.

Kanan et al. (2019: 622-624) had a general definition of ANLP, as he defined ANLP as a division of computer science and artificial intelligence that deals with the studying and examination of text blocks in an automated process without the usage of any manual effort. The main target of using ANLP is to interpret Arabic text blocks automatically by using some NLP techniques. Some of these techniques are tokenization which is a procedure of dividing text blocks so that every given word in a text block is separated and differentiated, this division is done by separating every token via the spaces between the words. Another technique is part of speech tagging which is an algorithm of recognizing every word in a given text block with some rules like its form, shape, and position in every sentence, whether it can be a noun, verb, adverb, or adjective. Clustering can help in the ANLP process as it collects words with high similarity and elements in a group called a cluster. The clustering process has two main types. The first type is hard cluster and its rule is that every given word shall only exist in one, and only one cluster, while soft cluster does not apply this rule, where it is accepted for a word to be in one, or more than one cluster.

2.3.3 Natural Language Processing Vendors

The first vendor we will talk about is Monkey Learn. It is a platform that is very easy to use and many businesses use it to extract knowledge from unstructured text files. Users

can enter the text file they want as an input and Monkey learn will perform multiple operations on it like sentiment analysis, keyword extraction, and topic classification. Also, users of Monkey Learn can develop a tailored machine learning model that can produce outputs that is suitable for the business needs of the client. Furthermore, monkey learn can be connected to other platforms like Excel and Google sheets (Top 10 NLP Tools & Services in 2022, 2020).

The second vendor is Amazon comprehend which is an NLP model developed by Amazon and uses the power of Amazon's cloud computing and infrastructure to operate, they have a user friendly API that people can use to do named entity recognition, sentiment analysis, etc. What is special about Amazon Comprehend is that they have a special NLP model Called Amazon Comprehend Medical which is main specifically to apply NLP tasks on medical data (Top 10 NLP Tools & Services in 2022, 2020).

The third vendor is Intel which developed a Python library that can perform natural language processing, Intel developed this Python library by using powerful AI techniques. The main advantage of that library is that it can be used in a wide range of scenarios and business fields (Zephin et al., 2022).

The fourth vendor is IBM which developed IBM Waston which is used by businesses for the tasks of Natural Language Processing, businesses use this tool for keyword extraction, sentiment analysis and named entity recognition. One of the main advantages of IBM Watson is that is has a free trail of 30 days, then the paying model is pay as you go model which is useful for small businesses that have difficulty financial wise and which to use NLP techniques (Top natural language processing (NLP) providers in 2022, 2022).

The fifth vendor is Stanford Core Natural Language processing which is developed by Stanford University. It is simple to use and it performs mist of the NLP tasks like tokenization, named entity recognition and part of speech tagging. Some of its main advantages are scalability and speed, but there is a drawback which is that you have to have JDK on the computer in order to use it as it operates using Java (Top 10 NLP Tools & Services in 2022, 2020).

2.4 General Natural Language Processing Applications

There are many practical and real life cases where NLP was a key player in achieving a certain business goal, as it helps in dealing with large sizes of text blocks whether it's in the structured or unstructured format. NLP can deal with text blocks that come from many sources like emails, social media, survey answers, etc. Using NLP in any business field will help in the fast and automated analysis of text which can give any company a competitive advantage. We are going to discuss three of the main NLP applications which are voice applications and chat bots, Customer Feedback, and Speech-to-text applications

2.4.1 Natural Language Processing Voice Applications and Chat Bots

Sudarsan and Kumar (2019: 133-135) stated that nowadays, voice bots have grabbed the attention of businesses, many banks use voice bots if a client called their hotline at the beginning of the call. These voice bots use NLP technologies to listen and understand the client's request. These voice bots use AI and NLP technologies in order to have the power of understanding human sentences and act upon them. These voice bots have to have some technologies that will be discussed in order to be efficient and effective. Some of these technologies are: the detection of the client's sentiment based on his/her words, gathering these types of data is important for quality control and marketing departments, and it can use sentiment analysis techniques in order to attain this technique. 2) Usages of any predefined banned words, this can be done by having a list of banned words so that the voice bot is triggered once it hears a word from that list. 3) Working in a service like client support, the bot must use greeting words at the beginning of the conversation. Having said that, it is important that the voice bot do this for every given call with a client. 4) Mentioning of a competitor's name, it is important for the voice bot to automatically detect any mention of the competitor's name, and this can be done just the same as the banned word technique where there is a list of the competitor's names. It is important to do so for marketing purposes.

Ayanouz et al. (2020: 2-3) explained another application of using NLP in chat bots. Chat bots use AI and NLP to help in the customer service process in a more efficient and effective way. The key player for implementing a successful chat bot is to use the most suitable NLP engine, the selection of the NLP engine depends on the type of chat bot

which needs to be implemented, whether it deals with structured or unstructured conversations with customers. Using structured type of chat bots can be easier to program and implement, but unstructured chat bots tend to solve customers' problems more compared to structured ones. There are many challenges that face chat bots, some of these challenges are grammatical errors made by customers while texting the chat bot, which will make the task of understanding the customer's concern more complicated for the chat bot, as well as difficulty in detecting the sentiment of customers, thus it is important to use sentiment analysis techniques

2.4.2 Understanding Customer Feedback

According to Ramaswamy and DeClerck (2018: 170), every business must have the ability to analyze and study customer satisfaction, as this will help them know their customer's needs and wants and act upon them, thus increasing customer satisfaction and maintaining competitive advantage. Knowing your customer's opinions matters in the manufacturing of the product and marketing-wise. There are many ways a business can collect the opinion of their customer whether online or offline, thus it is important to have a strong NLP and AI-based model that can classify and study your customer's feedback. This study is mainly concerned with unstructured data on social media whether it is text chats, comments, or voice recordings model implemented in it. It is important to analyze these data in order to know the company's weak points to solve them. Using deep learning, AI, and NLP help in the automation and the accuracy of the whole process.

Getting to the technical part of this application Nikitha et al. (2020: 24) applied this application by using the following steps: 1) Removing punctuation from reviews, punctuations are important for human beings as it helps us in understanding the sentences and their contexts, but for NLP models they are just excess and unwanted words so we delete them. 2) Tokenization: putting each word in an entity called "token" so that every token can be studied and analyzed alone. 3) Removal of excess filler words: just like step 1 filler words are important for humans as it helps in understanding the sentences, but in NLP models, a filler like a word "or" is useless, hence it is deleted. 4) Stemming: which is returning each given token to its simplest form in order to make it easier to deal with, this can be done by removing “ing”, “ly”, “s”, etc. After using all these previous steps, it is now easier to deal with and understand customer feedback.

2.4.3 Speech to Text Applications

According to Win et al. (2020: 112-113), speech-to-text applications have many uses nowadays, as it executes actions that are commanded by human voices. Thus, it has to be very well programmed to hear, interpret and understand vocabulary as well as having the ability to differentiate between speech patterns. This can be done by using the sciences of linguistics, computer science, and AI. The main mission of a Speech-to-text application acknowledges any human language given to it sound wise and act upon it as desired by the user. To do this, there are two main components. Which are natural language processing (text studying) and digital signal processing (electronic phonetic studying).

Vinnarasu and Jose (2019: 3642-3643) had a similar definition as they stated that speech is the main communication path between people, so it is very important to apply AI and NLP to speech in order to analyze it and make use of it. That is why Speech recognition techniques are used, which are the techniques that allow the model to acknowledge different speech from different people. Successful speech recognition models should interpret and understand differences in accents from one person to another. NLP and Text summarization techniques are key players in the speech recognition models, as they help in removing excess and unwanted words and leave only the needed ones to be studied in detail. The speech recognition process goes as follows: Firstly: voice recording, then preprocessing the input voice, extraction of the needed and important information, then changing the important information from the voice format to the text format. After that, we summarize the text format leaving only the important and needed information.

2.5 Research Gap

Badaro et al. (2019: 33-34) stated that their research had an unsolved problem/research gap which is that most of the ANLP systems deal with MSA, which is unpractical as every country/region has its own dialect, and words have different meanings depending on which country are you in. Having said that, it is important to develop different ANLP systems that deal with each different Arabic dialect and use each model on its specific dialect to have more accurate results and practicality.

According to Wahdan et al. (2021: 101) there is a major gap in the number of ANLP researchers compared to English NLP researchers, this gap causes slow development in the ANLP field, Arabic machine Learning programmers need to give attention to ANLP as the Arabic Language is one of the most spoken languages worldwide and there are many hidden benefits of applying machine learning algorithms and NLP on the Arabic language with its different dialects.

Obiedat et al. (2021: 152644) stated that their research faced a challenge in finding Arabic datasets easily with different types and themes, as per the study, most of the data sets available are hotel booking or book reviews, which is not enough if we need to develop multi use ANLP models, so we should focus on constructing new Arabic data sets with different themes and business fields.

Farha and Magdy (2019: 196) stated that there are many Arabs who are using the English alphabet to write Arabic words and expressions on social media, and this is called Arabizi. Even though ANLP is only concerned with the Arabic script. There is a lot of loss of data and information due to the use of Arabizi. So it is important to also develop NLP systems that can deal with and understand Arabizi.

The most similar gap to our thesis of the previously stated four is the research made by (Badaro et al., 2019: 33-34) which states that most ANLP systems deal with MSA which is the Modern Standard Arabic, which is not practical as there are many different Arabic dialects, vocab and sentence structures according to each Arabic country, so our study will mainly focus on developing in ANLP system that is mainly centered around the Egyptian Arabic only, because Egyptians has some unique vocabulary that can be different from the MSA.

To fill this gap we have research questions and hypothesis that we will investigate to produce useful findings:

Research Question 1: How Egyptian Arabic can be processed using Arabic Natural Language Processing (ANLP) techniques to increase the quality of e-services?

Research Question 2: What are the benefits of Categorizing Egyptian Arabic texts using ANLP techniques on the quality of e-services?

Research Question 3: What Categorizing techniques can be used on Egyptian Arabic Texts in order to increase e-services qualities?

Research question 4: How can dividing Egyptian Arabic text into topics of interest improve the quality of e-services?

Hypothesis: Using ANLP techniques on Egyptian Arabic text will improve the eservices qualities.

This hypothesis can be linked to research questions number 1 and 3 as our main target in this thesis is to develop a sentiment analysis model that classifies Egyptian Arabic books and movies reviews into positive and negative sentiments, as this can help in understanding customer's feedback and improving e services qualities.

3 Application: CRISP-DM Model for E-services quality NLP analysis

This thesis focuses on implementing sentiment analysis techniques to tackle the business problem of understanding customer's feedback and identifying their needs in the Egyptian Arabic language. By analysing the sentiment expressed in customer reviews and comments, valuable insights can be gained to enhance marketing campaigns and improve its overall efficiency. The study aims to bridge the gap in understanding customer sentiments and translate them into actionable strategies for businesses operating in the Egyptian market. As stated above, this can be done by implementing sentiment analysis techniques that helps us in understating whether a customer has a positive or a negative sentiment, our sentiment analysis technique was built by applying CRISP-DM model which will be discussed in details in the following section.

Our methodology which we will be using to fill the research gap discussed above is the CRISP-DM methodology which are six main steps that data mining models should follow in order to attain reliable results. The first step is Business understanding, which is knowing what business problem or challenges this data mining project aims to solve, this should be done by considering the inputs that will be used and the desired output, this is the step where the planning for the whole project happens. The second step is data understanding, which is studying the available data that the model will be applied to. For example, to get to know its size and attributes. Then data preparation happens, which is processing the data to make it ready for modelling. For example, dealing with noisy data and imputing missing attributes (Schröer et al., 2021: 527).

The fourth step is modelling, which is choosing the most appropriate modelling techniques, there are many modelling techniques available and some of them are stated above in the theoretical considerations section, the most appropriate one should be chosen carefully depending on the model's target and the business problem that the researchers are trying to solve and it also depends in the dataset type. Then the evaluation step takes place, where the accuracy of the model built in the previous step is evaluated. The evaluation depends on one predefined targets and the business problem itself. Were we compare our desired results with our actual ones. After making sure that the requirements are met, deployment takes place which is using the model to solve the predefined business

problem, inefficiencies may arise, so it is important to have scheduled maintenance of the model in order to improve its accuracy and efficiency as well as constant monitoring if the model's output (Schröer et al., 2021: 527).

The following figure illustrates the order of the steps and how they are connected to each other

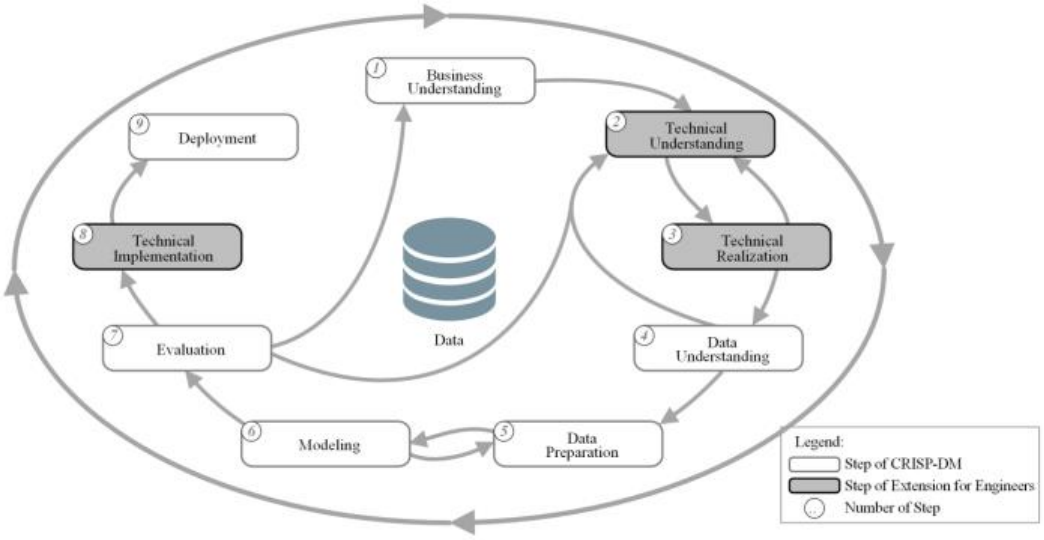


Figure 1 The CRISP-DM process, (Huber et al., 2019: 406).

4 Research Methodology

In this section, we will state how the scientific steps of CRISP-DM methodology that were discussed in detail in the previous section were applied to our sentiment analysis model as well as the challenges we faced, how we did data cleaning and pre-processing in order to ensure accurate results and the different models that were used.

4.1 Business understanding

One of the main challenges this thesis aims to face is that there is low attention worldwide that is given to the topic of ANLP which is applying NLP models and techniques to the Arabic text, most attention is given to the English language where there are numerous studies and sentiment analysis applied models. Having said that, we aim to do an accurate sentiment analysis model that classifies 100k Arabic movies and book reviews into positive and negative sentiments. Our main goal is to implement an accurate model that can be used to automatically classify Egyptian Arabic reviews that can be used by Egyptian businesses to study and classify their reviews. We faced several challenges, one of the main challenges was being able to have a reliable dataset which can be used to implement an accurate model based on it

4.2 Data understating

The dataset on which we are applying our NLP model was found on Kaggle[1] platform, and its name is "Arabic 100 Reviews" which contains 100k reviews written in the

[1] Dataset link: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews

Github link that contains the workbook:
https://github.com/Ahmedtarekyoussef/Machine-Learning-Models/blob/main/AhmedTarek_49_4345%20(3).ipynb
Workbook link:
https://colab.research.google.com/drive/1EuPeu9hq1sWfdJEqnrATeAKoooLctnAW?usp=sharing

Egyptian Arabic language about movies and books, it contains only two columns the first one is the column which holds the review, and the second one is the sentiment of this review whether it is positive, neutral or negative sentiment. After data exploration, we found these facts about the dataset. Firstly, it is balanced as it contains an equal number of reviews per sentiment (33333 reviews per sentiment) which totals to 99999 reviews, we used a count plot graph in order to make sure that it is balanced where the 3 sentiments were on the x-axis and the count per sentiment is on the y axis, also we checked if there are any null values and we found out that there is not. Lastly, we found out that there is not any duplicated record.

4.3 Data Preparation

In this Stage, we are mainly concerned about one of the prominent steps in any NLP model which is data cleaning and pre-processing, as stated above we already are sure that the dataset we are using is balanced and we visualized this using a count plot that shows the number of reviews per sentiment, we also are sure that there are not any missing values so there is no need of using any imputation techniques, the first step which we made was tokenization, which is breaking down the review in each record into smaller units known as tokens. Each token holds a word from the original sentence, this breaking down helps in attaining high accuracy results as it is easier to manage small tokens rather than dealing with the whole review at once. We did tokenization by importing ' TreebankWordTokenizer' which is a part of the NLTK (Natural Language Toolkit) library. Which is a popular library that can be used in several NLP tasks. We used the command "nltk.download('punkt')" to download the library.

We decided to remove the stop words, as they appear in every review and they are not useful as they do not carry a sentiment that can be analyzed, so we will remove them from each review in the reviews column to focus only on the important words which could carry a sentiment, to do that we downloaded a list of Arabic stop words from NLTK library and assigned this list to a variable that we called "stop_words". After that, we noticed that the stop words in the list have diacritics (تشكيل), and our original dataset does not contain any diacritics, so we decided to remove the diacritics from the list by using regular expressions, we faced some difficulties trying to construct the code on our own

so we used the help of Github. After we imported the regex library, we used a regular expression to iterate over each stop word in the list to remove diacritics from it, after that we discovered that some words in the stop words list are negation words that carry sentimental meaning, for example, the word "جميلة" which is beautiful in English if there is a negation word before it like "ليست جميلة" which in English is not beautiful that will carry a whole different meaning, so we decided to remove the most popular negation words which are ['ليست', 'لم', 'لن', 'غير', 'ليس', 'لا'] from the stopwords list. Finally, we used a lambda function to remove from all the reviews in the dataset any token which is also in the stop_words list, so now there are not any stop words in the reviews column except the negation stop words as they can carry a sentimental meaning.

The next step was removing any punctuation as they are useless to our model and we want to focus only on tokens that can carry a sentimental meaning for our sentiment analysis model. We used a lambda function that replaces any punctuation character like (.,-,_,) with "" (which is an empty space or empty character), to simply delete any punctuation character. The last thing we did in this step was stemming as we believe that stemming will help in achieving better results as it is the process of transforming words to their base or parent word for better analysis results. We did that by using the nltk library and the ISRIStemmer class and then applying the stemming to the reviews column, we faced some difficulties in constructing the lemmatization code on our own so we used the help of Chat-GPT to tell us the steps of implementing this code. After the code was implemented, most of the words were returned to their parent word. For example, the word "جمال" was returned to its parent or root word which is "جمل".

4.4 Modelling[2]

We chose two techniques of modeling which are multinomial Naive Bayes and decision tree classifier, which were discussed in detail in the theoretical considerations section.

[2] Dataset link: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews

Also, we used 10-fold cross-validation to make sure that each record participated in both roles which are the train and test roles. In both models, we splitted the data into training and test sets using the train_test_split function from scikit-learn. We performed 10-fold cross-validation on the training data using cross_val_score to evaluate the model's performance. Then, we trained the model on the entire training data using pipe.fit. Lastly, we predicted the labels for the test set using pipe.predict. We use the F1 score evaluation technique to evaluate the accuracy of our model. When the data set had 3 types of sentiments (positive, neutral, and negative). The F1 score of both modeling techniques was unsatisfactory, then we discovered that having neutral sentiments in our dataset can cause high inaccuracies, as neutral sentiments make it more challenging for the model to understand what words lead to positive sentiments and what words leave negative ones. Having said that, we decided to delete any record that have neutral sentiment, leaving only positive and negative sentiments, after retrying the two modeling techniques, the accuracy has significantly increased, we will talk more in detail about the F1 score and the accuracy of the models in the results chapter.

4.5 Deployment

We advise Egyptian E-businesses to use our sentiment analysis model in order to understand their customers' feedback and analyze it in an automated procedure, rather than reading each and every review to classify it into positive and negative sentiments manually as this will save them much time, effort, and money. Moreover, it is important to use this model in order to know the customer’s specific needs and wants as it is important for any business to understand the needs of its customers in order to modify

---

Github link that contains the workbook:
https://github.com/Ahmedtarekyoussef/Machine-Learning-Models/blob/main/AhmedTarek_49_4345%20(3).ipynb
Workbook link:
https://colab.research.google.com/drive/1EuPeu9hq1sWfdJEqnrATeAKoooLctnAW?usp=sharing

their production to fill these needs as well as issuing more successful marketing campaigns that addresses directly the needs and wants of it potential customers. In addition, these types of NLP models are not widely used in the Egyptian market, so any business that will apply our model will leverage a competitive advantage compared to the other businesses. One of the main advantages of our model is that it does not require complex computational power in order to be executed, so small businesses can use it freely as long as they have a computer and internet connection. Also, they need to have some sort of API that collects automatically their customers' reviews from their website and stores it in a CSV file, as our sentiment analysis model takes input a data set of CSV format.

5 Results

We used F1 score to evaluate the accuracy of the models, its equation is as follow where is tries to find the balance between precision and recall

$$\text{F1\_Score} = 2 * \frac{Precision + Recall}{Precision * Recall}$$

Figure 2 How F1 score is calculated (Maseer, 2021: 22358)

As discussed above we used Multinomial Naive Bayes and decision tree classifier, we tried both models when there were 3 sentiment types (positive, neutral, negative), and the F1 score for both models were as follows:

Table 1: F1 score of both models while having three types of sentiments

	F1 Score	Cross-validation mean	Cross-validation standard deviation
Multinomial Naive Bayes	61.76%	61.36%	0.82%
Decision tree classifier	55.43%	52.35%	0.22%

After we found out that the F1 scores were less than our expectations, we discovered that having neutral sentiments in the dataset can have a considerable negative impact on the accuracy of the model, so we decided to delete any record that holds a neutral sentiment, and the new results are as follows

	F1 Score	Cross-validation mean	Cross-validation standard deviation
Multinomial Naive Bayes	81.98%	81.59%	0.39%
Decision tree classifier	74.98%	73.75%	0.47%

Table 2: F1 scores of both models after deleting the neutral sentiment

As the tables show, using multinomial Naïve Bayes showed higher F1 score and accuracy, so by comparing the accuracy of both models, we believe that using multinomial Naïve Bayes and more suitable while dealing with our ANLP sentiment analysis model.

Github link that contains the workbook:
https://github.com/Ahmedtarekyoussef/Machine-Learning-Models/blob/main/AhmedTarek_49_4345%20(3).ipynb

Google Collab Workbook link:
https://colab.research.google.com/drive/1EuPeu9hq1sWfdJEqnrATeAKoooLctnAW?usp=sharing

6 Discussion

As we discussed above, we used two types of models (decision tree classifier and multinomial Naïve Bayes) and we used the F1 score to evaluate the accuracy of both models, we believe that the multinomial Naïve Bayes performed better as it achieved higher F1 score. In addition, we discussed the practical details of our NLP model and its F1 score, we recommend future researchers to try and find other datasets that may contain a more diverse and higher quantity of reviews that are not just mainly books and movies reviews, a more advanced approach for applying NLP techniques on Egyptian Arabic text would be applying clustering techniques which can be used to classify words in reviews into multiple categories. For example, categories that classify the type of complaints whether it is complaints about the quality of a product, delivery timing, or price using our model or even developing a more advanced one can help companies in Egypt in classifying their reviews into positive and negative ones easily, and to know their customer's feedback in an easy and efficient way. Also, it can help companies in Egypt to identify their strong and weak points and what should they improve. If a company has 10,000 reviews for example, it would take much time to classify those reviews manually whether they are positive or negative, but by applying our model, the classification will be made in a few seconds with more than 80% accuracy, future researchers are advised to try to improve the F1 score of the model may be using more advanced data cleaning and pre-processing techniques that could have a positive impact on the model accuracy.

7 Conclusion

To conclude what we studied in this thesis, we examined what is artificial intelligence and its different types and forms and uses, we stated three of them which are machine learning, deep learning, and neural networks. Also, we defined what text mining is and what the advantages of using it are. The next chapter was talking about what is data mining, why is it important nowadays, and how data pre-processing is one of the most important steps in any data mining model. Also, we talked about the two main data mining learning models which are supervised learning that contains linear regression, Naïve Bayes, and decision trees, and we also talked about unsupervised learning that contained clustering and association rules. Also, we talked about Natural language processing and how businesses use it in their operations and why is it important with its different tasks like sentiment analysis, named entity recognition, and text summarization. Then we talked about the state of the art of Arabic NLP and the challenges that it faces compared to English NLP. After that, we discussed five different and popular NLP vendors and their strong and weak points. Moreover, we discussed different real-life applications of the usage of NLP. We stated our research gap, we stated our research questions, and our hypothesis that worked on trying to solve them and reach useful findings in the field of Egyptian-Arabic natural language processing. We displayed our NLP model which aims to classify Egyptian Arabic text into positive and negative sentiments, we discussed that using multinomial Naive Bayes was the best modelling technique possible as it is the one that showed the highest accuracy results, and we discussed our findings. Lastly, we gave advice to future researchers who may be interested in our topic on how they can improve their research and reach their desired objectives.

References

Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text summarization: a brief review. Recent Advances in NLP: the case of Arabic language, 1-15

Agarwal, M. (2019). An overview of natural language processing. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 7, 2811-2813.

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. Ieee Access, 7, 31883-31902

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

Ariza Colpas, P., Vicario, E., De-La-Hoz-Franco, E., Pineres-Melo, M., OviedoCarrascal, A., & Patara, F. (2020). Unsupervised human activity recognition using the clustering approach: A review. Sensors, 20(9), 2702

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020, March). A smart chatbot architecture based NLP and machine learning for health care assistance. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security (pp. 1-6).

Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K. B., Habash, N., ... & Hamdi, A. (2019). A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3), 1-52.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655-1674.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 1-22.

Farha, I. A., & Magdy, W. (2019, August). Mazajak: An online Arabic sentiment analyser. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 192-198).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6), e1332.

George, N., Muiz, K., Whig, P., & Velu, A. (2021). Framework of Perceptive Artificial Intelligence using Natural Language Processing (PAIN). Artificial & Computational Intelligence/Published Online: July.

Gunawan, O., & Aldridge, T. (2018). Text mining of Scottish post-emergency and training exercise debrief reports.

Haque, S., Rahman, T., Shakir, A. K., Arman, M., Biplob, K. B. B., Himu, F. A., ... & Islam, M. S. (2020, February). Aspect based sentiment analysis in Bangla dataset based on aspect term extraction. In International Conference on Cyber Security and Computer Science (pp. 403-413). Springer, Cham.

Harahap, F., Harahap, A. Y. N., Ekadiansyah, E., Sari, R. N., Adawiyah, R., & Harahap, C. B. (2018, August). Implementation of Naïve Bayes classification method for predicting purchase. In 2018 6th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-5). IEEE.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp*, *79*, 403-408.

Jarek, K., & Mazurek, G. (2019). Marketing and artificial intelligence. Central European Business Review, 8(2), 46.

Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AlZu'bi, S., Elbes, M., ... & Alia, M. A. (2019, April). A review of natural language processing and machine learning tools used to analyze arabic social media. In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) (pp. 622-628). IEEE.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25.

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. Business Horizons, 63(1), 37-50.

Kieuvongngam, V., Tan, B., & Niu, Y. (2020). Automatic text summarization of covid19 medical research articles using bert and gpt-2. arXiv preprint arXiv:2006.01997.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. Digital Education Review, (33), 235-266.

Marie-Sainte, S. L., Alalyani, N., Alotaibi, S., Ghouzali, S., & Abunadi, I. (2018). Arabic natural language processing and machine learning-based systems. IEEE Access, 7, 7011-7020.

Maseer, Ziadoon Kamil, et al. "Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset." *IEEE access* 9 (2021): 22351-22370.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE access, 8, 131662-131682.

Mostafa, S. M. (2019). Imputing missing values using cumulative linear regression. CAAI Transactions on Intelligence Technology, 4(3), 182-200.

Nandy, A. (2018). Association Rule Mining with Eclat on A Malaysian Retail Store. International Journal of Research in Science and Technology, 32-49.

Nikitha, G. N., Chandana, C., Neelashree, N., Nisargapriya, J., & Vishwesh, J. (2020). Bank customer complaints analysis using natural language processing and data mining. International Journal of Progressive Research in Science and Engineering, 1(3), 22-25.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., ... & Habash, N. (2020, May). CAMeL tools: An open source python toolkit for Arabic natural language processing. In Proceedings of the 12th language resources and evaluation conference (pp. 7022-7032)

Obiedat, R., Al-Darras, D., Alzaghoul, E., & Harfoushi, O. (2021). Arabic AspectBased Sentiment Analysis: A Systematic Literature Review. IEEE Access.

Parlina, I., Arnol, M. Y., Febriati, N. A., Dewi, R., Wanto, A., & Lubis, M. R. (2019, August). Naive Bayes Algorithm Analysis to Determine the Percentage Level of visitors the Most Dominant Zoo Visit by Age Category. In Journal of Physics: Conference Series (Vol. 1255, No. 1, p. 012031). IOP Publishing.

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78.

Phan, Q. T., Wu, Y. K., & Phan, Q. D. (2021, September). An Overview of Data Preprocessing for Short-Term Wind Power Forecasting. In 2021 7th International Conference on Applied System Innovation (ICASI) (pp. 121-125). IEEE.

Puschmann, C., & Powell, A. (2018). Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. Social Media+ Society, 4(3), 2056305118797724.

Quarteroni, S. (2018). Natural language processing for industry. Informatik-Spektrum, 41(2), 105-112.

Ramaswamy, S., & DeClerck, N. (2018). Customer perception analysis using deep learning and NLP. Procedia Computer Science, 140, 170-178.

Ranganathan, G. (2021). A study to find facts behind preprocessing on deep learning algorithms. Journal of Innovative Image Processing (JIIP), 3(01), 66-74.

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Rezaii, N., Wolff, P., & Price, B. H. (2022). Natural language processing in psychiatry: the promises and perils of a transformative approach. The British Journal of Psychiatry, 220(5), 251-253.

Rygielski, C., Wang, J. C., & Yen, D. C. (2020). Data mining techniques for customer relationship management. Technology in society, 24(4), 483-502.

Salmi, N., & Rustam, Z. (2019, June). Naïve Bayes classifier models for predicting the colon cancer. In IOP Conference Series: Materials Science and Engineering (Vol. 546, No. 5, p. 052068). IOP Publishing

Sapountzi, A., & Psannis, K. E. (2020). Big data preprocessing: an application on online social networks. In Principles of Data Science (pp. 49-78). Springer, Cham.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. (2019). Challenges in Arabic natural language processing. In Computational linguistics, speech and image processing for arabic language (pp. 59-83).

Singh, V., Vijay, D., Akhtar, S. S., & Shrivastava, M. (2018, July). Named entity recognition for hindi-english code-mixed social media text. In Proceedings of the seventh named entities workshop (pp. 27-35).

Sudarsan, V., & Kumar, G. (2019). Voice call analytics using natural language processing. Int. J. Stat. Appl. Math, 4, 133-136.

Talukder, M. A. I., Abujar, S., Masum, A. K. M., Faisal, F., & Hossain, S. A. (2019, July). Bengali abstractive text summarization using sequence to sequence RNNs. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Teng, X., & Gong, Y. (2018, July). Research on application of machine learning in data mining. In IOP conference series: materials science and engineering (Vol. 392, No. 6, p. 062202). IOP Publishing.

Top 10 NLP Tools & Services in 2022. (2020, March 11). MonkeyLearn Blog. Retrieved December 27, 2022, from https://monkeylearn.com/blog/natural-languageprocessing-tools/

Top natural language processing (NLP) providers in 2022. Datamation. Retrieved December 28, 2022, from https://www.datamation.com/artificialintelligence/natural-language-processing-providers/

Turner, J. (2018). Robot rules: Regulating artificial intelligence. Springer. Vinnarasu, A., & Jose, D. V. (2019). Speech to text conversion and summarization for effective understanding and documentation. International Journal of Electrical and Computer Engineering, 9(5), 3642.

Wahdan, A., Salloum, S. A., & Shaalan, K. (2021, March). Text Classification of Arabic Text: Deep Learning in ANLP. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 95-103). Springer, Cham.

Walczak, S. (2019). Artificial neural networks. In Advanced methodologies and technologies in artificial intelligence, computer simulation, and humancomputer interaction (pp. 40-53). IGI global.

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., ... & Lei, M. (2019). Machine learning in materials science. InfoMat, 1(3), 338-358.

Win, K. M. N., Hnin, Z. Z., & Thaw, Y. M. K. K. (2020). REVIEW AND PERSPECTIVES OF NATURAL LANGUAGE PROCESSING FOR SPEECH RECOGNITION. International Journal Of All Research Writings, 1(10), 112-115.

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. Ieee Access, 7, 51522-51532.

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470.

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., ... & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. Journal of Evidence-Based Medicine, 13(1), 57-69.

Yoshimura, Y., Sobolevsky, S., Bautista Hobin, J. N., Ratti, C., & Blat, J. (2018). Urban association rules: uncovering linked trips for shopping behavior. Environment and Planning B: Urban Analytics and City Science, 45(2), 367-385.

Yu, J., Bohnet, B., & Poesio, M. (2020). Named entity recognition as dependency parsing. arXiv preprint arXiv:2005.07150.

Zephin LivingstonZephin Livingston is a content writer for eWeek, Livingston, Z., & Zephin Livingston is a content writer for eWeek. (2022, September 22). Top natural language processing companies 2022. eWEEK. Retrieved December 28, 2022, from https://www.eweek.com/big-data-and-analytics/natural-languageprocessing-companies/

Appendix: Code sections

```
#To check if the dataset is balanced
sns.countplot(x = 'sentiment', data = df, palette = 'deep')
```

```
<Axes: xlabel='sentiment', ylabel='count'>
```

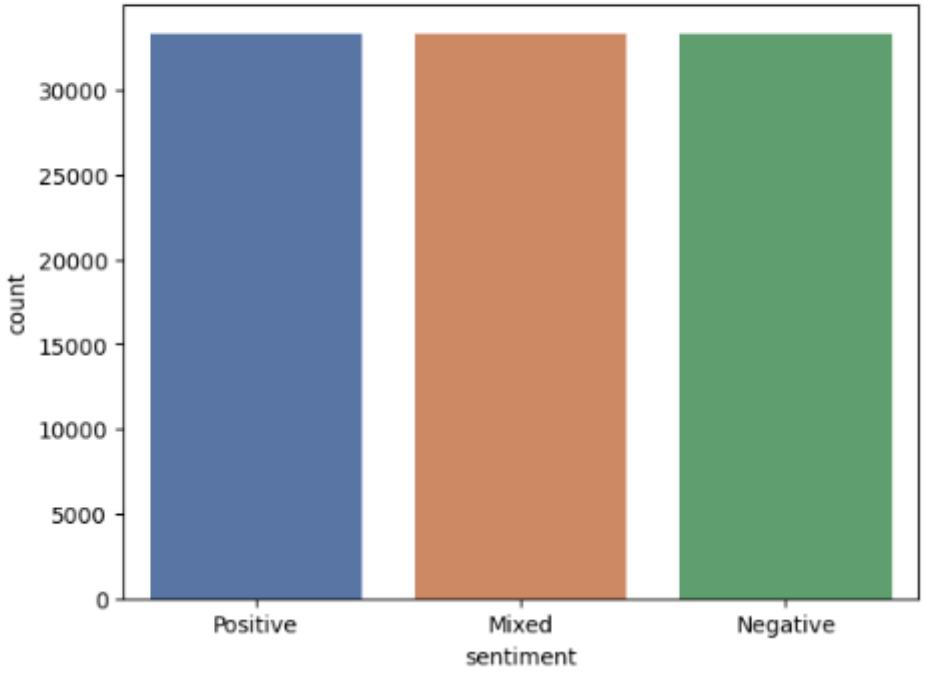


Figure 3 This count plot shows that the dataset is balanced

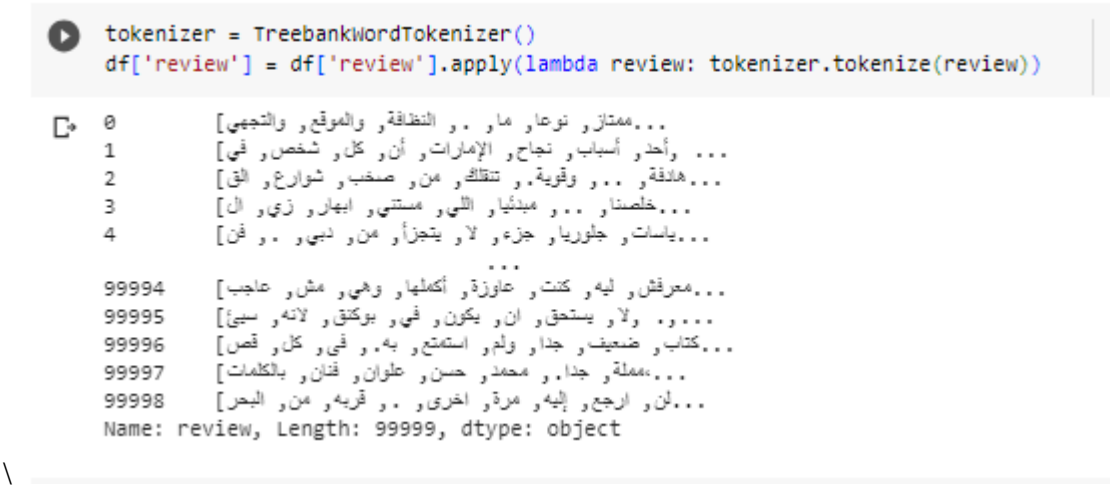
```
# to check if there is any null values
df.isnull().sum()
```

```
sentiment    0
review       0
dtype: int64
```

```
#to check if there is any duplicates
df.duplicated().sum()
```

```
0
```

Figure 4 This shows that the dataset has no duplicates or null values



\

Figure 5 This is the code that was used for tokenization

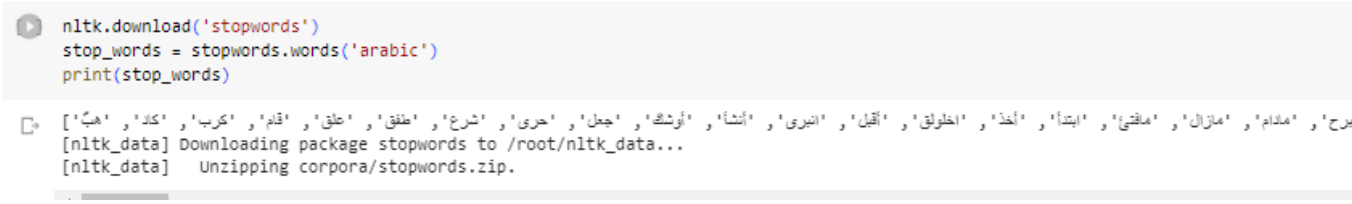


Figure 6 The code used to download the Arabic stop words from the NLTK library

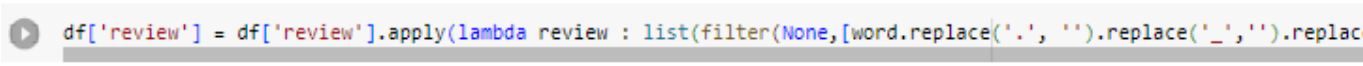


Figure 7 This is the code that was used to remove unwanted punctuation marks

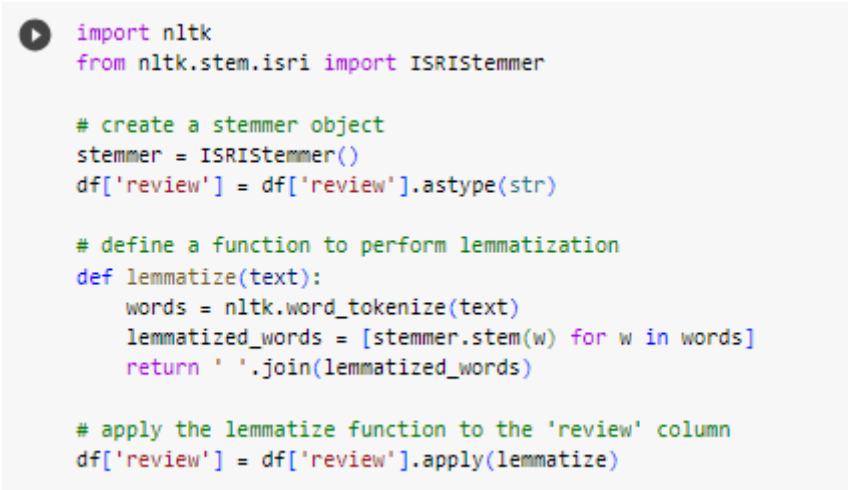


Figure 8 This is the code that was used for stemming

```python
# Split the data into train and test sets
x_train, x_test, y_train, y_test = train_test_split(df['review'], df['sentiment'], test_size=0.2)

# Create a pipeline that vectorizes the text data and trains a Naive Bayes classifier
vectorizer = CountVectorizer()
clf = MultinomialNB(alpha=1)
pipe = make_pipeline(vectorizer, clf)

# Perform 10-fold cross-validation on the training data
cv_scores = cross_val_score(pipe, x_train, y_train, cv=10, scoring='f1_macro')

# Train the pipeline on the entire training data
pipe.fit(x_train, y_train)

# Make predictions on the test data
y_pred = pipe.predict(x_test)

# Evaluate the performance of the classifier using F1 score
f1 = f1_score(y_test, y_pred, average='macro')

# Print the F1 score for the test data and cross-validation
print('F1 Score (Test):', (f1*100),"%")
print('Cross-Validation Mean F1 Score:', (np.mean(cv_scores)*100),"%")
print('Cross-Validation Standard Deviation:', (np.std(cv_scores)*100),"%")
```

```
F1 Score (Test): 81.98055876079111 %
Cross-Validation Mean F1 Score: 81.58867334691993 %
Cross-Validation Standard Deviation: 0.3947440216172387 %
```

Figure 9 This is the code that was used for multinomial Naive Bayes modelling with 10 folds of cross validation

```python
# Split the data into train and test sets
x_train, x_test, y_train, y_test = train_test_split(df['review'], df['sentiment'], test_size=0.2)

# Define the pipeline
vectorizer = TfidfVectorizer()
clf = DecisionTreeClassifier(random_state=42)
pipe = make_pipeline(vectorizer, clf)

# Perform 10-fold cross-validation on the training data
cv_scores = cross_val_score(pipe, x_train, y_train, cv=10, scoring='f1_macro')

# Train the model on the entire training data
pipe.fit(x_train, y_train)

# Predict the test set labels
y_pred = pipe.predict(x_test)

# Evaluate the F1 score
f1 = f1_score(y_test, y_pred, average='macro')

# Print the F1 score and cross-validation results
print("F1 Score:", f1*100,'%')
print('Cross-Validation Mean F1 Score:', (cv_scores.mean()*100),'%')
print('Cross-Validation Standard Deviation:', (cv_scores.std()*100),'%')
```

```
F1 Score: 74.98097850755275 %
Cross-Validation Mean F1 Score: 73.75367932159538 %
Cross-Validation Standard Deviation: 0.46927650166033325 %
```

Figure 10 This is the code that was used for decision tree classifier using 10 folds of cross validation

Declaration

I herewith declare that this report is in full accordance with the Plagiarism Guidelines of the Faculty of Management & Technology at the GUC.

Signature