PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

IBN KHALDOUN UNIVERSITY OF TIARET
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTMENT

# THESIS

**Submitted to the Faculty MATHEMATICS AND COMPUTER SCIENCE**

**Department of Computer science**

**To Obtain a Doctorate in Computer science**

**Presented by: Mrs. AOUMEUR Nour El Houda**

## TITLE

# THE PURPOSE OF DEEP LEARNING MODEL USING EMBEDDING TECHNIQUE IN ARABIC SENTIMENT ANALYSIS

**Board of Examiners:**

| | | |
|---|---|---|
| **Prof. MEBAREK Bendaoud** | **Ibn-Khaldoun Tiaret University** | **Chairman** |
| **Prof. Li Zhiyong** | **Hunan University** | **Supervisor** |
| **Dr. Mostefaoui Sid Ahmed Mokhtar** | **Ibn-Khaldoun Tiaret University** | **Co-Supervisor** |
| **Dr. AID Lahcen** | **Ibn-Khaldoun Tiaret University** | **Examiner** |
| **Dr. KHARROUBI Sahraoui** | **Ibn-Khaldoun Tiaret University** | **Examiner** |
| **Dr. HAMDANI Mostefa** | **Tissemsilt University** | **Examiner** |

**2023- 2024**

# Dedication

*I would like to thank God, for letting me through all the difficulties; I have experienced your guidance day by day.*

*I dedicate my thesis work*

*To my dear father, lovely mother and my sweet sister a special feeling of gratitude has never left my side.*

*You've been my biggest supporters throughout many years of studies and researches up until this day; you've encouraged me and stood by my side during the ups and downs, and given me confidence to keep going.*

*Thank you for your love, endless support and encouragement. I am truly blessed for having you in my life, without you I would not have the courage to keep going on this journey in the first place.*

*And To my dear Husband who has been a constant source of support and encouragement during the challenges of graduation and life and to pursue my dreams and finish my thesis.*

*Thank you*

# Acknowledgement

**I would like to first say a very big thank you and express my appreciation and gratitude to the rector of the IBN Khaldoun Tiaret University Professor BELGOUMANE Berrezoug for all the support, encouragement, and help, This Ph.D. would not have been achievable without his help.**

**I wish to thank most sincerely my Professor LI ZHIYONG for supervising this work. I am forever grateful to you for giving me the opportunity to pursue my interest in activity theory and for helping me so well to shape and express my thoughts.**

**I would like to acknowledge and give my warmest thanks to my co-supervisor Dr. MOSTEFAOUI Sid Ahmed Mokhtar who made this work possible. His guidance and advice carried me through all the stages of writing my thesis.**

**I would like to express my sincere appreciation and gratitude to the Professor MEBAREK Bendaoud for his support and encouragement. He helped me to arrange my search. I thank him for patiently reviewing my work.**

**I am also grateful to the board of examiners for their willingness to evaluate my work and to provide helpful comments and remarks:**

**Prof MEBAREK Bendaoud from Ibn-Khaldoun Tiaret University**

**Dr. AID Lahcen from Ibn-Khaldoun Tiaret University**

**Dr. KHARROUBI Sahraoui from Ibn-Khaldoun Tiaret University**

**Dr. HAMDANI Mostefa from Tissemsilt University**

**A special thanks to The Vice Rector Professor AIT AMAR Meziane Mohamed for his hospitality and assistance help.**

**Also, I am sincerely thankful to the Dr. LARABI Doyen of faculty mathematics and computer science of the university Ibn Khaldoun, Dr. OUARDANI Abderrahmane and Dr. CHADLI Abdelhafid for their support and hospitality.**

**Thanks to everyone who helped me from the university administration workers.**

# Abstract

Social media, widely used by Internet users to express their opinions on a given topic, has become one of the main sources of information for analysts. Sentiment analysis (SA) is a growing area of research of natural language processing (NLP) and machine learning (ML) tools to identify and label opinion text. Sentiment analysis is an important task in fields related to data analysis and information mining.

In this study, the books of the most famous Arab authors were read and each sentence was manually extracted and labeled. This research aimed to generate a new Classical Arabic dataset (CASAD). In addition, feature extraction from these datasets is generated using word embedding techniques equivalent to Word2vec, which can extract deep relationships representing features of formal Arabic languages. Some machine learning techniques, such as support vector machine (SVM), logistic regression (LR), naive bayes (NB), K-nearest neighbor (KNN), latent Dirichlet allocation (LDA), and classification tree and regression are used to evaluate the features for classical Arabic (CART). In addition, statistical techniques such as validation and reliability are used to evaluate the labels of this dataset. Finally, using six machine learning algorithms for 10-fold cross-validation, our tests evaluated the classification rate of the feature extraction matrix into two and three classes, and the results showed that the Logistic regression with Word2Vec was the most effective in predicting the occurrence of polarizing topics.



**Keywords**: Arabic Sentiment Analysis, Classical Arabic, Feature Extraction, Machine Learning, Word2Vec, Deep learning.



الملخص

وسائل التواصل الاجتماعي، التي يستخدمها مستخدمو الإنترنت بشكل واسع للتعبير عن آرائهم حول مواضيع معينة، أصبحت واحدة من المصادر الرئيسية للمعلومات لدى المحللين. تحليل المشاعر (SA) هو مجال بحث نامٍ في مجال معالجة اللغة الطبيعية (NLP) وأدوات تعلم الآلة (ML) لتحديد وتصنيف النصوص التي تحمل آراءً. إن تحليل المشاعر مهم في المجالات المتعلقة بتحليل البيانات وتنقيب المعلومات.

في هذه الدراسة، تم قراءة كتب أشهر الكتَّاب العرب، وتم استخراج كل جملة يدويًا وتسميتها. كان هدف هذا البحث إنشاء مجموعة بيانات عربية كلاسيكية جديدة (CASAD). بالإضافة إلى ذلك، يتم إنشاء استخراج الميزات من هذه المجموعات باستخدام تقنيات التضمين الكلموية المعادلة لـWord2vec ، التي يمكنها استخراج العلاقات العميقة التي تمثل ميزات اللغات العربية الفصحى. يتم استخدام بعض تقنيات تعلم الآلة، مثل آلة الدعم

النوعي(SVM) ، والتحليل اللوجستي (LR) ، والبايز الساذج (NB) ، وأقرب الجيران (KNN) ، وتوزيع لاتنت ديريكليه(LDA) ، وشجرة التصنيف والتحليل (CART) لتقييم الميزات للعربية الكلاسيكية. بالإضافة إلى ذلك، يتم استخدام تقنيات إحصائية مثل التحقق والموثوقية لتقييم تسميات هذه المجموعة البيانات. وأخيرًا، باستخدام ستة خوارزميات لتعلم الآلة في التحقق المتقاطع لعشر مرات، قامت اختباراتنا بتقييم معدل التصنيف لمصفوفة استخراج الميزات في فئتين وثلاث فئات، وأظهرت النتائج أن التحليل اللوجستي بتقنية Word2Vec كان الأكثر فعالية في توقع حدوث مواضيع تحمل طابع القطبية.

,الكلمات المفتاحية: تحليل المشاعر العربية، العربية الكلاسيكية، استخراج الميزات، تعلم الآلة، ، التعلم العميق Word2Vec

.

# List of figures:

# List of tables:

## List of Abbreviations

**SA:** Sentiment Analysis
**POS:** Part of Speech
**ABSA:** Aspect-based sentiment analysis
**ML:** Machine Learning
**SVM:** Support Vector Machine
**KNN:** K-Nearest Neighbor
**TR:** Training Dataset
**TS:** Test Set
**TF-IDF:** Term Frequency-Inverse Document Frequency
**NB:** Naive Bayes
**HL-NBC:** Hybrid dictionary-Naïve Bayesian Classifier
**LDA :** Latent Dirichlet Allocation
**LR :**Logistic Regression
**CART:** Classification and Regression Trees
**TP:** True Positive
**TN:** True Negative
**FP:** False positive
**FN:** False Negative
**AUC:** Area Under the Curve
**ROC:** Receiver Operating Characteristic
**TPR:** True Positive Rate
**FPR:** False Positive Rate
**DL:** Deep learning
**AI:** Artificial Intelligence
**DNN:** Deep Neural Networks
**RNN:** Recurrent Neural Networks
**DBN:** Deep Belief Networks
**ANN:** Artificial Neural Networks
**LSTM:** Long-Term Memory
**GRU:** Gated Recurrent Units
**ANN:** Artificial Neural Network
**NLP :**Natural Language Processing
**NER:** Named Entity Recognition
**NLU:** Natural Language Understanding
**NLG:** Natural Language Generation
**CBOW:** Continuous Bag of Words
**SG:** Skip-gram
**BoW:** Bag of Words
**PV-DBOW:** Distributed Bag-of-Words Paragraph Vector
**PV-DM:** Distributed Memory Model of ParagraphVector
**TF:**Term Frequency
**IDF:** Inverse Document Frequency
**DF:** Document Frequency
**BERT:** Bidirectional Encoder Representations from Transformer
**PN:** Producer Node
**CN:** Consumer Node
**ASA:** Arabic Sentiment Analysis

**CA:** Classical Arabic
**MSA:** Modern Standard Arabic
**DA:** Dialect Arabic
**MT:** Machine Translation
**WSD:** Word Sense Disambiguation
**OCA:** Corpus for Arabic
**HARD:** Hotel Arabic Reviews Dataset
**LABR:** Large-scale Arabic Book Review
**BRAD**: Book Reviews in Arabic Dataset
**CASAD:** Classical Arabic Sentiment Analysis Dataset
**ASTD:** Arabic Sentiment Tweets Dataset
**MASC :** Multi domain Arabic Sentiment Corpus
**TASC:** Tunisian Sentiment Analysis Corpus

Table of contents

# General Introduction

# General introduction

There are a lot of things on the Internet that can help people decide how to solve their problems. People often try to find other people's opinions online about a business, a country they want to visit and live in, or a movie they want to watch at the cinema. This makes it possible to collect and analyze many thoughts and opinions on certain topics on these pages. Therefore, the language filter needs to be changed. It's better for people to directly express their thoughts and opinions on a topic rather than searching for and reading reviews to get a final opinion. Sentiment analysis has proven to be very useful for various natural language processing (NLP) tasks, such as response systems and information extraction. The goal is to extract information relevant to a specific topic or user need. For example, people today tend to use the Internet to spread their thoughts and ideas on topics or issues through forums or other social networks. Some of these thoughts are positive, while others are more violent in nature and content. This idea of spreading emotions online opens up a new field of text analysis and expands the scope of research from traditional fact- and information-focused views of text to emotion-aware applications.

The Arabic language boasts an impressive number of speakers, with over 220 million individuals utilizing it in more than 57 countries worldwide. Unlike European languages like English, Arabic boasts a more intricate morphological structure. However, Arabic presents its own set of unique challenges that require specialized attention. Due to its complexity and the limited resources available, researchers find Arabic natural language processing an alluring topic. Consequently, it is essential to address the importance of this language. In Arabic, the fundamental tools of NLP such as the morphological analyzer, part of speech tagger, and syntactical parser are being heavily researched and developed.

Sentiment analysis is a relatively new topic in Arabic compared to other languages such as English. Research in this area isn't going anywhere anytime soon. This research contributes to the field of Arabic sentiment analysis. We discuss the interesting and challenging aspects of this language in the context of its history, strategic importance to the country and its culture. Furthermore, language limitations in this field begin with resources and end with tools. This work creates a corpus and can be used by other academic researchers to advance the field. We identify and evaluate the performance of methods that can be used. Feature extraction in these datasets is generated using word embedding techniques that extract deep relationships that represent formal Arabic language features. The corpus also promotes research on Arabic sentiment analysis. Although this study is limited to Arabic, this approach and method can be applied to other low-resource languages.

Natural language processing for Arabic has improved significantly over the past two decades, but it is still considered underpowered compared to English. Therefore, the research community can benefit from additional resources such as annotated corpora and sentiment dictionaries. There are two main obstacles to effective classification:

1. Limited number and accuracy of tools such as morphological parsers, part-of-speech (POS) taggers, and stemmers.
2. These include the lack of tagged corpora available for experiments, and the limited research being done in this area compared to what has been done for English.

The research objectives proposed in this study are:

- A new Classical Arabic Sentiment Analysis Dataset (CASAD) used to determine the polarity of feature extractions generated using word embedding techniques to extract deep relationships representing formal Arabic features. This corpus contains sentences with emotional tendencies.
- Identify classical techniques used in sentiment analysis, focusing on the Arabic language to extract aspects of author reviews written in the Arabic language.
- Compare the performance of machine learning classifiers such as Support Vector Machine (SVM) , Naïve Bayes (NB) , and other classifiers in the Arabic sentiment with the same methods that are used in English, or does Arabic require other methods .
- Used to extract aspects of reviews written by authors in Arabic. Specifically, we worked with Arabic authors reviews by evaluating the performance of existing aspect extraction techniques that have been used for other languages, identifying methods that work best and apply it for sentiment analysis purposes.

- Achieving these goals is expected to result in a coherent sentiment analysis framework proposed in this study, which aims to improve sentiment classification performance and provide detailed.

The thesis consists of four chapters including a general introduction. The remaining chapters of this thesis are as follows:

- **Chapter 1** presents a background and literature review of opinion mining and sentiment analysis levels, challenges, and approaches that show the importance of this domain; We consider how sentiment analysis capabilities are configured into applications that can play an important role in social media.
- **Chapter 2** consists of two parts; first, machine learning methods for sentiment analysis and second, deep learning methods for sentiment analysis; both of them introduce the approaches to sentiment analysis, applications, methods, algorithms, and performance that can be used.
- **Chapter 3** is a summary of the characteristics of the Arabic language and provides the context and challenges needed to improve Arabic sentiment analysis; it deals with the writing, syntax and morphology of Arabic. The chapter describes the techniques used to preprocess Arabic text using NLP and the Arabic sentiment analysis methods currently being built in this research.
- **Chapter 4** deals with the evaluation of the approach framework proposed in this study and the published article. It provides detailed analysis of the results and compares them with the ML result classifier. The final section concludes with suggestions for future research and improvement directions.

To conclude, the general conclusion section provides a summary of the main points discussed throughout the thesis in order to answer the research questions based on the results obtained and then provide recommendations for further research in the same field of study, and suggestions for further research.

# Chapter 1

# Background and Literature Review

# I. Chapter 1 Background & Literature Review

## I.1. Introduction:

The Internet holds great importance in our everyday existence. There are lots of different things on the Internet that give us a lot of information. Some of it is organized and easy to understand, but some of it is not. The data from social media is important because it holds a lot of people's opinions. Because this information is important in different areas like politics, business, and personal matters, it would be helpful to have the ability to analyze opinions automatically. Analyzing these posts can provide insights into users' emotions and thoughts related to particular things, such as subjects, occurrences, topics or issues, events, products, individuals, and others. We use sentiment analysis for this.

This chapter gives basic information about important ideas connected to the research described in this thesis. This includes: sentiment analysis and well-known methods previously used for SA.

## I.2. Sentiment Analysis (SA)

The word “sentiment” is defined as an attitude, thought or judgment prompted by feeling in [1] . In the literature, several terms such as opinion, textual subjectivity, feeling, emotion, evaluation, belief, and conjecture are used interchangeably to refer to this concept. All of these terms refer to a private state in which objective observation or verification [2], Because of the variety of terms, the computation analysis areas are known as opinion mining, sentiment analysis, or subjectivity analysis [3] , it is also defined as "a particular view or idea." "Opinions" and "Feelings" The word "opinions" is usually used to mean "a mentally formed view, judgment, or assessment of a particular matter." This can be confusing for readers and newcomers to the field. So sentiment analysis and opinion mining seem to refer to the same research field. There are several other names and slightly different tasks, such as opinion extraction, sentiment mining, subjectivity analysis, effect analysis, emotion analysis, review mining, and so on [4]. In industry, the term sentiment analysis is far more commonly used, while in academia, both sentiment analysis and opinion mining are commonly used. Therefore, the term sentiment analysis and opinion mining are used synonymously in this work.

## Chapter 1: Background and Literature Review

In the field of information gathering, sentiment analysis involves computers identifying opinions from texts and classifying them as positive or negative. With so much data, it becomes critical to ensure that there is enough data for analysis and utilization, and this is a major task of today's data warehouses and relational databases. Emotions expressed in comments, feedback, or criticism serve different purposes and can be categorized by polarity. Use polarity to determine if a review is overall positive or negative. As an example: A positive mood is expressed in subjective sentences:"I loved that romantic movie." The sentiment threshold for the word "I loved it" indicates that the phrase expresses positive feelings about a romantic movie. As a result, the word "loved" has a positive numerical threshold. Subjective Negative Mood: "The Greatest is a flop movie" (predefined phrase) expresses negative feelings about the movie "The Greatest" as defined by the emotional threshold of the word "flop". As a result, the word "flop" has a negative numeric threshold.

Textual information generally falls into two categories. The first type of information is factual information, which simply contains factual or objective descriptions of entities or events. The second type of information is subjective information, reflecting the creator's actual feelings and opinions about the entity or event. It has become a very big task for the data warehouse and the relational database due to the enormous amount of data, as it becomes very crucial to analyze and use sufficient data for their functionality [5].Velocity, volume, intensity are the three main ingredients of big data analytics. A large amount of data collected simultaneously from various sources is called a volume. The velocity of the data that allows us to measure it is called its intensity [6]. Text, audio, video, image data, in addition to structured and unstructured data from various sources are classified as data types [7].

Sentiment analysis is a technique for capturing people's sentiment (feelings or opinions) about a specific topic. This field relates to machine learning, natural language processing, and computational linguistics. In other words, it usually captures people's emotions and tries to extract them from the text. Subjective analysis, opinion mining, review mining, and evaluative extraction are all labels for SA in the literature. The first step in sentiment analysis is to determine or classify the content of a text as subjective or objective. This is called subjective classification. The second task is to analyze subjective texts to determine which mood polarities they contain. The strength of this polarity depends on your position. For example, user reviews of a product should be classified as positive or negative for the target group. This indicates binary polarity. The task becomes even more difficult when the polarity extends to more than two elements, for example to the neutral.

In this case, the task would consist of a series of classification challenges. The classification process can be performed at different levels of text such as terms, phrases, sentences and documents. Normally the output of each level is used as input for the next level. For example, extend document classification output sentence ratings. This is also a kind of mood with the emotion of [8].

Sentiment analysis is about emotional goals or their discovery. Most sentiment analysis work focuses on identifying sentiment around a specific topic or goal, such as user reviews of a movie or product. Since such reviews are intended to relate to a specific product, the subject matter of such a review can be easily determined. With unknown goals, such as a feature-based sentiment analysis, it becomes even more difficult. It's difficult to find out what product features users have written about and get their opinions on them. Therefore, a research is first performed to determine which features the user is described using a feature extraction approach [9].

While sentiment analysis and opinion mining have received a lot of attention from researchers and marketers lately, the potential interest in sentiment analysis persists today. Much of the early work on textual information processing focused on collecting and retrieving fact, such as information retrieval, text classification and text clustering. One of the main reasons for the lack of sentiment analysis research is the lack of opinion-forming texts before the advent of the World Wide Web. Ever since technological advances made the internet popular, people have been able to express their opinions and feelings by posting reviews of products and services online. This online word-of-mouth behavior represents a new and measurable source of information with many practical applications. In addition, three other factors that have contributed to the surge in sentiment analysis research are identified in [3].The rise of machine learning methods in natural language processing and information retrieval. As a result of the growth of the World Wide Web, especially review summary sites, datasets have become available for training machine learning algorithms and an appreciation for fascinating intellectual challenges and opportunities.

## Chapter 1: Background and Literature Review

The figurative nature of language impacts sentiment analysis and is frequently taken advantage through the use of literary features such as sarcasm and irony. Sentiment analysis is a challenging yet fascinating research area due to its complex composition of multiple tasks.

The explosion of Web 2.0 platforms such as blogs, Twitter, Facebook, and various other types of social media provides unprecedented channels and opportunities for individuals to share their opinions and brand experiences about products or services. In addition, companies can change their marketing strategies by monitoring and analyzing social media. However, finding and monitoring sources of opinion on the World Wide Web can be a daunting task due to the multitude of different sources of information, including online forums, discussion groups, and blogs. In addition, each source may contain large amounts of user-generated content that shows feelings and emotions. This work is known as opinion mining. Sentiment analysis in natural language processing (NLP) encompasses various aspects of how information about emotions, attitudes, perspectives, and social identities is conveyed in language [10] .We also conclude that the purpose of sentiment analysis is to determine consumer attitudes and opinions about products through automated analysis of product review texts. According to [11] an ideal opinion-mining tool would process a set of search results for a specific product, create a list of product attributes (quality, features, etc.), and generate opinions (poor, mixed, etc.) about each of those attributes.

To start generating more accurate insights, sentiment analysis solutions use consistent criteria. A machine learning model, for example, can be trained to recognize two aspects with opposing sentiments. It would take the overall sentiment as neutral, but it would also keep track of the specifics. Hundreds of megabytes of text can be analyzed by sentiment analysis algorithms in a matter of minutes. Now that you don't have to manually analyze data in spreadsheets, you can use your time on more worthwhile pursuits. Sentiment analysis only provides a signal. However, if you quickly and easily receive this signal, you will have time to develop the ideal plan. Algorithms and methods for sentiment analysis are constantly improving. By giving them higher-quality, more diverse training data, they become better. Researchers are also creating new algorithms to improve the use of this data. Thematic keeps track of your results and evaluates any errors. We supplement existing training data with additional, more precise information as necessary. As a result, sentiment analysis is improving precision and providing insights that are more focused.

Opinions are personal statements that reflect a person's feelings, thoughts, and beliefs about defined a specific topic in [12]. According to the authors, each opinion has a source, which is commonly referred to as the opinion holder. The item or topic that prompts the opinion keeper to express their opinion is another aspect of opinions. The authors describe that the final component of opinions is the opinion holder's view, which is influenced by their sentiments about the object. According to [13] an opinion is described as a quadruple of four essential components: $(\boldsymbol{t}, \boldsymbol{s}, \boldsymbol{h}, \boldsymbol{dt})$ where $\boldsymbol{t}$ represents the sentiment or opinion target, $\boldsymbol{s}$ represents the sentiment towards the target, $\boldsymbol{h}$ represents the opinion holder, and $\boldsymbol{dt}$ represents the date and timestamp of the opinion.

Sentiment analysis (SA) is widely used through research areas and practitioner domains, and is often regarded as "omnipresent as a concept," but it is fraught with analytical, domain-specificity, methodology, and interpretation challenges [14]. There have been many different definitions of SA. It has been defined as "a generic name for a great amount of opinion and actually effect nece ssary activities [15] as "a research area that aims at comprehending the fundamental sentiment of unstructured content" [16] , as "an active area of study to demonstrate feelings and to instantly explore the sentiments expressed inside the text" [17]. Many studies use the terms as synonyms, with no regard for their underlying meanings. For example, SA has been defined as "the gathering of people's perspectives on any event occurring in real life understanding people's emotions are extremely important" [18].

### I.2.1. Sentiment Analysis Level

Sentiment analysis can be performed at several granularity levels, including documents, sentences, words or phrases, aspects, and users. The following paragraphs explain the various aspects that can be performed on three primary levels, as shown in Figure 1.

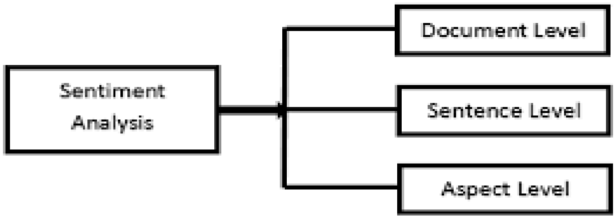


**Figure 1: Level of sentiment analysis.**

### I.2.1.1. Document level:

The main purpose of document-level sentiment analysis is to classify the entire document and determine whether it is positive or negative. This assumes that the document only conveys a feeling about one of its objects. Therefore, it is not suitable for documents that evaluate multiple entities or aspects. This has therefore been criticized as impractical as a given document can contain multiple emotions. For example, a sentiment analysis system classifies the overall polarity of customer reviews for a particular product. Sentiment analysis typically produces only two (positive and negative) or three results, so this level of sentiment classification does not assume that a single result will be reflected in a single document. However, multiple opinions are often expressed in one of his documents. Therefore, it does not apply to documents expressing opinions on multiple products. Many researchers have performed document-level sentiment analysis [19]. Your main concern is the isolation of positive sentences.

### I.2.1.2. Sentence level:

Sentence-level analysis examines sentences that express a single opinion and attempts to define their direction. Each sentence is assumed to convey a single emotion. This assumption does not apply to every sentence in the text. In fact, there are many sentences that lack clear meaning. It is important to distinguish between subjective and non-subjective sentences, since non-subjective sentences provide no information to the classifier [20].

Sentence-level sentiment analysis determines whether each sentence expresses a neutral, positive, or negative opinion [21]. There is no fundamental difference between a document-level sentiment analysis and a sentence-level sentiment analysis, since a sentence is just a short document. This usually involves two subtasks. The first is to determine whether the text is subjective or objective. The second, if subjective, determines whether the sentence expresses a positive or negative opinion [22].

This level relates to the subjective classification [23], used to distinguish between subjective sentences expressing feelings or opinions and objective sentences expressing facts. Subjective classification is very important as it filters out sentences that do not contain an opinion. Sentence-level sentiment classification assumes that a single sentence expresses a single opinion of a single opinion-maker [24], On the other hand, it is established that a single sentence can contain not only several opinions, but also subjective and factual sentences. It is therefore important to identify factual statements and emotional strengths.

A value of "neutral" usually indicates objective or non-opinions writing. Also note that subjectivity is not synonymous with feeling. Objective sentences like this that imply an opinion are classified as opinionated sentences. Compound sentences do not qualify for this sentence-level classification, even if they are comparable or contain grouped opinions about different aspects of an object [25].

A subjective statement is a statement that expresses an opinion. Separating statements containing opinions from statements containing facts are called perception of subjectivity. Labeling subjectivity prevents misleading and irrelevant sentences from influencing sentiment classifiers, even in idiosyncratic texts [26] .Therefore, subjective classification is used to improve the performance of sentence-level sentiment classifiers [27], [28] . When different opinions need to be incorporated into the document, the sentence level is preferred over the document level. Texts are made up of different types of sentences, each with their own characteristics and which can be processed and classified differently, for example a conditional and comparative sentences. It is argued that there is no single taxonomy or strategy that works for all types of sentences or even entire texts. As a result, combining different strategies based on different sentence types improves classification accuracy. In general, opinion holders can express overall positive or negative opinions in the commented text or reviews. However, the author of the review can mention both good and bad aspects of the object.

#### I.2.1.3. Aspect level:

Categorizing opinions at the document or sentence level is useful in many cases, but it is not possible to identify emotional goals or relate opinions to those goals, making it impossible to provide the necessary detail for applications. Affirmative writing about an object does not imply that the author agrees with all aspects of the subject. Sentiment has been often an intermediate step because, in addition to classification at the sentence level, it is more meaningful to know which features or entities of an object the opinion refer to. Therefore, we need an aspect level with more detailed analysis.

The aspect level is also referred to as the entity level or feature level in some studies [29]. The aspect level of sentiment analysis focuses on opinions rather than document structures such as paragraphs, sentences, and phrases. It is not enough to merely create a polarity of opinions. It is also necessary to identify the subject of the opinion.

The task of sentiment analysis at the aspect level can be divided into two parts: Aspect extraction and aspect mood classification [4]. The aspect extraction task aims to extract the aspects in which opinions are expressed, so it can also be considered an information extraction task. For example: "This Samsung S6 has a great screen, but the battery life is too short." The entity "Samsung S6" represents aspects of the entity "screen" and "battery life". Aspect extraction is finding common nouns or noun phrases that are defined as aspects. Aspects of the text are then classified as positive, negative, or neutral [30].

However, aspect-level sentiment analysis can still cause problems. Most studies are based on previously specified aspects using keywords [31] and propose dictionary-based methods for aspect analysis, but assume that the aspects are already known. According to [4] , existing algorithms cannot process complex sentences, so accuracy at the emotional level is still considered low. Therefore, minor-level sentiment analysis is more difficult than document-level or sentence-level classification. In [32] proposes a framework for classifying aspects, but the aspects are predefined before classification and no results or assessments are available.

In order to analyze product ratings in more detail, a number of statistical models have recently been developed to determine customer opinions on specific product characteristics [33]. This is called aspect-level sentiment analysis. This is the process of determining the sentiments of corresponding opinions on relevant aspects of the product being evaluated [34]. This assumes that all opinions are fundamentally directed towards a specific topic/subject and none of the above frames precisely and consistently aim at that. For example, in a movie review, aspects extracted may include music, actors, lighting, and so on. When clients write about films, they comment on various aspects such as the choice of actors and music.

Furthermore, some researchers [35], [28] have added two additional levels to sentiment analysis:

- Comparative sentiment analysis aims to identify sentences containing comparative opinions and to extract positive entities from each opinion.
- Acquisition of a sentiment lexicon: A manual approach is usually not possible since each domain requires its own lexicon, which would take an enormous amount of time.

A dictionary-based approach, on the other hand, is more effective as it allows us to start with a small number of sentiment words appropriate to the target domain. This sentence can be expanded with synonyms and antonyms.

### I.2.2. Type of Sentiment Analysis:

It is important to understand different sentiment analysis techniques. Sentiment analysis has several uses, as shown in Figure 2.


**Figure 2: Types of Sentiment Analysis.**

1. Aspect-level sentiment analysis aims to identify and aggregate sentiments about a document or an entity mentioned in an aspect of a document. A detailed overview of the current state of the art is given and it can be seen that significant progress has already been made in identifying both the target, which may be itself or a part of it, and the corresponding emotion. Aspect-level sentiment analysis provides extremely detailed sentiment information that can be used in a variety of applications [36]. Aspect-level sentiment analysis is increasingly used in the design and development of various real-world applications. For example, manufacturers can identify which parts or aspects of their products attract, retain or even enhance consumers [37]. Generally, in reviews, users discuss multiple aspects of a given entity and express different feelings about each, making this level more appropriate for real-world applications [38] , [39]. Aspect-based analysis focuses on specific characteristics of a product or service. For example, laptop manufacturers ask users about their experiences with touchpad's, keyboards, and graphics. Use sentiment analysis tools to associate customer intent with hardware-related keywords.
2. Emotion detection involves analyzing a person's psychological state while they are composing the text. Emotion detection is a more complicated method than sentiment analysis because it goes beyond simply categorizing data. The sentiment analysis models in this approach attempt to convey different emotions, such as joy, anger, sadness, and regret, through the person's choice of words. Emotion detection detects specific emotions as opposed to positivity and negativity.
3. Intent-based analytics recognize both the action and the sentiment behind the text. It focuses on specific characteristics of a product or service. For example, the online comments expressing dissatisfaction with battery replacements prompt customer service to reach out to customers to resolve specific issues, and laptop manufacturers urge customers to seek advice on sound, graphics, keyboards, and touch pads. Use sentiment analysis tools to associate customer intent with hardware-related keywords.
4. A finer-grained sentiment analysis provides a finer level of polarity by further categorizing it, typically from very positive to very negative. This corresponds to the opinion of a 5-star rating system. Fine-grain sentiment analysis related

to the classification of text intent into different emotional levels. In general, the process involves rating user sentiment on a scale of 0 to 5 , with each equal segment expressing very positive, positive, neutral, negative, and very negative sentiment.

### I.2.3. Sentiment Analysis Methods:

Machine learning has influenced the growth of SA technology due to its superior performance in many areas. In particular, it is directly related to the polarity classification task and proposes learning techniques that accurately identify sentiments of opinion. The methods of these algorithms are:

- **Data collection**: Developing methods for extracting emotions from texts requires large amounts of data. This information is typically obtained from social networking sites, blogs, web scraping, social media, news channels, e-commerce sites, forums, weblogs and other websites or platforms. The first stage of sentiment analysis is data collection. Depending on the sentiment analysis results, text data can be combined with other data types such as video, audio, and location information. Primary data collection sources include:
  - ✓ **Social Media:** Information collected through social networks is referred to as social data. Users interact with your product by accessing, publishing, and sharing information. Social media is used as an interactive source of data in scientific research on people, groups and behavior. It refers to web or mobile-based internet applications that allow users to create, access, and share user-generated content.
  - ✓ **Forums:** Users can use the message boards to explain a range of subjects, swap ideas and opinions, and text for help. Because user-generated content is dynamic, forums are a fascinating source for sentiment analysis. Researchers can also conduct sentiment analysis on a particular domain.
  - ✓ **Weblog:** A short weblog can consist of passages containing articles, information, personal journal entries or links. Posts are edited and organized in chronological order, with the most recent entries displayed first. A useful tool for conducting sentiment analysis for many companies is a blog.

- ✓ **Web sites for electronic commerce:** These are sites where customers can comment on and rate a specific company or organization. In this case, websites that are not specifically reviewed sites, like e-commerce sites that feature product reviews or professional review sites, have millions of reviews.

- **Sentiment labeling:** The algorithm requires labeled data (the class) to understand correlations between the input and the output and to learn how the class is distributed. The class here could be multiple or binary (positive or negative) (positive, neutral, negative, etc.). This process is open to the person who holds the opinion. If this is not the case, a crowd sourcing tag must be used to evaluate the degree of disagreement among various experts.
- **Feature selection:** The fact that the data in these processes is unstructured is one of their most crucial features. The most important words in each polarity or class are featured. In the past, metrics like the term frequency-inverse document frequency have been used to select the most pertinent terms from a collection of documents (TFIDF). Using text mining techniques like lemmatization, stemming, part-of-speech, tokenization, etc., terms and words are also changed.

It is important to remember that relevant feature must be found in the data set before creating a classification model. This allows the words to be decoded and added to the feature vector during model training. The method is known as a "uni-gram" when only one word is taken into account, a "bi-gram" when only two words are taken into account, and a "tri-gram" when three words are taken into account. Bi-gram and uni-gram together are useful for analysis, and the context feature aids in obtaining the most accurate results.

- ✓ **Practical features**: are features that focus more on usage than on theoretical underpinnings. Pragmatics in linguistics and related fields studies the relationship between context and perception. The study of phenomena such as entailment, speech act, association, and discourse is called pragmatics.
- ✓ **Emojis:** are facial expressions used in sentiment analysis to express emotions. A wide range of human emotions is represented by various emoticons. Emojis help convey a person's tone when composing sentences and aid in sentiment

analysis. Replace the emoji with its meaning. Ratings show different emotions such as joy, sadness and anger.

- ✓ **Punctuation marks:** also known as exclamation marks are used to emphasize the strength of a positive or negative remark. Other punctuation marks include the apostrophe and the question mark.
- ✓ **Slang expressions:** These are often used to make statements more humorous. Given the nature of opinion tweets, it's safe to assume that the slang in the text indicates sentiment analysis. Replace the slang with its meaning.
- ➢ **Classification problem:** On the one hand, machine learning approaches can be used to study sentiment analysis. If an opinion is evaluated, it is subject to monitoring. The goal of this problem is to create a model that uses selected features to classify new unlabeled opinions. Decision trees, linear classifiers (SVM and neural networks), association rules and probabilistic classifiers are the main classification methods (naïve Bayes, Bayesian networks and maximum entropy). If neither the opinion owner nor other users provide feedback, the issue is considered unattended. The goal in this case is to determine the polarity of the text using linguistic rules or heuristics.

The main drawback of sentiment analysis methods would be that they heavily rely on the learning strategy and are content-dependent. Using text from a particular domain, such as movie reviews, restaurant reviews, product sentiments, etc., these methods are created from scratch. As a result, one of these methods will perform poorly when tested in a different domain from where it was trained. Words can, in fact, be used in various contexts to signify opposing polarities. One of the SA community's top concerns and an actual challenge is this lack of generalization.

In addition, one of the requirements for developing a polarity classification SAM is to create a specific corpus corresponding to the problem being addressed. This method is laborious as each document, sentence or phrase has to be tagged individually. And make sure that not just one expert, but several experts agrees on the labeling.

### I.2.4. Sentiment Analysis Challenges:

In natural language processing, sentiment analysis or classification is considered a subset of text classification. Sentiment analysis has a small number of classes, but the process of sentiment classification is more difficult than traditional topic text classification [3]. Topic text classification uses keywords for classification, but this does not always work well for sentiment analysis [40].

The problem of training set accuracy is usually at the root of the difficulty of sentiment analysis. Systems often struggle with objectivity and impartial comments, and are often misunderstood. Recognizing moods can also be difficult if the system cannot understand the context and tone; Questionnaire results or responses to questionnaires, etc., where the circumstances are not specified, Such as "Nothing" or "all" can be interpreted either positively or negatively, depending on the question. Like sarcasm, sarcasm often leads to mislabeling of emotions and cannot be trained.

Computers have a very difficult time deciphering the meaning of sarcastic phrases. Consider the following statement: *"Yeah, that's great."* I didn't receive my order for another three weeks. If you use the word *"great,"* the computer will rate the experience as positive unless it thoroughly analyzes the sentence and understands the scenario. Negation is the process of using negative words to reverse the meaning of a sentence. For example, I can't say that the subscription was expensive. Such sentences can be difficult for sentiment analysis algorithms to understand, especially when the negative spans two sentences, such as *"I thought the subscription was affordable, Not at all."*

Computer programs struggle when encountering emojis and irrelevant data. To avoid incorrect labeling of texts, special attention must be paid to training models with emojis and neutral data. It is possible for people to make conflicting claims. Most reviews contain both positive and negative feedback, which can be managed somewhat by examining each sentence individually. But the more informal the medium, the more likely people is to mix up different viewpoints in one sentence, making it harder for a computer to parse them.

When more than one emotion is expressed in a sentence, this is known as multi-polarity. The following sentence might appear in a product review: "I'm happy with the sturdy build but not impressed with the color." The software finds it challenging to determine the underlying sentiment. Each entity and its corresponding emotion must be extracted using aspect-based sentiment analysis.

The nature of the problem contributes to further difficulties in sentiment analysis. Negative. It is also difficult to decide whether the text should be called objective or subjective. One of the most difficult tasks in sentiment analysis is determining the mood of the opinion of your reader or the person expressing the emotion in the text. The data range has a major impact on sentiment analysis. Words can have a positive in one part and a negative in another part. Some writing styles such as sarcasm, sarcasm and negative sentences can make sentiment analysis difficult.

Additionally, text properties pose particular challenges for sentiment analysis. For example, users of social media texts (like Twitter) often use unusual spelling, grammar, and sarcasm to express their opinions. According to [41] Sarcasm is a type of sentiment where people express negative feelings in positive or very positive language. Sarcasm is viewed as a means of expressing feelings, but it is difficult to use because the literal meaning affected by the use of sarcasm must be recognized and treated as the opposite of what is being said.

Sentiment analysis can pose classification problems. This reflects many of the difficulties that such an approach poses as it involves some degree of NLP [42]. The challenge in this area is to improve the ability of machines to understand the text in the same way as human readers. In a highly competitive world, it is very important for businesses to pay attention to what customers are saying about your products and services.

There are some challenges when using sentiment analysis, text preprocessing affects the performance of sentiment classification [43]. Data preparation can be slower and more difficult than data mining [44]. Furthermore, this is a domain and context dependent. Positive text in one area can be negative text in another. For example, "go read a book" is positive for books, but negative for movies [45].

Other difficulties were also identified by [46] the purpose of co-reference resolution is to identify word references that refer to specific entities, such as pronouns. Another common problem is dealing with negative feedback. This includes reversing the placement of opinion words. For example, the positive polarity associated with the word *"good"* is usually reversed with the phrase *"not good."* Negatives include not only syntactic negations (not and nor), but also linguistic patterns such as prefixes *(un- and dis-)* and suffixes *(less)*, making finding negations a difficult task. It is considered. The reviews contain many comparative opinions. Instead of having a direct and accurate opinion about an entity or its characteristics, users can express their opinion by comparing different entities [47].

### I.2.5. Sentiment Analysis Approaches:

Sentiment analysis is gaining increasing attention in academia and business because of its tremendous value and potential for practical applications, especially in the Web 2.0 era [48].Traditionally, sentiment analysis has been viewed as a binomial classification of opinions. In [49] it is shown that mood classification can be divided into three different subtasks.

1. A determination of subjectivity that determines whether a given text contains factual or subjective information.
2. Determines the alignment or polarity of the text. For example, they determine whether a given subjective sentence expresses a positive or negative opinion.
3. Determines the strength of this placement.

Due to the large amount of subjective data on the web, automated sentiment analysis is required to address this issue. Sentiment classification examines the emotions expressed in the text to determine the degree of their polarity. This section describes different machine learning techniques for detecting emotions. Semantic orientation and machine learning are his two main approaches to sentiment analysis and investigation [50] .

- **The machine learning or corpus-based approach:** Use vectors to represent positive or negative data. These vectors are processed by a classifier that is trained to create features that are assigned to specific classes. The focus is on the selection of suitable machine learning algorithms and text features to classify text polarity [51].Text classification methods used in machine learning use two sets of documents: a training set and a test set. The test set is used to evaluate the classifier's performance, while the training set is used to learn the document's distinguishing characteristics. The machine learning-based sentiment classification method has the following characteristics:

✓ ***Term Presence and Frequency:*** This includes -grams or n-grams and their presence or frequency.

- ✓ ***Part of Speech Information:*** Used to distinguish meanings and serve as a guide for feature selection.
- ✓ ***Negations:*** A word or phrase that expresses a positive or negative mood. You have the power to reverse sentiment.

However, we have found that applying a sentiment classifier trained in one domain to another domain often fails [52], [53]. For example, when an in domain support vector machine classifier (90.45% accuracy) trained to validate movie data was used to validate book reviews, a much lower accuracy rate of 70.29% was reported. This clearly shows that moods are situational and can vary depending on the topic [52].

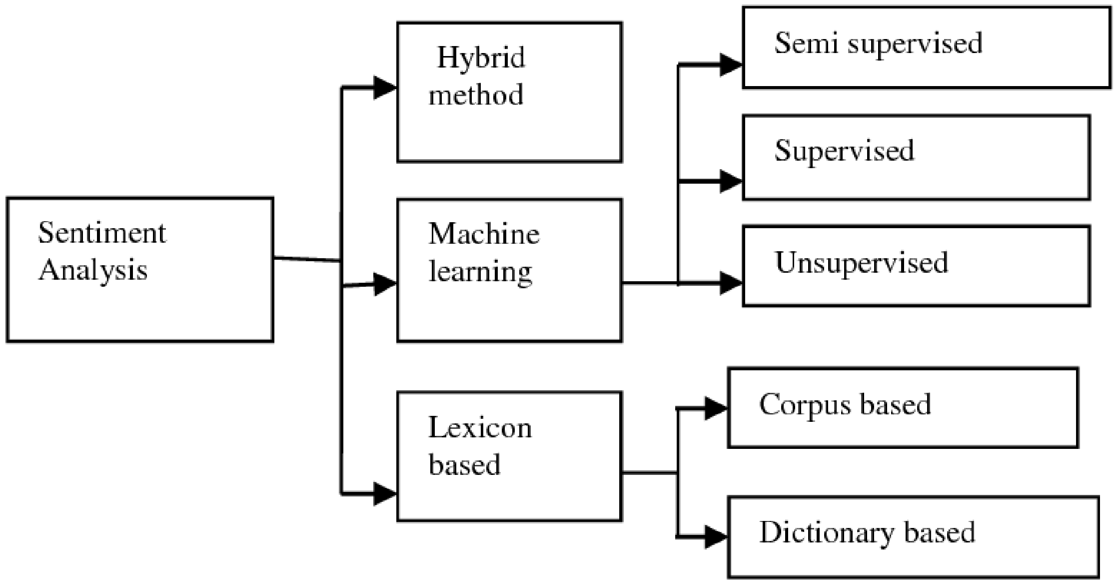


**Figure 3: Approaches of Sentiment Analysis.**

Aspect-based sentiment analysis commonly uses both supervised and unsupervised machine learning techniques to determine sentiment scores. Some approaches also allow the use of lexicons as part of the training data for machine learning techniques. Machine learning techniques determine sentiment by training on known datasets and using multiple learning algorithms.

- **Semantic-oriented approaches, also known as lexical-based approaches:** focus on words and phrases as semantic-oriented indicators. The overall polarity of the text is the average sum of the polarities of the indicators. [54].Dictionary-based approaches keep track of words by creating linguistic resources such as dictionaries. We usually assign a final score for each aspect using an aggregation step that combines the scores for each word and a mechanism that takes into account grammatical rules and surrounding words. The sentiment polarity of a review is determined by the semantic placement of words or phrases within the review. A text's subjectivity and opinion is measured by its "semantic orientation".

To determine polarity, dictionary-based approaches compare opinion words from sentiment dictionaries with available data. Sentiment dictionaries can be created using one of three methods: manual creation, corpus-based methods, or dictionary-based methods. It takes a lot of effort and time to build it by hand. Corpus-based techniques can be used to generate opinion terms with considerable accuracy. Finally, the idea behind the dictionary-based approach is to first manually collect a small set of opinion words with known biases, and then search the WordNet dictionary for synonyms and antonyms for those words.

As dictionaries are created, simply counting the various lexical entries reveals the polarity of the text [55] . The main challenge in building a lexicon is that it is difficult to create manually, and while it is domain dependent, no annotated data set is required to detect sentiment [56]. Unsupervised and supervised learning are terms used by other scientists and researchers to describe these two approaches [57]. A method for adapting sentiment analysis vocabulary to specific problems was proposed in his aspect-based sentiment analysis (ABSA) [58]. In particular, they extended his two vocabulary generation techniques for aspect-based problems using statistical methods and genetic algorithms. This enables the authors to achieve better experimental results compared to existing methods.

In addition, hybrid classification systems for mood analysis are also proposed by combining both approaches [59].

- **The linguistic approach, also known as the hybrid approach:** The two methods mentioned above are combined. Also known as semi-supervised or weakly supervised methods, these techniques use lots of unlabeled data and sparse annotation to improve the classifier. The placement of text is usually determined by the syntax of words and phrases and the layout of the text [60]. Part of Speech (POS) is a technique used in linguistic approaches. These are grammar type designations assigned to each token based on the sentence's language and context information [61].By combining machine learning and lexical-based approaches in a hybrid approach, the performance of sentiment classification can be improved. Using these different approaches has some advantages and disadvantages as shown in Table 1. The main benefit of machine learning techniques is the ability to create pre-trained models that are tailored to specific goals and contexts. However, the main disadvantage is that they integrate general knowledge into the classifier that may not be learned from the training data. Is difficult. Furthermore, trained models are often based on domain-specific features of the training data and are often poorly adaptable across domains or different text genres.

Lexical-based techniques cover a wider range based on general knowledge and feelings. However, these methods have two major drawbacks. First of all, the limited number of words in the dictionary can lead to problems when trying to extract moods from highly dynamic environments. Second, sentiment dictionaries often assign fixed sentiment placements and ratings to words, regardless of how the words are used in the text.

The symbiosis of vocabulary and learning, the perception and measurement of mood at the conceptual level, and the reduced sensitivity to changes in the subject area are the main advantages of the hybrid approach. The main disadvantage of this method is the inability to detect mood. As a result, reviews that contain a lot of noise (words unrelated to the subject of the review) often receive a neutral rating.

**Table 1: Sentiment classification approches.**

| **SENTIMENT CLASSIFICATION APPROACHES** | | **FEATURES/ TECNIQUES** | **ADVANTAGES** | **LIMITATIONS** |
|---|---|---|---|---|
| **Machine Learning** | Bayesian Networks<br>Naive Bayes Classification<br>Maximum Entropy<br>Neural Networks | Term presence and frequency<br>Part of speech Information<br>Negations<br>Opinion words and phrases | Ability to customize and create trained models for specific purposes and contexts | Less applicable to new things, requires labeled data to be available but can be costly or prohibitively |

| | | | | |
|---|---|---|---|---|
| | Support Vector Machine | | | expensive |
| **Lexicon Based** | Dictionary based approach Novel Machine Learning Approach Corpus based approach Ensemble Approaches | Manual construction, Corpus-based Dictionary based | coverage for a longer period of time | A finite number of words in the lexicon and fixed assignments of mood direction and word count |
| **Hybrid** | Machine learning Lexicon based | Sentiment Lexicon. Constructed using public resources for initial sentiment detection. Sentiment words as features in the machine learning method | Symbiosis of vocabulary and learning, perception and measurement of mood at a conceptual level, and reduced sensitivity to changes in the thematic area | loud reviews |

These three strategies can be used individually or in combination. This model classifies unlabeled data using a supervised classifier trained on labeled data. This is the case when machine learning and linguistic techniques are combined; For example, when the features chosen for training only serve as selling points. This means you can create a dictionary of adjectives and classify it as positive or negative. Nonetheless, that supervised and semi-supervised approaches are difficult to analyze and unsupervised or weakly supervised approaches are likely to be even more difficult in [62].

Hybrid approaches is to aspect-based sentiment analysis have been proposed in several articles, aiming to combine the advantages of lexicon-based and machine learning approaches (especially deep learning).;Aspect-based mood analysis combining lexical element domain ontology and a neural attention model [63]. A hybrid approach to categorizing the sentiment of social media posts. The authors deliberately use domain-specific multilingual dictionaries to tag articles and comments with extremist phrases. It then categorizes the sentiment of posts and comments published on social networks using a machine learning approach. This technique reduces the effort and time required to annotate a large corpus of text to train a classifier [64] .

### I.2.6. Sentiment Analysis Applications:

Sentiment analysis is still an active area of research and ever-increasing areas of application, linguistic nuances, different contexts, and even cultural factors make it difficult to assess the sentiment of a text become difficult. Humans' information needs have always included a desire to understand other people's feelings about objects. Individuals traditionally ask friends and family for sentiment information, whereas organizations, businesses, and governments conduct surveys and opinion polls [65].

The reputation of a product, which is derived from the opinions of others, is crucial information when consumers must decide or choose a product. The results of a sentiment analysis can be used to gauge customer opinion. Thus, the first use of sentiment analysis is to offer advice and suggestions for choosing products based on the collective wisdom of the market. When choosing a product, you are frequently drawn to particular features of the item. A single overall rating may be misleading. Sentiment analysis can classify reviewers' viewpoints and estimate ratings on particular product attributes. Businesses that want to know what their customers think about their products can use sentiment analysis in this way. Then, they can improve the areas that the customers thought needed improvement. The importance of various factors for customers can also be ascertained through sentiment analysis. Other technologies have been suggested to include sentiment analysis as a component. One suggestion is to enhance information mining in text analysis by excluding a document's most arbitrary portion or by automatically suggesting internet ads for goods that align with the viewer's preferences (and removing the others). Numerous opportunities exist in the field of human-machine interfaces when we understand how people think.

## Chapter 1: Background and Literature Review

Sentiment analysis, which examines how people feel about a given product, focuses on product reviews. Reviews are a reflection of the content created by actual product users as a result; professionals must receive this feedback [26]. Sentiment analysis has various uses. The use of sentiment analysis in a variety of areas, such as items and services, stock market prices and variances, and even student opinions, can lead to breakthrough overall business success. This research focuses on three such instances: student feedback, product reviews, and the stock market.

Online shopping is a more convenient way for customers and retailers to exchange goods than in-store shopping, but it also has disadvantages: online customers cannot see your product or service directly. Therefore, product reviews are becoming an increasingly important tool for customers to share their experiences. Therefore, there is a growing desire to identify and extract useful customer feedback for marketing purposes [66].

Businesses understand the value of the Internet in gathering user feedback and opinions about their products and services in the marketplace. Businesses value time more than regular users and businesses typically need automated systems to help them understand the feelings and opinions of users of their products and services. However, regular users usually spend hours surfing the web to find out the opinions of these users. Other users, Businesses can benefit even more from tools that can collect and analyze user reviews to understand overall sentiment. With this tool you can collect feedback and ideas from your customers to improve your products and services. For those looking for knowledge and information, the World Wide Web is an excellent resource. You don't need to ask a friend for assistance when you want to buy a product, go on vacation, or need certain services. All you need is the internet to search through this unstructured data. As a result, sentiment analysis should be capable of searching through this data and presenting it to end users in a proper format.

People now use forums and other social networks to share their thoughts and ideas on a variety of topics. Some of these ideas are more uplifting in spirit and substance than others. Extremist groups, according to [67], use the Internet to incite violence and hatred among other groups. As a result, sentiment analysis may be more useful in these situations for monitoring groups' online sentiment. This makes it easier for the government to detect violence early on and deal with it before it escalates. A new method to assess violent emotion in two groups: Middle East and American militants deployed [68].

Social media posts are unsolicited and therefore often contain current opinions on companies, products and services. With sentiment analysis software, you can quickly sort through all of this data and examine both individual and overall public sentiment on any social platform. Sentiment analysis goes beyond simple definitions to identify sarcasm and common chat acronyms like *"lol and ROFL"*. Fix common mistakes like misspellings and misused words. Sentiment analysis on micro-blogs like Twitter requires a different processing method than textual information. Setting a maximum text length often forces users to express their opinions clearly. Irony and irony, on the other hand, present these systems with challenges.

Tweets are full of internal slang, abbreviations, and emoticons, making it difficult to capture opinions. In addition, their short length limits the presence of contextual cues that are typically present in dialogs and documents. Most sentiment analysis studies use only the text modality. However, the availability of user-generated videos on online platforms such as YouTube and Facebook has brought multimodal sentiment analysis to the forefront of sentiment analysis research. This increase is driven by commercial interest as company's base their product decisions on user sentiment in these videos. Multimodal fusion has become a focus of multimodal sentiment analysis as more studies propose new fusion methods. These include multiple kernel learning, tensor-based nonlinear fusion, and memory networks.

One of the most common applications of sentiment analysis in business is brand monitoring. Negative reviews pile up quickly online, and the longer you ignore them, the worse things get. The sentiment analysis tool instantly alerts you to negative brand mentions. Additionally, you can monitor your brand image and reputation over time or at specific points in time to track your progress. By searching for brand news, blogs, forums and information on social media, you can turn this data into useful information and statistics. By tracking trends and predicting outcomes, machine learning can also help you stay ahead and move from a reactive to a proactive approach.

If you focus on your customers and listen to them, you can interact with them effectively. The Repustates Sentiment Analysis solution uses the Text Analytics API to find recurring or overlooked topics and extract information from polls, posts, emails and other data sources. Product quality, product segregation and customer service are just a few examples of how our solutions can help you be more competitive and successful. Because a company's success ultimately depends on its ability to adapt to its market base, it is natural that Voice of the Customer (VOC) measurements be conducted on a regular basis.

Customer feedback from online surveys, chats, call centers and emails needs to be collected and analyzed. With sentiment analysis, you can organize and categorize this data to find patterns and recurring issues and themes. Listen to your customers and learn how to communicate with them, including what works and what doesn't.

To reduce human bias, an increasing number of businesses are abandoning antiquated methods of collecting employee feedback in favor of cutting-edge artificial intelligence technologies. Continuous surveys and feedback are used to collect data throughout an employee's tenure at a company. These feedback programs include audio, chatbots, and videos. Repustate's sentiment mining tool employs aspect-based sentiment analysis (ABSA) to provide businesses with powerful insights that allow them to strengthen their relationships with their employees. The solution enables you to make better decisions by assisting you in measuring employee satisfaction and identifying risk factors to prevent employee attrition by getting to the root of the problem.

Analyze never-before-seen feedback over the years or find out what the public thinks of a new product after it's launched. Aspect-based sentiment analysis allows you to search for keywords related to specific product features (interface, UX, features) to find exactly the information you need. Sentiment analysis tells you how your target audience views your product, what aspects of your product need improvement, and what your most valuable customers are happy with.

Utilize sentiment analysis, users can perform a wide variety of competitive and market research. It can mean the difference between breaking into a new market and gaining an advantage over your competitors. By reviewing your product online, you can compare it to that of your competitors. You can also exploit your competitors' weaknesses by analyzing them. In sociology, psychology, and political science, sentiment analysis are used to analyze trends, opinions, ideological bias, gauge reaction, and so on. Many of these sentiment analysis applications are already operational.

Measuring the patient experience using innovative voice-of-the-patient techniques provides valuable insights for hospitals, pharmaceutical companies and health insurers. A study published in the New England Journal of Medicine found that 97% of physicians felt that listening to patients is important to improve patient care. Feedback and patient information can be collected through email, online forms, phone calls, and other methods. With Repustates' Voice Sentiment Analysis technology, you can continue to learn from the feedback you receive through voice or video consultation recordings. Audio sentiment analysis examines a speaker's mood from an audio signal by capturing representative features known as audio sentiment vectors (ASV). With the help of automatic speech recognition, the speech is converted into a transcript.This transcript is fed into Repustate's text analysis pipeline, which automatically identifies emotionally relevant linguistic features and provides important and comprehensive insights.

## I.3. Conclusion

This chapter described various features of sentiment analysis (SA). It provides comprehensive background information about sentiment analysis and helps you gain a solid understanding of SA. It also shows that different areas of SA are considered to provide a comprehensive understanding of the fundamentals, SA approaches, challenges, and application areas. This helps in creating a comprehensive and holistic view of the various SA areas, making it easier to apply the required processes in later phases.

# Chapter 2

# Machine Learning and Deep Learning For Sentiment Analysi

# II. Chapter 2 Machine Learning and Deep Learning for sentiment analysis

## II.1. Introduction:

Nowadays, when companies use artificial intelligence programs, they typically rely on machine learning. These terms are often used in place of each other and can be confusing. Machine learning is a part of artificial intelligence. It helps computers learn new things without someone telling them exactly what to do.

Human brains have many neurons that work together to learn things. Deep learning uses software nodes to create neural networks with many layers that work together. Deep learning models are created by using a lot of labeled information and neural networks. So, in this chapter; we will talk about two things: using machine learning and using deep learning for sentiment analysis. We will explain how these methods work, what they can be used for, the steps involved, and how well they perform.

## II.2. Machine learning approaches for sentiment analysis:

Machine Learning (ML) is a vast field with many applications including information technology, statistics, probability, artificial intelligence, psychology, neurobiology and more. Simple problems can be solved with machine learning by developing models that represent a given data set well. ML has evolved from teaching computers that mimic the human brain to a broader field of developing basic statistical computational theories of learning. Machine learning is the development of algorithms that enable computers to learn. The process of discovering statistical rules and other patterns in data is called learning. Machine learning algorithms [69] are designed to mimic the human approach to learning tasks. These algorithms can also reveal the relative difficulty of learning in different environments.

Machine learning is a branch of artificial intelligence that uses statistical techniques to enable computer systems to learn from data. By developing self-learning algorithms, machine learning has solved many problems for which there are no known algorithms.

ML-based approaches have long been widely used and preferred for SA applications due to their performance and reliability.

Sentiment analysis based on machine learning has gained popularity in this field, particularly because of its independence manually constructed rules. Despite best efforts, it was impossible to list all the rules, which limited the ability to generalize. The potential to learn general representations emerged with machine learning. ML-based approaches, both supervised and unsupervised, have used a variety of algorithms, including SVMs and Naive Bayes classifiers' nearest neighbors, in combination with features like bag-of-words (including weighted variants), lexicons, and syntactic features like parts of speech.

### II.2.1. Machine Learning Applications:

Machine learning is a rapidly growing field. Although we may not realize it, we are using machine learning on a daily basis in applications such as Google Maps and have numerous application domains and sub-domains Below is a list of some of the top real-world machine learning applications:

- **Computer vision**: The sub-domains of the computer vision domain are object recognition, object detection, and object processing. Classification, analysis, and recommendation are some of the different sub-domains that are covered here. Machine learning has been used to successfully implement text classification, document classification, image analysis, medical diagnosis, network intrusion detection prediction, and denial of service attack prediction.
- **Natural language processing, semantic analysis, information retrieval:** Semantic analysis is the process of linking the syntactic structure of sentences, paragraphs and words to the overall level of the sentence. Computers can process natural language data correctly if they are programmed to process natural language. The science of finding information in documents, in metadata that describes documents and data, and in databases, including sounds and images, is called information retrieval.

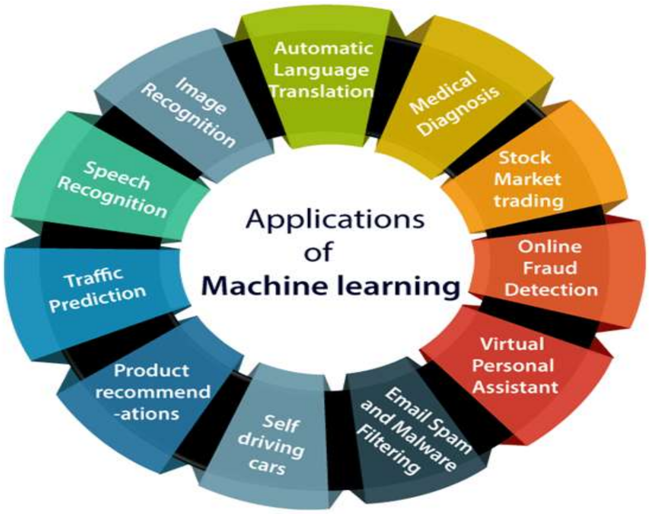


**Figure 4: Machine learning applications.**

- **Image Recognition:** One of his most popular uses of machine learning is image recognition. Used to identify digital images, people, places, objects, etc. Auto-tagging friends is a common use of an image and face recognition. Facebook offers suggestions to auto-tag your friends. Face recognition and recognition algorithms used in machine learning to automatically provide tagging suggestions with your Facebook friend's name when you upload a photo.
- **Speech Recognition:** Google allows you to do "voice search", a popular machine learning application that corresponds to speech recognition. Speech recognition, also known as "speech-to-text" or "computer speech recognition » is the process of converting spoken instructions in the text. Today, machine learning algorithms are used extensively in speech recognition applications.
- **Traffic prediction:** When you want to travel to a new place, Google Maps can help you find the best routes and predict traffic conditions. Predict traffic conditions using two methods Is it sunny, slow moving, crowded, etc.:

✓ Real-time vehicle location with Google Maps app and sensors

✓ Average time taken for the same hour over the past few days.

- **Product Recommendations:** Machine learning is widely used by Amazon, Netflix and other e-commerce and entertainment companies recommend products to users. Thanks to machine learning, you'll see ads for the same products every time you search for products on Amazon while browsing the web in the same browser.

- **Self-driving cars:** Self-driving cars are one of the most exciting applications of machine learning. Self-driving cars rely heavily on machine learning. Tesla, the most famous automaker, is developing self-driving cars. Using unsupervised learning techniques, car models are trained to recognize people and objects while driving.
- **Email Spam and Malware Filtering:** All new emails you receive are automatically classified as important, frequent or spam. Spam emails always end up in the spam folder, while important emails always show up in your inbox with an important icon. It is a technology based on machine learning. The spam filters used by Gmail are: content filters, header filters, general blacklist filters, rule-based filters, permissions filters.
- **Virtual Personal Assistant:** All new emails you receive are automatically classified as important, frequent or spam. Spam emails always end up in the spam folder, but important emails always show up in your inbox with an important icon. It is a technology based on machine learning. The spam filters used by Gmail are: content filters, header filters, general blacklist filters, rule-based filters, permissions filters.
- **Online Fraud Detection:** Machine learning fraud detection ensures safe online transactions. Any time you conduct an online transaction, fraudulent transactions can occur in a variety of ways, including using fictitious accounts and ID cards, and stealing money during transactions. A feed forward neural network can help detect this by determining whether a transaction is legitimate or fraudulent.
- **Stock Market trading:** Trading on the stock market frequently makes use of machine learning. A long short term memory neural network is used in this machine learning project to predict stock market trends because there is always a risk of ups and downs in share prices.
- **Medical Diagnosis:** Machine learning is used in medicine to detect diseases. For this reason, medical technology has developed rapidly and enables the creation of 3D models that can be used to localize brain lesions. It helps in the early detection of brain tumors and other brain-related diseases.

- **Automatic Language Translation:** Nowadays, it doesn't matter if you're traveling to a new place and don't speak the local language, machine learning can even translate texts into your native language. Sequence-to-sequence learning algorithms combined with text translation and image recognition from one language to another is the technology behind automatic translation.

Machine learning is used to automate sentiment analysis. This implies that businesses can obtain information instantly. This can be very useful for identifying problems that require immediate attention. For instance, a bad story trending social media can be identified immediately and addressed. If one customer complains about a problem with their account, others may experience the same issue. Companies can avoid negative experiences by immediately alerting the appropriate teams to resolve this problem.

Determine the sentiment polarity of target text using a pre-trained ML decision model. This ML-based decision model is created by training an ML algorithm on an emotion dataset that comes from an emotion-specific corpus. There are two types of ML algorithms used to solve two different problems: classification and regression; if the target text can be classified into one of three sentiment categories (positive, negative, or neutral), SA can be considered an ML classification problem; if the target text can be classified into one of three sentiment categories (positive, negative, or neutral), SA can be considered an ML classification problem. SA is also an ML regression problem where the ML model predicts a number representing the sentiment strength value of the target text [70]. Formally, a predefined ML algorithm defined as [71] an ML-based computer program is said to learn from experience **E** with respect to any class of task **T** and performance measure **P**, if its performance of tasks in **T** as measured by **P** improves with experience **E**.

### II.2.2. Machine learning methods:

In machine learning, a machine is presented with a large amount of data so that it can learn and make predictions, find patterns, or classify data. The three types of machine learning are supervised, unsupervised, and reinforcement learning.

#### II.2.2.1. Supervised Learning:

Classification problems are a common formulation of supervised learning tasks. The learner must learn a function (or approximate its behavior) that assigns vectors to one of several classes by looking at some example function inputs and outputs [72] . The algorithm already knows the correct data output. The goal is to find patterns in the input that characterize the desired output. Two prominent examples are regression algorithms (continuous output) and classification algorithms (discrete output).

Supervised learning is effective at resolving classification and regression issues, such as figuring out what news article category it falls under or forecasting the amount of sales for a specific date in the future. Making sense of data within the context of a particular question is the goal of supervised learning. Training is the basis of supervised learning. During the training phase, the system is fed a labeled data set that tells the system, how to map the outputs to each input value. Data that has been labeled, but not yet published to the algorithm shows test data from the trained model. The purpose of the test data is to evaluate how accurately the algorithm performs on unlabeled data.

Although supervised learning has many advantages, such as improved automation and deep data an insight, developing sustainable models poses several challenges, including the following: Because of the increased likelihood of human error in supervised learning datasets, algorithms may learn incorrectly. Supervised learning allows you to collect data and generate data output from previous experiences, allowing you to use your experiences to refine performance criteria. And it helps solve a variety of real computer problems.

The downside is that it can be difficult to classify large amounts of data. The training of supervised learning requires a lot of computing time. Therefore, it is very time consuming. Detailed structuring of this type of model requires expertise, and model training can be time-consuming.

Analysis of customer sentiment Massive amounts of brand sentiment data can contain critical information that can be identified and characterized by supervised machine learning algorithms, providing detail about emotion, context, and intent with little to no human involvement in improved brand engagement efforts. Objects in images and videos can be found, isolated, and classified with the aid of ML algorithms. In order to provide deeper insights into a variety of data points, predictive analytics systems are developed with the aid of supervised learning models. Due to their ability to pivot in favor of the brand, defend choices, or predict outcomes based on particular output variables. Modeling is a type of spam detection. Users can efficiently train databases to separate correspondence related to spam from correspondence unrelated to spam based on patterns or anomalies in new data by implementing supervised classification algorithms in machine learning.

### II.2.2.2. Unsupervised Learning:

Unsupervised machine learning techniques is particularly useful for describing tasks, as they aim to discover relationships in data structures without requiring measurable results. Unsupervised machine learning is so named because it lacks a response variable that can be monitored in the analysis. Unsupervised learning aims to find hidden dimensions, components, clusters, or trajectories in data structures [73]. In this case, the algorithm does not know what the result will be. The purpose is to produce a structured output from the input. Clustering algorithms are a good example of this type.

Unsupervised learning is a method that uses unlabeled datasets to find structure and similar patterns in input data. When collecting reliable annotated datasets is complicated, but collecting unlabeled data is easier, the unsupervised method is usually used. It does not cause any problems when retrieving new domain-dependent data.

ML has proven to be very effective on a variety of real-world problems. Because it learns from large amounts of data, it helps discover patterns that humans can't see. It also shows how you can learn to draw informed conclusions from data and trends. These algorithms are used today for a variety of purposes, including product recommendations, weather forecasting, and disease detection. In addition, it is also used in the NLP and SA area, for example to help with spam detection. Among the SA applications, the development of sentiment analysis methods for polarity classification is the most challenging [74].

If we have unlabeled data that needs to be classified, we should create a classification model for the classification problem that is trained on a set of labeled data and their target classes. There are two kinds of classification problems: binary classification and multiclass classification. If we have two target classes, we are dealing with a binary classification problem, such as a positive or negative target class for a text document. Otherwise, if there are more than two target classes, multi-class classification is used. This model can be trained using target vectors; each vector represents a set of features chosen with the target class in mind. Each vector of features in text classification represents a document or sentence and includes extracts feature values from these documents. For example, a document can be classified as positive if the result value of its included function is greater than or equal to zero; otherwise, the document can be classified as negative.

The main difference between supervised and unsupervised learning is the learning process of the algorithms. The algorithm takes unlabeled data and uses it as a training set for unsupervised learning. Unlike supervised learning, where the correct output values are given, unsupervised learning looks for patterns and similarities in the data rather than correlating the data with external measurements. In other words, the algorithms are free to learn about data and discover interesting or unexpected results that humans didn't want. Clustering applications (the process of identifying groups in the data) and associations benefit greatly from unsupervised learning (the process of predicting rules that describe data). Although supervised learning models have some advantages over unsupervised methods, they also have disadvantages. Supervised learning systems make more humane decisions.

The following Table 2 are shown the primary distinctions between supervised and unsupervised learning:

**Table 2: The distinctions between supervised and unsupervised learning.**

| Supervised Learning | Unsupervised Learning |
|---|---|
| **Labeled data is used to train supervised learning algorithms.** | Unsupervised learning algorithms are trained on unlabeled data. |
| **A supervised learning model uses direct feedback to determine whether it is predicting the correct output.** | Unsupervised learning models do not accept feedback. |
| **The output is predicted by a supervised learning model.** | Unsupervised learning models are used to discover hidden patterns in data. |
| **In supervised learning, the model is fed both input and output data.** | Only input data are provided to the model in unsupervised learning. |
| **The goal of supervised learning is to train a model to predict outputs when presented with new data.** | Unsupervised learning looks for hidden patterns and useful insights in unknown datasets. |
| **Supervised learning requires supervision to train a model.** | Unsupervised learning does not require a teacher to train the model. |

| | |
|---|---|
| **Classification and regression problems are two types of supervised learning problems.** | Clustering and Associations problems are two types of unsupervised learning problems. |
| **Supervised learning can be used when both the input and the corresponding output are known.** | Unsupervised learning can be used when you only have input data and no corresponding output data. |
| **A supervised learning model yields reliable results.** | Unsupervised learning models may produce less accurate results than supervised learning models. |
| **Supervised learning differs from true artificial intelligence because a model must be trained for each data set before it can predict the correct output.** | Unsupervised learning is similar to real artificial intelligence in that it learns the same way a child learns everyday life the experience. |
| **It includes various algorithms like linear regression, logistic regression, support vector machines, multi-class classification, decision trees, Bayesian logic and more.** | It contains various algorithms such as clustering, ANN and a priori algorithm. |

### II.2.2.3. Reinforcement machine learning:

Reinforced machine learning is a machine learning model similar to supervised learning, but the algorithm is not trained on sample data. The model learns over time through trial and error and by finding mistakes and rewards. Delayed gratification through trial and error is a key feature of reinforcement learning. Using this method, machines and software agents can automatically determine their ideal behavior in a given context to maximize performance. A set of successful results is enriched to create optimal recommendations and guidelines for specific problems. Beneficial outcomes are encouraged or reinforced, and negative outcomes are discouraged or penalized. If the program finds the right solution, the interpreter improves the solution by rewarding the algorithm. In the case of unfavorable results, the algorithm should be repeated until better results are found. In most cases, reward systems are directly related to the effectiveness of the results. For common reinforcement learning using cases, such as finding the shortest route between two of your points on a map, the solution is not absolute. Instead, an effectiveness rating is given as a percentage. The higher this percentage, the more reward the algorithm gets. The program is therefore designed to provide the best possible solutions for the best possible rewards.

Practical applications of this type of machine learning are still under development. Some use cases include: teaching cars to park and drive themselves, and dynamically controlling traffic lights to reduce congestion. You can also teach your robot to learn policies using raw video images as input and use them to replicate the actions shown.

Reinforcement learning problems can be formalized using a Markov decision process (MDP). With MDP, agents constantly interact with the environment and perform actions. The environment is responsive and generates a new state in response to each action.

Reinforcement learning is mainly classified into two types of techniques/algorithms:

- ✓ **Positive Reinforcement learning:** refers to the fact that adding something increases the likelihood that the requested behavior will occur again. It strengthens the strength of the agent's actions and influences them positively.
- ✓ **Negative Reinforcement Learning:** Works the exact opposite of positive RL. Avoiding negativity increases the tendency for certain behaviors to reoccur.

Here are some examples of real-world use cases for reinforcement learning:

- ✓ **Video Games:** Reinforcement learning algorithms are much more popular in gaming applications. Used to perform super-human feats.
- ✓ **Resource management:** When used on a computer, it automatically learns and schedule resources to wait for different jobs in order to minimize the average job slowdown.
- ✓ **Robotics:** Reinforcement learning is commonly used in robotics applications. Robots are used in industry and manufacturing, where they become more efficient through reinforcement learning. Various industries have visions of building intelligent robots using AI and machine learning technology.
- ✓ **Text mining:** One of his great uses of NLP, text mining, is now implemented using reinforcement learning from Sales force Company.

The learning model of reinforcement learning is similar to human learning. This will help you find the most accurate results and get long-term results. It helps solve complex real-world problems that are difficult to solve with common techniques. On the other hand, reinforcement learning algorithms are not suitable for simple problems and require large amounts of data and computation. Too much reinforcement learning can overload the state and lead to weak results.

### II.2.3. Machine learning algorithm:

The use of machine learning techniques in natural language processing has helped make sentiment analysis research more widely accessible. Machine learning approaches used text feature representations and a range of algorithms such as Nave Bayes, Support Vector Machines (SVM), KNN, LDA, and LR commonly used to create classifiers for sentiment analysis.

#### II.2.3.1. Support Vector Machine (SVM):

Support Vector Machine (SVM) is a supervised learning algorithm with a wide range of applications that has fueled its adoption by many researchers. In addition, it outperforms other classifiers and is therefore widely used in text classification and SA problems [75]. The main idea of SVM is to find the best way to separate different classes using a specific linear delimiter. This helps in the analysis, detection and classification of data.

A linear learning technique called Support Vector Machine uses a binary classifier. The main purpose of SVM is to distinguish between positive and negative cases [76]. To achieve this, the function $f(x)$ is sought. If you're using an SVM that classifies$x$, you don't need to know the function. Where $x$ is an input vector with $n$ attributes. Positive if $f(x)$ is not negative, negative otherwise. The input is split into two parts by the hyper-plane of the function, one containing the positive instances and one containing the negative instances. The graph shows how the SVM linear function distinguishes between positive and negative cases. That is, the SVM tries to find a decision boundary with the largest margin of error separating the two classes. If the two classes cannot be linearly separated, the boundary is determined by transforming the input space (all instances of the data set D) into a n-dimensional space rather than hyper-planes, so the separation is planar and not linear. Another important feature of SVMs is their ability to handle linear- non-separable problems. In-product operations impact classification capabilities [77].

SVM separates and constructs a maximum margin hyperplane that can be used for classification using training data [78]. It is defined in formula (1) as follows:

$$\boldsymbol{f(x) = \omega . x + b \ , \omega \ \epsilon R^{d} \ , b \ \epsilon \ R} \qquad \textbf{(1)}$$

Where $\boldsymbol{d}$ is the dimensionality of data space, $\boldsymbol{\omega}$ is the weight factor, $\boldsymbol{b}$ denotes bias of the function, and $\boldsymbol{x}$ is the training data vector.

The optimal hyper-plane of SVM is defined in formula (2) as:

$$\|\boldsymbol{\omega}.\boldsymbol{x} + \boldsymbol{b}\| = \boldsymbol{0} \qquad \textbf{(2)}$$

Where $\boldsymbol{y}$ is a vector of class labels (positive or negative). Then the process of classification is a function in formula (3) as:

$$\boldsymbol{y} \mapsto \boldsymbol{sgn}(\boldsymbol{\omega}.\boldsymbol{x} + \boldsymbol{b}) \qquad \textbf{(3)}$$

The SVM kernel is a function k defined as a dot product of data inputs $X_i$ and $X_j$ :

$$\boldsymbol{K}(\boldsymbol{X}_i^T, \boldsymbol{X}_j) = \boldsymbol{\varphi}(\boldsymbol{X}_i^T).\boldsymbol{\varphi}(\boldsymbol{X}_j) \qquad \textbf{(4)}$$

Where $\boldsymbol{\varphi}$ is some linear or non-linear data transform.

However, there are a few disadvantages to consider when using SVM:

- Because SVMs operate in real space, all non-numeric attributes must be converted to numeric before using SVMs. This can be achieved by replacing each attribute with a boolean value (1 if the attribute is present, 0 if the attribute is not present).
- SVM is a binary classifier. If you have more than two classes, you can't use SVM directly, so you need to make significant changes before using SVM.
- SVM's hyper-planes are complex, and humans struggle to understand them visually.

To improve on the usual n-gram-based classifier that ranks posts based on their sentiment polarity, the Twitter Sentiment Classifier uses an SVM-based classifier that integrates an internal n-gram function with the external sentiment vector. Conducted tests using three distinct weighting methods and four different sets of n-gram characteristics [79]; used Support Vector Machine to build a sentiment classifier with n-grams and various weighting schemes on benchmark datasets. Applying chi-square feature selection will greatly increase the classification accuracy for both datasets, according to the results [80].

### II.2.3.2. K-Nearest Neighbor Algorithm (KNN):

The K-Nearest Neighbor (KNN) algorithm is one of the most basic pattern recognition methods and classifies an unknown instance based on the class of the closest training instance in the feature space. Based on the same principle, a very good and practical supervised learning method is to use the last k. It is based on the calculation of the page [81]. A k-NN algorithm predicts a target output for a new object based on the results of nearby samples or multiple nearby objects in the training set feature space.

One of the most fundamental machine learning algorithms, the KNN algorithm, is a nonparametric technique that can handle both [82]: classification and regression problems. The samples are classified by estimating the neighborhood majority vote and assigning a new object class with the highest frequency among the nearest neighbors.

Let TR be a training dataset and TS be a test set; both are made up of a fixed number of samples, n and t, respectively. Each sample $x_p$is a vector $\left(x_{p1}, x_{p2} \ldots\ldots x_{pn}, \omega\right)$ in which $x_{pf}$ is the value of the f-th feature of the p-th sample. TR samples all belong to a known class, whereas TS samples do not. The k-NN algorithm computes the distance between each sample in the TS and each sample in the TR. The most common distance function is the Euclidean distance. Thus, k-NN selects the k closest samples in TR by ranking them in ascending order based on distance. The simplest approach, then computes a majority vote using the class label of the k nearest neighbors [83].

This algorithm performs a class classification based on the nearest neighbor class as the given k-value. The KNN algorithm performs vector classification using a vector n with known the k in the training cluster closest to the sample is chosen as the class to test. The sample under test is found to be a member of the class that includes the greatest number of samples in the chosen sample group. The distance between samples is calculated using the Euclidean distance. The Euclidean distance formula (5) gives the distance between two n-dimensional points [84] as:

$$\boldsymbol{d}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\{\sum_i^n (\boldsymbol{x}_i - \boldsymbol{y}_i)^2\}} \qquad \textbf{(5)}$$

To compare KNN with various techniques and determine the ideal K for KNN using the movie review dataset [85]. In comparison to NB, SVM, and Random forest, KNN with an Information Gain feature selection offers the best performance approach with 96.8% accuracy; the ideal K is 3. The output of the k-Nearest Neighbors model that was created using sentiment analysis of tweets from PT PLN (Persero) can be used or evaluated using a new twitter data file. Based on the training data that was already used, the k-Nearest Neighbors model will forecast the test data. By including training data with cleaner data quality, this model can be enhanced even more [86] . Using the K closest neighbors (KNN) and term frequency-inverse document frequency (TF-IDF) algorithms, this study categorizes talks into two groups: "satisfied" and "dissatisfied"; Testing with a confusion matrix yields results with accuracy, precision, and recall values of 74.00%, 76.00%, and 73.08% respectively.

Customers are more satisfied with service and meeting their needs in conversations with the label "satisfied," while they are less satisfied with waiting times in interactions with the label "not satisfied" [87]. Using the KNN algorithm to conduct sentiment analysis of Twitter users on issues related to government online learning policies Word weights using TF-IDF will be classified into two types of sentiment values: Positive and negative. After testing with K of 20, the highest accuracy result is obtained when K = 10 with an accuracy value of 84.65%, an accuracy of 87%, a recovery of 86% f measured is 87%, error rate 0.12% and trend [88].Using raw Twitter data on wildfire government sentiment, an automated labeling process and sentiment analysis were performed using VADER polarity detection and K-nearest neighbors. A workflow combining VADER's lexical polarity detection with K nearest neighbors allows us to achieve 79.45% accuracy when analyzing raw Twitter data [89].

### II.2.3.3. Naive Bayes Algorithm (NB):

Naive Bayes is one of the most efficient inferential learning algorithms for machine learning and data mining. Its performance in classifying collisions is undesirable because the basic assumption of conditional independence is rarely true in real-world applications [90].

For real learning problems, Naive Bayes is based on assumptions that are rarely true. When considering predicted values, the attributes used to generate predictions are independent of each other. NB gives the probability that each potential value is within the target range. It then creates a single prediction by combining the distributions of likely outcomes.

Using a naive Bayes classifier for word training is convenient and simple, especially for word features with a lot of data to evaluate and classify. Furthermore, it has the ability to create complex models. The naive Bayes classifier, in contrast hand, works best when the testing attributes are independent [91]. These are very simple Bayesian networks consisting of a directed acyclic graph with a single parent node (representing the unobserved node) and a set of child nodes (representing the observed nodes). They strongly assume the independence of child nodes with respect to their parent nodes. This estimation is thus the basis for the independence model (Naive Bayes). Bayesian classifiers usually have lower accuracy compared to other more complex learning algorithms (e.g. ANN) [92].

NB is dependent on the specifics of the probability model; as a result, the classifier can be trained effectively using the maximum probability method in a supervised learning environment [93], [94] the following is an explanation of NB:

The NB classifier is motivated by initially ignoring the Bayes rule by providing to a given document $\boldsymbol{d}$ the class **c* = arg maxc** $\boldsymbol{P(c|d)}$ formula (6) as:.

$$\boldsymbol{P(c|d) = \frac{P(d|c)P(c)}{P(d)}} \qquad \textbf{(6)}$$

Where $\boldsymbol{P(d)}$ Plays no role in electing **c*** to guess the term $\boldsymbol{(d|c)}$ . NB decomposes it by assuming $\boldsymbol{f_i}$ 's are conditionally independent, given $\boldsymbol{d}$ 's class in formula (7) as:

$$\boldsymbol{P(c|d) = \frac{P(c)\left(\prod_{i=1}^{m} P(f_i|c)^{n_i\,(d)}\right)}{P(d)}} \qquad \textbf{(7)}$$

Given the existence of functional dependencies, the success of Naive Bayes can be explained as follows. It is not required that the fit to the probability distribution be optimal (i.e. that the independence assumption is true) in terms of zero-one loss (classification error).Instead, the best classifier is determined when the actual and predicted distributions match in the most probable class [93]. In other words, the NB model treats each component independently and offers many advantages. In general, it's quick to sign up, easy to build, doesn't require extensive planning, and has less over-customization than other models.

In an attempt to classify the feelings of film critics and hotel critics, a comparison of two supervised machine learning algorithms, Naive Bayes and ANN, was conducted [95], According to the test results, the classifier performed well in film reviews, with the naive Bayesian method outperforming the ANN method, achieving over 80 % of accuracy. The Naive Bayes classifier can be used effectively to rate film reviews. A hybrid dictionary-na ïve Bayesian classifier (HL-NBC) sentiment analysis method has been proposed [96]. Prior to the sentiment analysis engine classifying tweets and removing irrelevant tweets, topic classification is carried out. The proposed method is compared to Lexicon and Naive Bayesian classifiers for uni-gram and bi-gram features. Among the different approaches, the proposed HL-NBC method enables better mood classification with an accuracy of 82% comparable to other methods.

Facebook Page Sentiment Analysis explores the sentiments and attitudes of people and investors by analyzing and classifying positive and negative comments using a simple Bayesian classifier. Useful for consumer behavior, marketing analysis, information dissemination, and attitude research [97].

### II.2.3.4. Latent Dirichlet Allocation (LDA) :

Latent Dirichlet membership is a probabilistic generative model for discrete data collections such as text corpora. Using a three-level hierarchical Bayesian model called LDA, each item in the collection is represented as a finite mixture over a subset of topics. Given an underlying set of subject probabilities, each subject is in turn modeled as an infinite mixture. Topic probabilities give the document an explicit representation in the context of text modeling.

LDA is a statistical technique based on the assumption that the author of the document has a specific topic in mind. When writing about a topic, choose terms that are likely to be specific to that topic's vocabulary. The entire document can then be presented as a collection of many topics. If the document was authored by one of them, these topics reflect that person's language and perspective on the topic. When multiple users annotate a resource using a tagging system, the resulting topics represent a common, collaborative view of the document, and the topic tags are a shared vocabulary for describing content. It will be displayed. In general, LDA accounts for data similarities by grouping the properties of the data into invisible sets.

The process of utilizing the data corpus from latent variables is called generative LDA. Inferential LDA, on the other hand, is the process of determining the probability value of a topic and reporting that value. The basic idea of LDA is that documents are modeled as an ad-hoc mixture of potential topics, each defined by the distribution of words within the document. The goal of LDA is to determine the number of topics in a corpus and the word distribution for each topic [98] Although NLP (natural language processing) tasks such as extracting latent themes from text documents are most commonly associated with LDA, it is now being used more widely in other fields such as remote sensing, biology, and genomics.

A Dirichlet prior provides the distribution of active topics in documents in Latent Dirichlet Allocation (LDA) [99]. Because they enable documents to contain terms from multiple topics instead of just one, LDA and similar models have the extensive representational capacity. When just the words are seen and the accompanying subjects are hidden, the unsupervised estimation issue becomes more difficult as a result of the enhanced representational power.

The LDA modeling procedure can be characterized as the discovery of a mixture of themes for each resource, i.e., $\mathbf{P(z|d)}$, with each topic being specified by words following a different probability distribution, i.e., $\boldsymbol{P(t|z)}$. Formally, this is formula (8) as follows:

$$\boldsymbol{P(t_i|d) = \sum_{i=1}^{z} P(t_i|z_i = j)\, P(z_i = j|d)} \qquad \textbf{(8)}$$

Where $\boldsymbol{P(t_i|d)}$ is the probability of the $i^{th}$ term for a given document $\boldsymbol{d}$ and $\boldsymbol{z_i}$ is the latent topic. $\boldsymbol{\sum_{i=1}^{z} P(t_i|z_i = j)}$ is the probability of $\boldsymbol{t_i}$ within topic $\boldsymbol{j}$. $\boldsymbol{P(z_i = j|d)}$ is the probability of picking a term from topic $\boldsymbol{j}$ in the document.

The number of latent topics $\mathbf{z}$ should be determined in advance, the degree of specialization of the latent topics can be adjusted. LDA estimates the topic term distribution $\boldsymbol{P(t|z)}$ and the document topic distribution $\mathbf{P(z|d)}$ from the unlabeled document corpus using the Dirichlet priorities of the distributions and a fixed number of topics.

Using the LDA and SVM methods to classify sentiment polarity in a Netflix application review gave the best results with a score of 78.15%, the number of topics of 40, and filtering stage removal in the preprocessing stage [100] ; a novel approach to developing dynamic energy lifestyles based on residential electricity demand data using LDA. This structure is infinitely scalable and extensible, as well as adaptable to different time intervals and entirely new sources of residential energy data from other locations and contexts [101]. LDA uses keyword filters to identify product attributes for online customer reviews. Compared with previous LDA applications, the proposed method improves the automation of keyword preprocessing. A case study of an Android Smartphone shows that the proposed method outperforms previous methods in terms of LDA results in identifying product attributes s [102].

LDA is a generative approach for discovering hidden semantic patterns in large datasets. It is a topic modeling-based model that is recurrently used for content analysis [103]. For these reasons, the LDA model was chosen over others and used for e-learning corpus topic modeling analysis. LDA was used to model topics regarding complaints from customers, which also described that LDA is an algorithm for unstructured data that is suitable for evaluating consumer behavior, product reviews, and other research.

Even though this method makes it possible to analyze large data sets, it is currently not possible to use LDA to carry out more in-depth analyses like systematic reviews. Deeper analysis of large data sets may be possible in the future with advancements in topic modeling algorithms, which could be anticipated to yield very significant insights for the field's researchers.

### II.2.3.5.LogisticRegression (LR):

Logistic regression (LR) is one of the most significant statistical and data mining methods used by statisticians and researchers for the analysis and categorization of datasets with binary and proportional responses Nevertheless, for linear regression, both the dependent variable and the model residuals must be continuous and regularly distributed. When the dependent variable is a binary event (such as yes/no, 0/1, etc.) rather than a continuous parameter, logistic regression is a version of linear regression that can be used; Maximum-likelihood Sampling Bias and Class Imbalance Rational Regression. The ability to naturally offer probabilities, the

extension to multi-class classification issues, and the fact that most of the techniques employed in the LR model analysis adhere to the same principles as those in linear regression are some of the key benefits of LR [104].
In logistic regression, the odds ratio is defined as the probability of an event occurring compared to not occur. In other words, the odds ratio in formula (9) as:

$$\frac{\boldsymbol{P}}{\mathbf{1}-\boldsymbol{P}} \qquad \textbf{(9)}$$

If the occurrence probability of the event (success factor) of " $\boldsymbol{P}$ " in this case is $\mathbf{1}-\boldsymbol{P}$, it indicates the possibility that the event will not occur (realization of the failure factor).

Depending on the ratio, the odds' ratio can be greater than 1 or less than 1. Analysis using logistic regression is considered nonlinear. The term "**logit**" refers to a key concept in logistic regression. The logarithm of the odds is called the logit The estimated logistic regression model can be written as follows from this point in formula (10) as:

$$\boldsymbol{logit}(\boldsymbol{P}) = \mathbf{log}\frac{\boldsymbol{P}}{\mathbf{1}-\boldsymbol{P}} = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{\mu} \qquad \textbf{(10)}$$

Here, $\boldsymbol{P}$ is the realization ratio of the case determined as success factor for the dependent variable; $xk$ is the number of the independent variables that involved in the independent variable matrix with dimension of $\boldsymbol{n} * (\boldsymbol{k}+\mathbf{1})$ ; $\boldsymbol{\beta} * (\boldsymbol{k}+\mathbf{1})$ represents the parameter vector $\boldsymbol{\mu}$ and is the error term. With the help of the odds ratio, the success factor of each dependent variable on probability can be obtained by the equation in formula (11) as:

$$\mathbf{P} = \frac{\mathbf{exp}(\mathbf{x}'\,\boldsymbol{\beta})}{(\mathbf{1}+\mathbf{exp}(\mathbf{x}'\,\boldsymbol{\beta}))} \qquad \textbf{(11)}$$

To accomplish the classification task, the probability value $\boldsymbol{P}$ obtained from the logical function transformation for a particular input is then mapped to two or more discrete classes.

In the LR independent variables could be assigned the values 0 and 1, signifying the lack and the presence of landslides, respectively. Model output, which reflects landslide susceptibility, ranges from 0 to 1. The logistic function$\boldsymbol{P_i}$, determined in formula (12) as:

$$\boldsymbol{P} = \frac{\boldsymbol{exp(K)}}{(\boldsymbol{1+exp(k)})} \qquad \textbf{(12)}$$

Where $\boldsymbol{P}$ denotes the probability associated with a particular observation and $\boldsymbol{K}$ can be defined in formula (13) as:

$$\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n} \qquad \textbf{(13)}$$

Where $\boldsymbol{\beta_0}$ represents the intercept of the algorithm, $\boldsymbol{\beta_i}$ represents the coefficient representing the contribution of the independent variable$\boldsymbol{X_i}$, and $\boldsymbol{n}$ represents the number of conditioning coefficients [105].

Tweets can be classified as positive, negative or neutral when logistic regression and classification are used as a classification model, and the anagram is used as a feature vector highlighting features. Labels are established using logistic regression. A word's "effective score" is employed [106]. In a comparison of all machine learning algorithms, they looked at the effects of two sentiment analysis features, TF-IDF word level and N-Gram, on the SS-Tweet dataset and found that logistic regression provided the best predictions of sentiments by having the highest output for all four comparison parameters and both feature extraction methods [107] . The accuracy of the logistic regression when the Bigram model was used was about 86% [108]. For Twitter sentiment, supervised machine learning algorithms perform better than logistic regression.

This Chabot was built using machine learning models, specifically the LR model. This appeared to offer a more diverse response system than, for example, a Chabot built with RasaNLU, where some responses ended in a loop [109]. Highly accurate sentiment analysis models for product, movie, and restaurant reviews on Amazon, IMDB, and Yelp. The pre-processed dataset is subjected to various ML-based classifiers to classify the results. The results show that DT, SVM and LR based classifiers perform well and offer higher accuracy. In all scenarios, the results show that the logistic regression classifier outperforms the other classifiers [110].

### II.2.3.6. Classification and Regression Trees (CART):

Classification and regression trees (CART) are widely used to study and model complex data. You will be able to deal with highly nonlinear relationships involving higher order interactions and variable types. With a randomly chosen remote sensing, sample as its foundation, the CART algorithm builds a binary tree, which it then prunes using the tested sample. One approach for creating decision trees is called the CART algorithm. Each non-leaf node has two branches, since it merely splits a sample into two sub-samples. By using this approach, a binary tree is a decision tree that was produced [111].

Modern techniques based on partitioning the data include classification and regression trees. They exhibit excellent learning performance, are computationally inexpensive, have straightforward conceptual underpinnings, and can be trained using routines and packages in well-known programming languages like Python and R. Because they are rule-based, they are also appealing in terms of interpretability [112].The structure of CART's classification or regression trees is invariant with regard to monotone transformations of independent variables, which is a crucial practical trait. The tree's structure won't change if any variable is substituted with its logarithm or square root value [113]. There are three components of the CART methodology: building the largest possible tree after choosing the appropriate tree size and using the created tree to classify new data.

CART is particularly useful when analyzing complex data as it provides a unique and informative way to represent results in the form of decision trees. This differs greatly from traditional statistical techniques, where linear combinations are the primary means of expressing relationships between variables. CART can process a variety of response variables. The goal is to minimize the sum of squares within the node, for example for continuous data [114].

The CART is a non-parametric statistical technique designed to analyze classification problems using either continuous or categorical dependent variables [115]. In the case of a categorical dependent variable, CART generates a classification tree. A regression tree results from a continuous dependent variable. The CART tree is built by repeatedly separating subsets of the data set into two child nodes using all predictor variables, starting with the complete data set [116]. Various impurity or diversity measurements are used to determine which predictor is the best. Producing data subsets that are as homogeneous as feasible with regard to the target variable is the aim. According to the improvement score or impurity reduction, the CART algorithm evaluates each predictor for each split to determine the appropriate cut point (continuous predictors) or groups of categories (nominal and ordinal predictors). The best-improving predictor is chosen for the split once the predictors have been compared. Recursively, the procedure continues until one of the halting rules is applied.

A binary recursive segmentation method is the CART algorithm. Since the current sample was split into two subsamples, each generated non-leaf node has two branches. As a result, the CART algorithm's decision tree is a simple binary tree. Since the CART algorithm represents a binary tree, there are only two possible outcomes at each step. Even if the features have multiple values, the data will be split into two parts. The CART algorithm consists of two main steps. In the first step the sample is split into a tree building process. The next step is to validate and clean up the data [117].

The CART principle is briefly explained in the section below. Let $\boldsymbol{x_1}, \boldsymbol{x_2} \dots \boldsymbol{x_n}$ represent a single sample's characteristics and indicate the category to which they belong. A one-dimensional space is split into non-overlapping rectangles using the CART algorithm. The splitting process goes like this:

1. After choosing an independent variable $\boldsymbol{x_i}$ and then the value $\boldsymbol{v_i}$ of $\boldsymbol{x_i}$ is selected with $\boldsymbol{v_i}$ the n-dimensional space into two parts, all of the points in the first part satisfying $\boldsymbol{x_i} \leq \boldsymbol{v_i}$ and all of the points in the second part satisfying $\boldsymbol{x_i} > \boldsymbol{v_i}$ for non-continuous variables. In other words, there are only two possible outcomes for the attribute's value: equal to or not equal to the value.
2. The processing is iterative. Once the entire n-dimensional space is partitioned, the two parts obtained above are reselected according to step 1.

The Midpoints of continuous variable attribute-value pairs act as splitting points for variable attributes. There will be $\boldsymbol{m-1}$ split points, each of which is the mean of two consecutive values, if a set of $\boldsymbol{m}$ samples contains $\boldsymbol{m}$ consecutive values. The amount of decontamination is calculated as the sum of impurities before separation minus the percentage of separated impurities of each node after separation. Each attribute is divided by the amount of impurities that can be reduced. Indicators of impurity are frequently measured using the $\boldsymbol{Gini}$ coefficient. The $\boldsymbol{Gini}$ impurity of a node can be defined as the formula (14), assuming a sample has a $\boldsymbol{G}$ class.

$$\boldsymbol{Gini}(A) = \mathbf{1} - \sum_{i=1}^{C} \boldsymbol{p_i}^2 \qquad \textbf{(14)}$$

In this case, $\boldsymbol{p_i}$ stands for the likelihood of falling under class $i$. All samples fall into the same class when$\boldsymbol{Gini}(\boldsymbol{A}) = \mathbf{0}$. When all classes are represented in the node with an equal probability, $\boldsymbol{Gini}(\boldsymbol{A})$ is maximized, and the resulting value is $C(C-1)/2$.

The actual recursive partitioning procedure is as follows, in accordance with the theoretical foundation mentioned above: This node becomes a non-leaf node when none of the samples in the current node belong to the same class or when there is only one sample left. Therefore, the attributes of each sample and the split points corresponding to each attribute are tested to determine the maximum split for the contamination variable. The ideal branch is the attribute division sub tree. It is clear that each branch node affects the CART method's accuracy. However, the branch is trimmed when the binary tree is created, which causes the branch data abnormality to persist throughout the entire classification process and lower CART classification accuracy.

CART employs a series of surrogate splits, which are split on different variables that stand in for the preferred split when the latter is not applicable due to missing values, to handle missing data values at a node. An importance score is also given for each $x$ variable using surrogate splits. These ratings, which gauge how well the surrogate splits forecasted the preferred splits can aid in the identification of masking. Linear splits, or splits on linear combinations of variables, are another technique that CART can use [118].

### II.2.4. Performance Metrics in Machine Learning

Evaluating the performance of a machine learning model is one of the most important steps in its development. Various metrics, also called performance metrics or evaluation metrics are used to assess model effectiveness and quality. You can use these performance metrics to assess how well your model handled the data it was given. Therefore, optimizing the hyper-parameters can improve the performance of the model. Performance metrics can be used to assess how well a generalized machine learning (ML) model performs on new or previously unexplored data. Every machine learning task or problem is classified again and again. Not all metrics are applicable to all problems. As a result, understanding the appropriate measures is critical. For the regression and classification tasks, different scoring measures are used. This section goes over data classification and regression.

#### II.2.4.1. Performance Metrics for Classification:

Based on the training data, a classification task identifies categories or data classes. The model learns from a certain set of data based on the training process and classifies the new data into classes or groups. The predicted output includes class labels such as "yes" or "no", "0" or "1", "spam", or "non-spam". The performance of the classification model is evaluated by different parameters. Among them are: Accuracy, Confusion Matrix, Precision, Recall, F-Score, AUC (Area Under the Curve).

##### *A. Accuracy:*

The accuracy metric can be calculated as the ratio of correct predictions to total predictions and is one of the simplest classification metrics to implement. It can be expressed in formula (14) as follows:

$$\mathbf{Accuracy} = \frac{\mathbf{Number\ of\ Correct\ Predictions}}{\mathbf{Total\ number\ of\ Predictions}} \qquad (\mathbf{15})$$

Easy to use and practice, but only works in situations where each class has the same number of samples. The classification accuracy is excellent, but I have the impression that a high level of accuracy is achieved.

##### *B. Confusion Matrix:*

The confusion matrix is a tabular representation of the predicted results of a binary classifier used to describe the performance of a classifier model for a test dataset when the true values are known. The confusion matrix is simply used. Typical binary classifier confusion matrix (but can be extended for use with classifiers with more than two classes). Table 3 is divided into four general terms, as follows:

**Table 3: Confusion matrix.**

| | | **Predicted condition** | |
|---|---|---|---|
| **Actual condition** | **Total population = P + N** | **Positive (PP)** | **Negative (PN)** |
| | **Positive (P)** | **True positive (TP)** | **False Negative (FN)** |
| | **Negative (N)** | **False positive (FP)** | **True negative (TN)** |

- ***True Positive (TP):*** In this case, both the predicted results and the reality agree.
- ***True Negative (TN):*** In this situation, both the prediction and the actual result are wrong.
- ***False Positive (FP):*** In this case, the predicted result is theoretically correct, but practically incorrect.
- ***False Negative (FN):*** The prediction in this situation is wrong, but the reality is true.

### *C. Precision:*

Accuracy metric constraints are overridden by the accuracy metric. The proportion of correct positive predictions is accurately determined. This can be calculated as "true positives" or as the percentage of positive predictions that come true (true positives and false positives) Calculated in formula (16) as:

$$\boldsymbol{Precision = \frac{TP}{(TP + FP)}} \qquad (16)$$

### *D. Recall or Sensitivity:*

It measures the percentage of falsely identified true positives and is comparable to the precision metric. The total number of positives can be calculated from true positives or actually correct predictions (true positives and false negatives), regardless of whether they were correctly predicted positives or falsely predicted negatives. This is how the recall is calculated in formula (17) as::

$$\boldsymbol{Recall = \frac{TP}{(TP + FN)}} \qquad (17)$$

According to the definitions, recall evaluates the classifier's performance on false-negative results while accuracy sheds light on false-positive result performance; regarding the aforementioned accuracy and recall. Therefore, the recall should be as close to 100% as possible to reduce false negatives, and the accuracy should be as close to 100% as possible to reduce false positives. In other words, increased precision leads to lower FP errors and increased recall leads to lower FN errors.

***E. F-Score:***

F-score (also known as F1-score) is a binary classification model evaluation metric based on positive class prediction. Calculations are made use of precision and recall. This is a type of composite evaluation that takes precision and recall into account. As a result, the F1 score is calculated as the harmonic mean of precision and recall, weighing both equally. The F1 score is calculated using the following formula (18) as:

$$\boldsymbol{F1 - score = 2 * \frac{precision * recall}{precision + recall}} \qquad (\mathbf{18})$$

Since both precision and memory are essential to the F-Score, it should be used when one of them (precision or memory) is slightly more crucial to take into account than the other; when false positives are typically more significant than false negatives, for instance, or the opposite.

***F. AUC (Area Under the Curve) ROC:***

AUC-ROC curves can be used when the results of a classification model need to be displayed on a graph. It is one of the most well-known and important metrics for evaluating the effectiveness of classification models. Let's start by understanding the ROC curve (Receiver Operating Characteristic Curve). An ROC curve is a graph that shows how well a classification model performs at different thresholds. A curve between the two parameters is shown as follows: True Positive Rate, False Positive Rate.

Since TPR or True Positive Rate is synonymous with Recall, it can be calculated in formula (19) as follows:

$$\boldsymbol{TPR = \frac{TP}{(TP + FN)}} \qquad (\mathbf{19})$$

FPR or False Positive Rate can be calculated in formula (20) as follows:

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{FP} + \boldsymbol{TN}} \qquad (\boldsymbol{20})$$

A logistic regression model can be evaluated multiple times with different classification thresholds to calculate values for each point on the ROC curve, but is not very effective. A useful way of doing this is by its AUC. Instead of focusing on the absolute value of the predictions, we should use the AUC to assess how well the predictions are ranking. In addition, the accuracy of the model prediction is evaluated without considering the classification threshold.

AUC is not always desirable as it requires adjustment of the stochastic output and is scale invariant. Additionally, AUC is not a useful metric when there is a large difference between false positive and false negative costs and it is difficult to reduce just one type of classification error.

#### II.2.4.2. Performance Metrics for Regression:

Finding relationships between dependent and independent variables is the goal of supervised learning techniques known as regression. Numeric or discrete values are predicted by predictive regression models. Regression metrics differ from classification metrics in a number of ways. This means that the accuracy metric (described above) cannot be used to evaluate regression models. Rather, the performance of regression models is reported in terms of prediction error. It lists the most commonly used metrics for evaluating the effectiveness of regression models: **Mean Absolute Error, Mean Squared Error, R2 Score, And Adjusted R2**.

## II.3. Deep learning approaches for sentiment analysis:

Deep learning (DL) extends classical ML by trying to add further ***"depth"*** (complexity) to the model and transforming the data with numerous functions that enable the data representation in a hierarchical form through many abstraction levels [119].DL is a research trend that is gaining traction in the increase of machine learning and artificial intelligence. It employs deep neural networks to predict the human brain's learning process and extract features from large-scale data (sound, text, images, etc.) in an unsupervised manner [120]. Deep learning has become an important research area of artificial intelligence because of its ability to provide more accurate and efficient results on various tasks. This is an evolving area of research with a wide range of applications that add to the cost benefits commonly expected during maintenance and repair work. Deep learning is a key technology and machine learning algorithms help to model abstract, high-level abstract views of data by processing layers with complex structures.

It is a rapidly expanding field that makes use of a variety of potent techniques and vast amounts of computing power to enable machines to recognize objects and instantly translate spoken language. The two main components of deep learning are: models with multiple non-linear information processing stages or layers, and supervised or unsupervised learning methodologies for feature extraction at an abstraction level [121]. Deep learning is a subset of AI and machine learning. The model learns possible spatial representations of the most basic forms of data; in existing deep learning approaches. The connection between deep learning, AI and machine learning as computational efficiency increases, research on artificial intelligence (AI) and deep learning (DL) is gaining momentum. In particular, AI, ML, and DL are inclusive relationships, as shown in Figure 5.

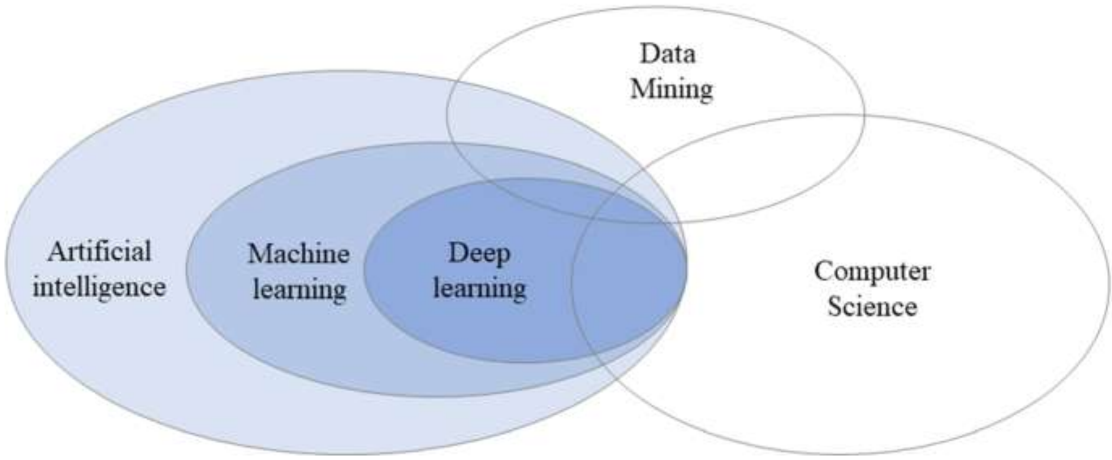


**Figure 5: The relationships between AI, ML and DL.**

Artificial Intelligence (AI) is a science like biology or mathematics that investigates how to create intelligent machines that can think outside the box to find solutions to issues. In terms of machine learning (ML), a branch of artificial intelligence that gives systems the capacity to automatically learn from experience and develop without being explicitly programmed More specifically, DL, or deep neural learning, is a type of ML that learns to represent the world as a nested hierarchy of concepts without manually extracting features. It achieves great power and flexibility in this way [122].

In deep learning implication, the input data is iterated into each layer, which multiplies the data using matrices. Typically, the input for one layer is used as the output for the next. The output of the final layer's processing of the data is either a feature or a classification output. A neural network is referred to as a deep neural network when it has numerous layers that are sequentially added to the model (DNN). When training a deep neural network, estimation occurs in reverse. Different iterations are performed through the layers, starting with the last layer and ending with the first layer, in order to optimize the parameters of each layer of matrix multiplication, taking into account the ground truth training conditions. The stochastic gradient descent is usually used as the algorithm. A "mini-batch" of samples is chosen at random for each pass in order Update the gradients in the direction of decreasing training losses (training losses are defined as the difference between prediction and ground truth). A training epoch is a single iteration through the entire training dataset.

Because the model is intelligent, it uses deep learning algorithms to automatically separate high-level latent dream features from the raw data. A deep learning model has multiple design layers. Lower-level layers are in charge of extracting simpler features of the input data, whereas higher-level layers are in charge of extracting more complex features. Accuracy and security in DL are impacted by the number of layers. The globalized data sequences are learned by the machine learning algorithms with the aid of data representation methods. For machine learning algorithms use time series data to make the right decisions, data quality is thus crucial [123]. More specifically, no matter how complex the algorithm is, the model will not work well with data that has insufficient descriptions. Therefore, to feature engineering is taken into consideration because it can help with data reconstruction by taking advantage of the sets of features in the raw data.

DL is particularly good and fast at solving more complex problems because it uses more complex models and allows massive parallelization. Given a large enough problematic dataset, these complex models used in DL can improve classification accuracy or reduce errors in regression tasks. Examples of different components that make up DL are convolutions, pooling layers, fully connected layers, gates, memory cells, activation functions, encoding/decoding schemes, etc. [124].

Both supervised and unsupervised learning can be done using deep learning. Depending on the application domain for which deep learning is used, different success measures are used. For example, when averaging across different object categories, the mean (map) average accuracy in object detection evaluates the overlap of the object's predicted location with that of the object's predicted location. Reality of the object the accuracy of the machine translation can be assessed using the BLEU score index, which compares the candidate translation with several basic truth reference translations. Throughput, latency, and power are examples of general system performance metrics that are not application-specific. Table 4 summarizes these metrics.

**Table 4: Common performance metrics.**

| **Performance metrics** |
|---|
| **Latency (s)** |
| **Energy (mW, J)** |
| **Concurrent requests served (#)** |
| **Network bandwidth (Mbps)** |
| **Accuracy (application-specific)** |

Deep learning techniques efficiently use complex non-linear functions to obtain distributed and hierarchical entity representations from labeled and unlabeled data. Deep Neural Networks (CDNN), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), Artificial Neural Networks (ANNs), and Long-Term Memory (LSTM). Various deep learning techniques and architectures such as: Convolution neural depth.

A neural network is made up of numerous neurons. These neurons are connected to one another through an intricate web of connections, which enables them to process information quickly and precisely. You can think of each neuron as a little information processing unit. To create the entire deep neural network, the neurons are connected to one another in a specific way. Because of the connections between the neurons, the network can process a lot of input and produce accurate results. Deep learning is the process by which the network's hidden layers multiply to form additional layers. Deep learning is more complex as it requires layer-by-layer initialization and batching to solve the difficult deep network training problem. Deep learning or inference iterates the input data in each layer and multiplies the data using matrices.

Typically, the input for one layer is used as the output for the next. Either a feature or a classification output is produced after the final layer has finished processing the data. A neural network is referred to as a "deep neural network" when it has numerous layers that are sequentially added to the model (DNN). We are dealing with a special case of DNNs where the matrix multiplication involves convolution filter operations, as is common in DNNs for image and video analysis. Such models are called Convolutional Neural Networks (CNN). Repetitive Neural Networks (RNNs) [125], with loops that maintain state on layer connections and allow prediction of sequential inputs are another type of DNN specifically designed for time-series prediction. The current state is supplied with the output of the previous state. RNN's hidden layers are capable of remembering data. Based on the output produced in the previous state, the hidden state is updated. Because RNNs have what is known as long short-term memory, which allows them to recognize previous inputs, they can be employed for time series prediction. The two most popular variations of RNN, referred to as gated RNN, are LSTM and Gated Recurrent Units (GRU) models. A gated RNN is a cell that contains an internal recurrence loop in place of the traditional RNN's interconnected is hidden units. Importantly, the gates in this model control the information flow. Modeling longer-term sequential dependencies are where Gated RNN excels.

### II.3.1. Types of Deep Learning Networks:

Depending on the techniques and architectures used for specific applications such as synthesis, classification, and detection, deep learning networks fall into three classes. They are categorized as follows:

I. Unsupervised deep learning network
II. Supervised deep learning network
III. Hybrid deep learning networks

When there is no defined target class, unsupervised deep learning networks gather higher-order correlation data for synthesis. We only have the input data and no corresponding output to map in unsupervised learning. By simulating the distribution of data, this learning aims to learn about the data. It is possible to use algorithms to unearth the intriguing structure that exists in the data. Association and clustering problems are solved by unsupervised learning. Unsupervised learning methods such as the K-Means algorithm [126], are used to solve clustering problems and the Apriori algorithm is used to solve correlation problems.

In unsupervised pre-training, the model is trained unsupervised before being applied to prediction. The dimension of the data is reduced using autoencoders, which are also used to solve problems involving novelty and anomaly detection. The first layer of the autoencoder is built as a coding layer and its counterpart is the decoder. Then use an unattended approach to training and reproduce your inputs. Attach weights to this layer after your workout. Once all layers of the Deep Net have been pre-trained, we move on to the next layer. Coming back to the classification/regression problem I originally tried to solve with deep learning, I start with the weights learned before training and optimize them using stochastic gradient descent. An autoencoder network consists of two components. The encoder converts the input into a latent space representation, which is indicated in formula (21) as follows:

$$\boldsymbol{h} = \boldsymbol{f}(\boldsymbol{x}) \qquad (\mathbf{21})$$

The decoder reconstructs the input from the latent space representation, which is represented in formula **(22)** as:

$$\boldsymbol{r} = \boldsymbol{g}(\boldsymbol{h}) \qquad (\mathbf{22})$$

In essence, autoencoders can be described as in Eq. (1).R is the decoded output, which will be similar to input in formula **(23)**:

$$\boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{x})\big) = \boldsymbol{r} \qquad (\mathbf{23})$$

A structure is created from the input data, and AE stores the most important derived dimensions in that structure. It is comparable to conventional methods for reducing the number of dimensions, such as Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). It is possible to train deep autoencoder networks using a technique called stacking. Among the Autoencoder subtypes are:

1. Sparse Autoencoder
2. Variational Autoencoder
3. Denoising Autoencoder

- ***Deep Belief Network:*** Learning how to function in the first layer is the first step in training a deep belief network. Then use the activation of the trained function on the next level. Continue until you reach the final layer. Restricted Boltzmann machine (RBM) is architecture for unsupervised learning that teaches the representation of input data. Although RBMs estimate the probability distribution of the available input data, it is almost identical to AE. As a result, it is considered to be a generative model where the underlying process generated the data.
- **Generative Adversarial Networks:** consists of a discriminator network and a generator network. The content is created by the generator, and it is verified by the discriminator. A discriminator determines whether an image looks natural and a generator produces a natural looking image. GAN is regarded as two-player minima algorithm convolutional and feed forward neural networks are used by GANs.

In supervised deep network learning, pattern classification identification is done by representing class distributions that are familiar with visible data. A deep network with discrimination is another name. Hybrid deep neural networks use both generator and discriminant elements. In addition, we use the convergence of homogeneous convolutional neural network (CNN) classifiers to build hybrid deep neural network models. The CNN classifier is trained to produce an output of 1 for the predicted class and an output of 0 for all other classes.

### II.3.2. Deep Learning Methods:

In the next section, we discuss some powerful methods that deep learning algorithms can use to reduce training time and improve models. The merits and demerits of each method are comprised in Table 5.

- ***Backpropagation:*** The main component of neural network training is back propagation, which is used to adjust neural network weights established in the previous epoch. As for the multilayer artificial neural network (ANN), it is a supervised learning technique used in pattern recognition, classification, and medical diagnosis. It can be used to calculate the slope of the function during iterations of the approach based on the optimization problem.

- ***Stochastic Gradient Descent:*** Convex functions can be used in gradient descent algorithms to find the best minimum without getting stuck in local minima. Depending on the value of the function and the learning rate or step size, the optimal value can be determined in different ways.
- ***Max Pooling:*** Max pooling applies a predefined filter to non-overlapping sub-ranges of the input, producing an output equal to the maximum of the values found within the window. Using max pooling not only reduces the dimensionality, but also reduces the computational load when learning multiple parameters [127].
- ***Dropout:*** Dropout methods can be used to solve overfitting problems in deep neural networks. With this technique, units and their connections are randomly removed during training [126]. To reduce overfitting and increase generalization error, Dropout provides powerful regularization techniques.
- ***Decay at Learning Rate:*** By changing the learning rate, the stochastic gradient descent algorithm performs better and requires less training time. A widely used technique is to gradually decrease the learning rate. This means making big changes first and then making changes throughout the training process.
- ***Batch Normalization:*** By minimizing covariate shift, batch normalization speeds up deep neural networks. When the weights are updated during training, the input is he normalized to one level per mini-batch. Regularization shortens training epochs while stabilizing learning. You can normalize the outputs of previous activation layers to improve the stability of your neural network. [128].
- ***Skip-gram:*** can be used to model word embedding algorithms. According to the Skip-Gram model, two lexical terms are equivalent when used in the same context. For example, the word "is a mammal" is used in the meaningful sentences "a cat is a mammal" and "a dog is a mammal". Skip-Gram can be implemented by considering a context window of n terms, omitting one of the terms to train a neural network, and using the model to predict the skipped term. [129].

- ***Transfer Learning:*** A model trained for one task is applied to the other related task. Another network that is trained on related problems can benefit from insights gained from solving specific problems. This will allow us to solve the second problem faster and with better performance [130].

**Table 5: The advantages and disadvantages of each method.**

| Method | Description | Merits | Demerits |
|---|---|---|---|
| **Backpropagation** | Used in the optimization problem | For calculation of gradient | Sensitive noisy data |
| **Stochastic gradient descent** | to solve optimization problems by determining the best minimum | keeps trapping to a minimum locally | Longer convergence time and more computation |
| **Learning rate decay** | Reduce learning rate gradually | improves performance while cutting down on training time | Computationally expensive |
| **Dropout** | Drops out units/ connection during training | Avoids overfitting | Increases the number of iterations required to converge |
| **Max pooling** | Applies a max filter | Reduces dimension and computational cost | Only takes into account the largest element, which can sometimes produce unacceptable results. |
| **Batch normalization** | Batch-wise normalization of input to a layer | It reduces covariant shifts; increases network stability, trains networks faster, and allows higher learning rates. | computation costs when training |
| **Skip-gram** | Used in word embedding algorithms | Can handle any raw text and uses less memory | The Softmax function is computationally |

| | | | |
|---|---|---|---|
| | | | costly, and training time is lengthy. |
| **Transfer learning** | Knowledge of the first model is transferred to the second problem | Improving performance, rapid progress in the second issue of training | Works with similar problems only |

### II.3.3. Deep Learning Applications:

Deep learning applications are being explored in new areas such as medical image processing, deep learning for sound generation, deep learning in art creation, robotics, prediction, computer games, self-driving cars, and big data. Deep learning can be applied to automatic coloring, automatic machine translation, automatic text generation, automatic handwriting recognition, automatic caption generation, advertising (through the development of data-driven predictive advertising, actual real-time , targeted display advertising), earthquake prediction, customer insights, automated speech recognition, self-driving cars, image colorization, computer vision, video colorization, deep learning AI, healthcare, speech recognition, mobile advertising [131].

### II.3.4. Deep learning in Natural Language Processing (NLP):

Natural Language Processing (NLP) is a branch of computer science, human language, and artificial intelligence, as shown in Figure 6. This technology is used by machines to understand, evaluate, and manipulate and human language perception. It helps developers organize knowledge for tasks like translation, machine summarization, named entity recognition (NER), language recognition, relationship extraction, and topic segmentation.

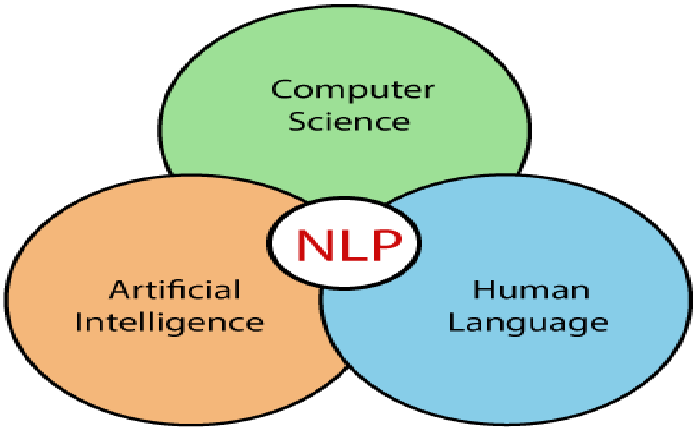


**Figure 6: Constituents of NLP.**

Natural language processing, also known as computational linguistics is the engineering of computer models and processes to solve practical issues related to human language understanding. These useful software approaches are employed to make. Even though it is occasionally difficult to determine which areas problems relate to, NLP tasks could be divided into two broad sub-areas: core areas and applications. The core area solves fundamental problems such as language modeling with a focus on quantitative measures of relatedness between natural source words morphological processing. It segments constituent word elements and recognizes the real parts of words when they are used. Syntactic processing or analysis that creates sentence diagrams as potential precursors to semantic processing that seeks to simplify the meaning of words, phrases, and sentences as they are used; Semantic processing attempts to extract the meaning of words, phrases and sentences as they are used. Application areas include concepts such as information extraction (e.g. Named entities and relationships), translation of text between languages, summarization of written work, automated question answering leading to answers, classification and clustering of documents. You often need to successfully master one or more core problem and apply those ideas and techniques to solve real-world problems.

There are two main processes in natural language processing: **Data Pre-processing, Algorithm development.**

- **Data pre-processing:** is the process of getting text data ready and clean for machine analysis. It highlights textual elements that an algorithm can work on and transforms data into a form that can be used. Pre-processing data are a part of the following methods.

- ***Tokenizing:*** It means dividing text or sentences into smaller parts. However, for the algorithm to understand these sentences, each word in the sentence must be captured and explained to the algorithm individually. Therefore, it splits the sentence into individual words and stores them. This is what tokenization means, each word is referred to as a token.
- ***Removing Stop Words:*** By removing common words and keeping unique words, you can speed up the learning process by eliminating irrelevant words that give little meaning to the statement and only make the statement more coherent. Stop words are words that can be removed, such as was, in, is, the.
- ***Stemming:*** Stemming is a technique for converting words into their base form, or root form. It is the process of obtaining a word's stem. When affixes are added to word stems, they generate new words. Jumps, jumped, and jumping, for example, all stem from the same root word, "jump." The main issue with stemming is that it sometimes produces a root word that has no meaning.
- ***Lemmatization:*** The lemmatization and Stemming are quite similar. It is employed to group the word's various lemmatized forms. The primary distinction between lemmatization and stemming is that the latter results in the root word, which has a meaning. The root indicated the new base form of a word already in the dictionary from which the word is derived. A word's tense, mood, gender, etc. can also be used to determine the root noun of that word.
- ***Part-of-speech tagging:*** Using part-of-speech tags Parts of speech like nouns, verbs, and adjectives are used to tag words. It describes how a word works grammatically and conceptually within sentences. Depending on the situation in which it is used, a word may belong to one or more parts of speech.

When preprocessing is finished, an algorithm is created to process the language. The following algorithms are commonly used:

- **Rules-based system:** Rules-based systems are an older practice that uses rules-based algorithms.
- **Systems based on machine learning:** Machine learning algorithms use statistical methods. They learn to complete tasks using hands-on data and refine their techniques as new data is processed. Using a combination of

machine learning, deep learning, and neural networks, natural language processing algorithms develop their own rules through iteration and learning.

NLP has two subdivisions: Natural Language Understanding (NLU) and Natural Language Generation (NLG).

- **Natural Language Understanding (NLU):** It is a technique that leverages content concepts, entities, keywords, emotions, relationships, and semantic roles to extract metadata useful for machine comprehension and human language analysis. NLU is primarily used in business applications to understand customer issues both verbally and in writing. Two main techniques are used to process and understand both text and speech. As follows:

***Syntax analysis, Semantic analysis.***

- ***Syntax:*** refers to sentence components. NLP uses syntax to analyze the language's meaning based on grammatical rules. The syntax method comprises the following components:
  - ✓ ***Parsing***: is the linguistic analysis of grammar.
  - ✓ ***Word segmentation:*** distinguishes between the word and the void of white space. In lengthy texts.
  - ✓ ***Sentence breaks:*** define the boundaries between sentences.
  - ✓ A word is divided **morphologically** according to the number of syllables it contains. Speech recognition and machine translation both benefit from this segmentation.
  - ✓ ***Stemming***: words are separated into their root words.
- ***Semantics***: Semantics deals with how words are used and what they mean. NLP employs algorithms to decipher the structure and meaning of sentences. Semantic methods comprise:
  - ✓ ***Word sense disambiguation:*** interprets the meaning of a word based on context.
  - ✓ ***Named Entity Tagging:*** The process of name detection involves identifying a named entity, such as a person's name, a movie's name, an organization's name, or the name of a place. This uses semantics to group words into categories.

- ✓ ***Language identification:*** Different languages are distinguished from one another based on their peculiarities and rules. Prior to the algorithm taking other actions, the language must be identified. Deep neural networks aid in language translation.
- ✓ ***Optical Character Recognition :*** This technology converts digital images or textual content into machine-readable text. You can use OCR in NLP to digitize physical documents by scanning them.
- ✓ ***Speech recognition***, also known as ***speech to text***, converts live or recorded audio into text documents. Machine learning algorithms can extract features such as intonation patterns, which trigger phoneme sequences associated with specific word types. It will produce better results than a transcription dictionary or language model.

- **Natural Language Generation (NLG):** As a translator, you convert computerized data into natural language representations. Basically, it consists of text planning, sentence planning, and text realization. New text is created. NLG focuses on his three main areas. They are:
  - ✓ Data to text takes unstructured data as input and outputs text.
  - ✓ Text-to-text: like summarization, combines sources to produce text output.
  - ✓ Dialogue: text is generated conversations, like a Chabot.

The differences between NLU and NLG are outlined in Table 6.

**Table 6: difference between NLU and NLG.**

| NLU | NLG |
|---|---|
| **The process of reading and interpreting languages is called NLU.** | The process of creating or generating languages is called NLG. |
| **Generate non-speech output based on natural language input.** | Generate natural language output by constructing natural language output from non-speech input. |

In particular, it makes recommendations for how to represent, handle, and examine the vast amount of natural language data. The signal processing community is increasingly engaged in research on document text and language. Deep learning helps model language by assigning probabilities to the order of linguistic symbols or words. When using punch cards and batch processing, analysis of a set can take up to 7 minutes. Today, thanks to Google and similar services, we can process millions of web pages in less than a second. [132]. NLP enables computers to perform a variety of natural language-related tasks at all levels, from parsing and part-of-speech tagging (POS) to machine translation and dialogue systems. Semantic processing, which aims to capture the meaning of words, phrases, and higher-level components in text, and syntactic processing (parsing) both produce sentence figures containing potential precursors to semantic processing. Application areas include topics such as information extraction (such as named entities). Language-to-language text translation, Language-to-language text translation, written work summarization, automated question answering using reasoning, Document classification and clustering, Chatbots, speech recognition, spell checking, machine translation, Sentiment analysis, spam detection. In order to successfully address one or more of the fundamental issues, one frequently needs to use those concepts and methods to address real-world issues [133].

NLP enables users to ask questions about any topic and receive a direct answer in a matter of seconds. It provides precise responses, meaning it does not provide extraneous or irrelevant information, and it enables computers to converse with people in their native languages. It saves a lot of time. The majority of businesses employ NLP to increase the accuracy and efficiency of documentation processes as well as to extract information from sizable databases. On the other hand, NLP might not display context. It is erratic and might necessitate more keystrokes. NLP has limited use and is unable to adapt to the new domain. Because of this, NLP is designed exclusively for a single, narrow task.

Deep learning certainly has advantages and challenges when applied to natural language processing, as summarized in Table 7.

**Table 7: Advantages and challenges of deep learning for natural language processing.**

| **Advantages** | **Challenges** |
|---|---|
| Excellent for pattern recognition problems | Difficulty in drawing conclusions or making decisions |
| Data driven and performant for many problems | Inability to process symbols directly, high computational cost of learning |
| Complete training: Little or no domain knowledge required to set up the system | Not suitable for small amounts of data as it consumes a lot of data |
| Learn expressions: Cross-modal processing possible | The long tail phenomenon is difficult to deal with and still lacks a rationale |
| Gradient-based learning: Simple learning algorithm | Models are usually black boxes and hard to understand |
| mainly supervised learning methods | We need to develop a method of unsupervised learning. |

End-to-end training and representation, learning are two benefits that set deep learning apart from more conventional machine learning techniques and make it an effective tool for processing natural language. End-to-end deep learning training is frequently possible for an application. This is so that information in the data can be effectively "encoded" in the model, thanks to the deep neural network's rich representability [134]. The representations of data in various formats, such as text and image, can all be learned as real valued vectors using deep learning. This enables the processing of information across various modalities.

Deep learning faces a number of difficulties that are more typical, such as the lack of a theoretical basis, the model's lack of interpretability, and the need for a lot of data and powerful computing resources. In addition, there are additional problems unique to natural language processing, such as the inability to process symbols directly, the difficulty of processing long tails, and the inefficiency of reasoning and decision making.

Data always have a power law distribution in natural language. As a result, for instance, the vocabulary grows in size in tandem with the growth of the data. This implies that there will always be cases that the system cannot handle, regardless of how much data there is for training. Deep learning is faced with a significant challenge in figuring out how to handle the long tail problem.

The most popular choice for modeling challenging natural language tasks is statistical NLP. However, when it first started, it frequently experienced the infamous dimensional space curse while starting to learn joint probability features of language models. The desire to learn words' feature representation in low-dimensional space was sparked by this.

This section looks at deep learning (DL) based work in sentiment analysis in general, because the primary goal of this research has been to show how deep learning can improve sentiment analysis; Convolutional neural network (CNN) model for sentiment analysis at the sentence level. The model was trained using sentence-level and character-level features. The model was tested on two different datasets, including movie reviews from the Stanford Sentiment TreeBank (SSTB), which had an accuracy rate of 85.7%, and tweets from the Stanford Twitter Sentiment Corpus (STS), which had an accuracy rate of 86.4%. A phrase recurrent neural network (PhraseRNN) uses a word2vec version pre-trained on the Google News dataset to recognize phrases using sentence dependencies and construction trees, as well as external word embeddings. Was trained to the model showed promising results when tested using a reference data set. We develop state-of-the-art methods to enhance hybrid neural networks (HNNs) with semantic features based on topic modeling for social sentiment classification. Latent Semantic Machine (LSM) models and collaborative transfer learning were used in this approach to extract semantic features for training the proposed hybrid neural network. Compared to related studies based on traditional neural networks, our approach using three reference data sets yielded significant results.

Several of the papers proved the benefit of DL in terms of reduced effort in feature extraction. Hand-engineered components require a significant amount of time, which DL automates. Furthermore, manually searching for good feature extractors is not always an easy and obvious task. The ability to create simulated datasets to train models that can be correctly created to address real-world problems is referred to as deep learning. So even though DL takes more time to learn than other traditional approaches (such as SVM and RF), it has highly desirable performance.

The requirement for large input data sets in the training process is a significant drawback and an impediment to the use of DL. But besides data-enhancement methods that dramatically improve a few datasets with transformations keeping the labels, hundreds of images are needed depending on the complexity of the problem under study (i.e. Number of layers, required precision, etc.). While the data annotation is the appropriate use of technology in most cases, some tasks are more complex and require experts (who can be difficult to recruit) to annotate input images. . In some cases, experts or volunteers are prone to mistakes when labeling data, especially when the task is difficult.

### II.3.4.1. Word embedding:

Word embeddings or distributional vectors essentially follow the distributional hypothesis, which states that words with comparable meanings tend to appear in similar contexts. Similar contexts frequently contain similar meanings. Thus, these vectors aim to represent the traits of a word's neighbor. The main advantage of distribution vectors is their ability to identify word similarities. Vector similarity can be measured using techniques such as cosine similarity [132].

Word embeddings are typically pre-trained and learned before being implemented on large texts to capture the syntactic and semantic features of text document collections. Learn word representations in context using immutable and reusable embeddings [135].

Word embeddings, which are real-valued representations of words generated by distributional semantic models (DSMs), are one of the most widely used tools in modern NLP, but their effect and limitations remain unknown. How to analyze and understand the performance of DSMs is among the most essential issues in the study of distributional semantics. There is still no scientific consensus about which analysis approach should be used: NLP engineers who are more ready to deal with regression tasks, For example, semantic role labeling commonly tests the efficiency of embeddings in such tasks, whereas computational language learners studying the nature of semantic information use experimental techniques derived from cognitive science and to study word embeddings.

In the past, shallow neural networks have been the models used to generate these embeddings; deep neural networks have not been required to generate effective embeddings. However, deep learning-based NLP models always use these embeddings to represent their words, phrases, and even sentences. In actuality, this is a key distinction between deep learning-based models and conventional word count-based models. For a variety of NLP tasks, word embeddings have produced cutting-edge results.

With advances in hardware technology and optimization technology, neural network models are gradually showing their advantages in various fields, and neural network-based word embeddings, which can more accurately reflect the words and their contexts, are common words. It is popular as It is a distributed representation technique and is important for various downstream tasks such as natural language inference, text classification, knowledge mining and named entity recognition [136], Noun Phrase Chunking, Sentiment Analysis, Shallow Syntax Parsing, Parse Tree Level Construction, Semantic Role Labeling, Negation scope, Part-of-Speech Tagging, Metaphor Detection, Paraphrase Detection, Textual Entailment Detection and input for artificial neural networks [137].

Word embeddings are considered *"good"* if they reflect a semantic understanding. At the same time, we are not sure whether our interpretation is correct. Furthermore, because there are many different types of word relationships, it is unclear which type of relationship the extracted features should represent (such as semantic relatedness and semantic similarity, the definitions of which are also quite obscure). It is also unclear if the model must be regarded *"bad"* if it doesn't take into account the semantic relatedness or similarity relationships between words. The majority of review datasets are not separated into the training and test sets. As a result, researchers prepare the word embeddings to adapt to the data, attempting to improve their quality. They are attempting to capture the relationships that exist in the data rather than the actual relationships between words. Statistically significant tests are not always carried out in key experiments involving new distribution models and evaluation methods.

The idea behind word semantic similarity is that the distances between words in an embedding space can be assessed using human heuristic decisions on the real linguistic distance and time between these words. Examiners are given a series of word pairs and asked to rate their similarity. The distances between these pairs are also collected in the word embedding space and her two resulting sets of distances are compared; the more similar, the better the embedding. It was also argued that the concept of semantic similarity acquired not only semantic connections between words, but also some morphological and grammatical characteristics of word representations.

Analogy of words in some works, this method (also known as analogical reasoning, linguistic regularities, and word semantic coherence) is the second most popular method of evaluating word embeddings. The major complaint leveled at this technique is the lack of a precise evaluation measure. If the cosine distance between word vectors seemed intuitively sufficient in the word semantic similarity task, its suitability for relationship transfer is called into question in this task.

Concept classification this method (also known as "word clustering") evaluates the space in a word's embeddings. The task is to divide a set of words into subsets of words belonging to different categories. The number of clusters must be determined. Potential criticisms of such methods could address the issue of choosing the best clustering algorithm or the best metric for evaluating clustering quality. The outlier detection method analyzes the same features of word embeddings as the word classification method (it also suggests clustering). However, the task is to define anomalous words with semantic information within already formed clusters, rather than a set of words within clusters. Subdivide a specific number of clusters.

Similar to the method of semantic similarity of words, synonym detection methods attempt to evaluate the embedding ability of words to form a vector space with predictable distances between words, but the absolute similarity is the proposed I don't. It is based on the concept that word similarity can be assessed by finding the most similar words compared to other sets of words.

Patterns of neural activation When a person reads, the meanings of the words are reflected in some patterns in her brain. As a result, such patterns could serve as input data for word embeddings. However, the consistency of such brain data has called into question because neural activation patterns do not correlate in many subjects due to differences in brain size and structure. Another issue is that it is unclear how much these patterns are related to lexical semantics and other linguistic data contained in words, such as the number of characters, stress location, syllable count, and so on.

### II.3.4.2. Word2vec:

Word2vec Google has suggested using a neural network to process the textual data. It consists of two learning models rather than just one algorithm.Word2vec a powerful tool that employs the Continuous Bag of Words (CBOW) and Skip-gram (SG) models to learn word representations from corpora. These two models can easily be applied to other downstream tasks and are effective at capturing word semantics [138]. The context words are supposed to be evenly situated to the target words within a window size in both directions. The prediction accuracy determines the word embedding dimension unsupervised settings. The accuracy of the prediction increases as the embedding dimension rises until it conforms at a certain point, which would be regarded the optimal embedding dimension because it is the shortest without sacrificing accuracy; and are important in a variety of downstream tasks such as natural language inference, text classification, knowledge mining, and named entity recognition.

Skip-Gram predicts the context of words, while CBOW predicts words based on context.Word2Vec creates word vectors by integrating a text corpus into a single learning model. In this method, Word2Vec first develops a vocabulary from a training text corpus and then learns vector representations of each word; the cosine distance between each word can also be determined by Word2Vec. As a result, grouping related words based on their distances is advantageous for us. The original feature dimension is projected to a new lower dimension by grouping related words.

The three components of the model are depicted in Figure 7, and the CBOW model's training goal is to discover word representations that can be used to predict the target word from its context words.

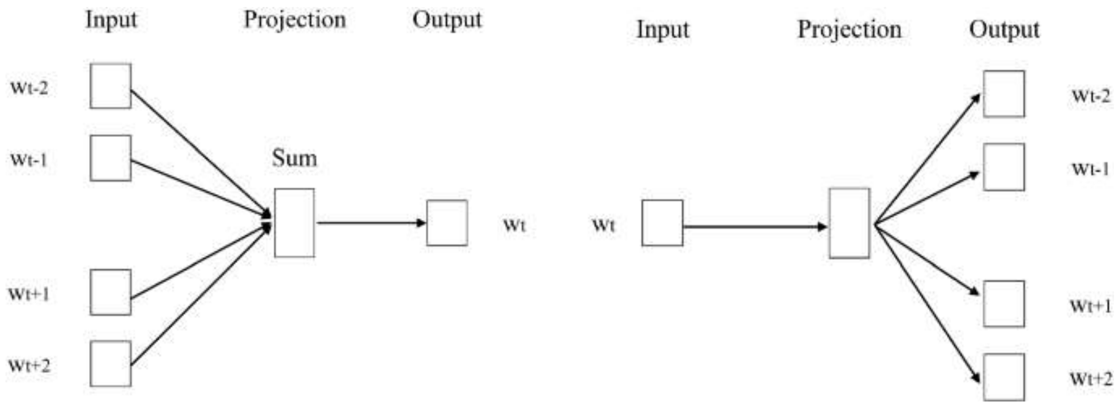


**Figure 7: The overall architecture of CBOW (left) and skip-gram (right) model.**

- **Skip – gram (SG):** In general, Word2Vec trains input with all other nearby words in context in one of two ways: Skip grams or continuous bags of words. Skip-Gram can predict the target context based on specific words [136]. The skip-gram model specifically seeks to increase the average log probability in formula (23) as:

$$\frac{1}{T}\sum_{t=1}^{T}\sum\nolimits_{-c\leq j\leq c\ j\neq 0} \log P(w_{t+j}|w_t) \qquad (23)$$

Given a set of words $w_1, w_2 \ldots\ldots w_t$ for training, $c$ is the size of the training context. The more training samples your model receives, the more accurate results it can provide. A basic Skip-Gram model defined below $P(w_{t+j}|w_t)$ the softmax function is presented in formula (24) as follows:

$$P(w_O|w_I) = \frac{exp(v'_{wO}T\ v_{wI})}{\sum_{w=1}^{W} exp(v'_w\ T\ v_{wI})} \qquad (24)$$

Where $v_w$ and $v'_w$ are the representations of vector inputs and outputs, $W$ is the amount of words in the vocabulary.

- **Continuous Bag of Words (CBOW):** This is a technology that uses context to predict target words. The word bag idea is based on this. NLP uses Bag of Words (BoW) techniques to optimize presentations. Represents text as a collection of words, regardless of grammar or word order [136]. By counting the frequency of each word as a feature, it is frequently used to train a classifier to categorize documents or texts. CBOW is a method of representing an unlimited number of features in a fixed-size vector when the number of features is not known in advance. The CBOW method works very much like the BoW method, which allows you to average or sum the embedding vectors of related vectors, with the following differences in formula (25) as:

$$\boldsymbol{CBOW}(\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots\ldots \boldsymbol{f_k}) = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} \boldsymbol{v}(\boldsymbol{f_i}) \qquad (\mathbf{25})$$

CBOW and SG have several similarities in their model architectures and training methods. Distributed vector representations of words are widely used in the machine learning and natural language processing communities at the time of writing. The majority of successful applications, on the other hand, are learned to use deep neural network architectures, which are slow to train on consumer hardware.

Word2vec creates a vector for each term, but additional work may be required to combine those vectors into a single vector or another format. In contrast, TF-IDF is a statistical measure that can be applied to terms in a document and then used to create a vector. Furthermore, word2vec considers the context of the words in the corpus, whereas TF-IDF does not.

#### II.3.4.3. Doc2vec:

The word embedding technique was created to allow computers to understand texts. Words are represented as vectors, and the system is based on artificial neural networks. Quoc Le and Tomas Mikolov created Doc2Vec, which creates a vector that represents the document in order to predict the target word [139]. The document's length is not taken into account when doing this. There are two different approaches. The doc2vec neural network-based document embedding learning algorithm is commonly in use in natural language processing (NLP) tasks such as text classification, sentiment analysis, information retrieval, and so on. The Doc2vec algorithm, which builds on Word2vec and uses a neural network to recognize words based on how frequently words arise together.

The ability to work with texts of varying length without losing their order or semantics when representing the documents in continuous vector space is one of the benefits of the doc2vec framework. A paragraph vector, or doc2vec, is a straightforward extension to word2vec for learning embeddings from words to word sequences. Doc2vec is unconcerned about the granularity of the word sequence. It could be a single word, a sentence, a paragraph, or a document. In earlier neural network approaches for word embedding, the next word in the set was attempted to be predicted by appending the inputs from several preceding word vectors. After the model had been trained, the word vectors were mapped into a vector space where words with similar semantic meanings have nearby locations and similar vector representations. We can quickly determine the degree of word similarity by applying basic algebraic operations to word vectors. For example, apply basic algebraic operations to word vectors to determine word similarity. For example, to find words that are synonymous with "largest" and similar to "small" and also similar to "large", use ***Vector X = Vector("largest") − Vector("large") + vector ("small")***. Then search the vector space for the word closest to X in terms of cosine distance [138]. For Doc2Vec, the model creates a new vector representing the entire document and combines it with the individual vectors for each word. The word order is maintained depending on the working mode, whether it is a distributed bag-of-words paragraph vector (PV-DBOW) or a distributed memory model of paragraph vector (PV-DM). PV-DM mode concatenates a paragraph vector and the corresponding word vector to predict the next word in the sequence, whereas PV-DBOW mode ignores the contextual words in the input and lets the model choose randomly from paragraphs predict the word output [140].

Numerous models that embed sentences or documents in Doc2vec have been proposed, including FastSent and Sentence-Bert. However, due to its quick learning time and high efficiency, document embedding using the Doc2Vec model is still widely used. Additionally, the document vector produced by Doc2vec is a vector that accurately depicts the order and context of words that appear in the patent document and is based on the co-occurrence frequency of words that appear simultaneously in a particular data window.

The domain-specific task has also seen attempts to use a Doc2Vec language model; a system that evaluates comments and their use in change-prone method analysis using a Doc2Vec-based language model. The study's findings suggest that Doc2Vec may be used for domain-specific tasks like comparing comments in source code to programming statements. A topic suggestion system built on a language model based on Doc2Vec. The Doc2Vec model-based automated topic recommendation system indicates that our proposed autonomous system for classifying text data related to cyber security may be implemented in practice [140].

### II.3.4.4. Term Frequency-Inverse Document Frequency (TF-IDF):

Mathematical concepts such as statistics, algebra, and calculus support machine learning algorithms. For example, consider numerical data in a two-dimensional array where rows are instances and columns are featured. The problem with natural languages is that the data are in the form of raw text that needs to be converted to vectors. Text vectorization is the process of converting text into vectors. This is a fundamental process of natural language processing, as no machine learning algorithm, even a computer, can understand the text. A popular approach to traditional machine learning algorithms, the text vectorization algorithm TF-IDF Vectorize helps vectorize text.

**The term Frequency-Inverse Document Frequency (TF-IDF)**: is an acronym for this term and a metric used in the information retrieval (IR) field. Machine learning evaluates the importance or relevance of string representations (words, phrases, lemmas, etc.) in documents within a group of documents (also called a corpus). The text vectorization process transforms the words in a text document into meaningful numbers. There are many different-vectorized scoring schemes for the text, but the TF - IDF is the most common. A word frequency in the corpus counteracts its importance and increases in proportion to its frequency in the text (data set).The TF-IDF algorithm, as its name suggests, vectorizes and scores words by dividing their Term Frequency (TF) by their Inverse Document Frequency (IDF).

**Term Frequency (TF):** A term or word in a document is measured relative to all other words in the document. The easiest way to determine this frequency is to simply count the number of times the word appears in the document. You can change the frequency using the length of the document or the frequency of the most common words in the document in formula (26) as:

$$\mathbf{TF} = \left(\frac{\textbf{number of time the terms appears in the document}}{\textbf{total number of terms in the document}}\right) \qquad (\mathbf{26})$$

- **Inverse Document Frequency (IDF**): Percentage of documents in the corpus that contain the term. Words that appear in a small portion of a document (such as technical terms) have higher importance scores than words that appear in all documents (such as a, the, and). It indicates how often or infrequently the word appears in the text. The closer a word is to 0, the more frequently it appears. This metric is calculated by dividing the logarithm of the total number of documents by the number of documents containing that word. So if the word is widely used and appears in many documents, this number tends to zero. Otherwise, it will look like this in formula (27) as:

$$\mathbf{IDF} = \mathbf{log}\left(\frac{\mathbf{number\ of\ the\ document\ in\ the\ corpus}}{\mathbf{number\ of\ documentsin\ the\ corpus\ containthe\ term}}\right) \qquad (\mathbf{27})$$

The $TF$ and $IDF$ scores are multiplied to determine a term's :$\boldsymbol{TF - IDF = TF * IDF}$

The fundamental assumption underlying $TF - IDF$ is that a term's significance is inversely correlated with its frequency across documents. $IDF$ provides information about the relative rarity of a term in the collection of documents, while $TF$ provides information about how frequently a term appears in a document. We can calculate our final $TF - IDF$ value by summing these values in formula (28) as:

$$\boldsymbol{TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)} \qquad \textbf{(28)}$$

Frequency in document d indicates how often the word t occurs. So, as you can see, it's only natural that words become more relevant when they appear in the text. The order of the concepts is not important, so vectors can be used to describe the text in the concept model pocket. The number of times a term appears in the document determines its weight in formula (29) as:

$$\boldsymbol{TF(t, d)} = \left(\frac{\boldsymbol{count\ of\ t\ in\ d}}{\boldsymbol{number\ of\ words\ in}}\right) \qquad (\mathbf{29})$$

- **Document Frequency (DF):** This evaluates the meaning of the text in the entire corpus collection that is very similar to $TF$. The only distinction is that in document$d$, $TF$ represents the frequency counter for a term $t$, whereas $DF$

represents the quantity of times the term $t$ appears in the document set $N$. In other words, there are $DF$ papers that contain the word in formula (30) as:

$$\boldsymbol{DF(t) = occurence\ of\ t\ in\ documents} \quad \textbf{(30)}$$

- **Inverse Document Frequency (IDF):** This test determines how relevant a word is. The primary goal of the search is to locate relevant records that meet the demand. Because $TF$ considers all terms to be equally relevant, the term's frequency can be used to calculate the term's weight in the paper.

To begin, determine the document frequency of a term $t$ by counting the number of documents that contain the term: $DF(t) = N(t)$ Where:

$\boldsymbol{(t) = Document\ frequency\ of\ a\ term\ t}$

and $\boldsymbol{N(t) = Number\ of\ documents\ containing\ the\ term\ t}$.

The IDF of the word is the number of documents in the corpus separated by the frequency of the text in formula (31) as:

$$\boldsymbol{IDF(t) = \frac{N}{DF(t)} = \frac{N}{N(t)}} \quad \textbf{(31)}$$

Although the element (most definite integers) is meant to be less significant than the more common word, it comes off as harsh. Next, we calculate the paper's inverse frequency's logarithm (with base 2). As a result, the term t's if becomes in formula (32) as:

$$\boldsymbol{IDF(t) = \log\left(\frac{N}{DF(t)}\right)} \quad \textbf{(32)}$$

The $TF - IDF$ metric is one of the best for determining how important a term is to a text in a series or corpus. The $TF - IDF$ weighting system gives each word in a document a weight based on its term frequency $TF$ and the reciprocal document frequency $TF - IDF$. Words with higher weight scores are thought to be more important.

Usually, the $TF - IDF$ weight consists of two terms: Normalized Term Frequency ($\boldsymbol{TF}$) and Inverse Document Frequency ($\boldsymbol{IDF}$) in formula (33) as:

$$\boldsymbol{TF - IDF(t, d) = TF(t, d) * IDF(t)} \quad \textbf{(33)}$$

Machine learning using natural language faces major obstacles. Its algorithms usually deal with numbers, and text is natural language itself. To convert this text to numbers, we need to go through a process called "vectorizing the text". This is a critical step in the data analytics machine learning process and the final result is highly influenced by the vectorization algorithm chosen. So make sure you get the desired result. TF-IDF scores can be fed into algorithms such as Naive Bayes and Support Vector Machines to convert words into numbers that machine learning algorithms can understand, greatly enhancing the output of simple techniques such as word count.

Using TF-IDF to determine the meaning of a document is similar to using Bag of Words: Preprocessing clean data entails finding TF and IDF for words, tokenizing words with frequency, normalizing the data (all lower case), lemmatizing data (all words to root words), and vectorizing vocabulary.

In the area of information retrieval, TF-IDF has applications, with search engines serving as one prominent example. Search engines can use TF-IDF to help rank search results based on relevance, with results most relevant to users with a higher TF-IDF score. Indeed, the TF-IDF can inform you of the relevance of a term according to a document. It can be used to extract keywords from the text as well. The words that received the highest scores were the ones that were most pertinent to the document, making them suitable to be used as keywords; easy to understand. This can be used to choose keywords (or even tags) for a document or to more effectively summarize articles.

Understanding how TF-IDF works will give you a better understanding of how machine learning algorithms work. Machine learning algorithms are traditionally good at dealing with numbers, but TF-IDF algorithms help decipher words by assigning them numbers or vectors. This was revolutionary for machine learning, especially in his NLP-related areas such as text analysis. The TF-IDF algorithm is suitable for data classification and keyword extraction in text analysis using machine learning. This means you can complete mundane tasks like support tickets, feedback lines, and date flagging in seconds. Only TF-IDF vocabulary levels are useful. Synonyms are not considered. This doesn't capture the semantics. The IDF gives higher weight to terms with lower DF and ignores the order of the terms, so the highest his TF-IDF score may be meaningless for the document's subject.

### II.3.4.5. Glove:

With regards to the word analogy and semantic relatedness tasks, GloVe [141], Log-bilinear regression models attempt to address the shortcomings of global factorization approaches such as latent semantic analysis and local contextual windowing approaches such as skip gram models. Global statistics about word co-occurrences from the corpus are used to train global GloVe vectors through unsupervised learning. Gloves' goal is to function the log count matrix and identify word embeddings that satisfy this ratio.

GloVe is a new word vector model based on word co-occurrence matrix theory. It can summarize the global and local statistical data of words to create word vectors as an improved performance to the word2vec model. GloVe model can correctly acquire word semantic information and improve the accuracy of word vector semantic information expression.

The Glove, a well-known model based on the global co-occurrence matrix, is suggested; each element $\boldsymbol{X_{ij}}$ in the matrix stands for the frequency with which the words $\boldsymbol{w_i}$ and $\boldsymbol{w_j}$ Co-occur in a specific context window. [141] Suggest the following formula to approximate the relationship between the two words in order to build an approximation of the relationship between a word embedding and a co-occurrence matrix in formula (34) as:

$$\boldsymbol{w_i^T\,\widetilde{w}_j + b_i + \widetilde{b}_j = log(X_{ij})} \qquad (\mathbf{34})$$

Where $\overrightarrow{w_i}$ and $\overrightarrow{w_j}$ are the corresponding embedding of $\boldsymbol{w_i}$ and $\boldsymbol{w_j}$ , $\boldsymbol{b_i}$ and $\boldsymbol{b_j}$ are their offset parameters. After that we can use the following loss function to train word embeddings in formula (35) as:

$$\boldsymbol{J = \sum_{i,j=1}^{V} f(X_{ij})\ (w_i^T\,\widetilde{w}_j + b_i + \widetilde{b}_j - log(X_{ij}))^2} \qquad (\mathbf{35})$$

Where $\boldsymbol{V}$ is the size of the vocabulary.

Where $\boldsymbol{f(x)}$ is a weight function used to control the weight, it does not increase excessively when the frequency is too high in formula (36) as:

$$\boldsymbol{f(x)} = \begin{cases} \boldsymbol{(x/xmax)^{\alpha}}, & \boldsymbol{if\ x < xmax} \\ \boldsymbol{1} & \boldsymbol{if\ x \geq xmax} \end{cases} \qquad \boldsymbol{(36)}$$

Where $\boldsymbol{\alpha}$ and $\boldsymbol{xmax}$ can be provided with empirical value.

Because they can accurately capture the syntactic and semantic relationships between words, these techniques have drawn a lot of interest in text and sentiment analysis. These techniques need to be improved because they have a number of shortcomings despite being very effective. For training and presenting a suitable vector for each word, Word2vec and GloVe both require very large corpora. For instance, Google has re-released pre-trained word vectors with 300 dimensions after using approximately 100 billion words to train Word2vec algorithms. Researchers are forced to use pre-trained word vectors like Word2vec and Glove because some datasets are so small that they may not be the best fit for their data.

#### II.3.4.6. Fast Text:

Methods like word2vec and GloVe ignore word morphology and assign a vector to each word in the vocabulary. When it comes to morphologically rich languages, the latter is a limitation because these languages have larger vocabularies that include more word forms with a lower frequency of occurrence, resulting in lower quality representations. Furthermore, higher vocabulary requirements result in more out-of-vocabulary words.The fast Text model, which is based on skip-grams. However, to fill that gap, we represent each word as a bag of letter n-grams. By representing each word as the sum of its corresponding n-gram embeddings, FastText learns full word and character n-gram embeddings [142].

FastText the technique was quick, allowing for rapid training the model on huge populations. It offers several benefits over previous versions, such as increased speed, scalability, and effectiveness [132].A vector representation for a given word is created by FastText specifically based on the composition of its morphological parts. The most practical way to do this is to think of a word as a collection of character-level n-grams. Working at the character level enables the FastText model to construct vector representations of words that are uncommon or possibly never seen in the training corpus as well as to share morphological information across words. Normally, the model assigns a 0 vector to a word of interest if the letter-level n-grams cannot construct the word. The goal is mostly similar to the SG model described above. Predict word context based on keywords. The FastText model is notable for representing both the focus word and context word vectors as character-level constructs of n-grams [142]. The objective functions in formula (37) as:

$$\sum_{t=1}^{T}[\sum_{c\in C_t} l(s(w_t\ ,w_c)) + \sum_{n\in N_{t,c}} l(-s(w_t\ ,n))] \qquad (37)$$

Where $t$ is the position in the text and $l$ is the logistic loss function defined in formula (38) as:

$$l(x) = log(1 + e^{-x}) \qquad (38)$$

And $s$ is the scoring function that computes the similarity between a word $w$ and a context $c$ in formula (39) as:

$$s(w, c) = \sum_{g\in G_w} z_g^T v_c \qquad (39)$$

Where $G_w$ is the set of character n-grams in the word $w$ and $z_g$ is the vector representation of n-gram $g$ , and $v_c$ is the context vector. The first term in the objective function considers the context words as positive examples, and the second term samples negative examples randomly from the dictionary. Many of the remarks made about hyper-parameter tuning above are directly applicable to this situation because the FastText model and the word2vec SG model are very similar.

#### II.3.4.7. Bidirectional Encoder Representations from Transformer (BERT):

Another model of contextualized word representation based on parallel attention layers rather than sequential recurrence is called BERT. It is based on a multilayer, bidirectional transformer encoder [143]. Google first proposed BERT at the end of 2018. After it became available, both science and business focused entirely on it and conducted additional research. Transformers and word embeddings are used to create BERTs. The word embedding is performed in the model using BERT is just a low-dimensional representation of the words projected onto a high-dimensional vector space. The Transformer, introduced by Google, is a new NN architecture that is more successful at simulating tokens' long-term dependencies in temporal sequences than conventional sequential models like LSTM, RNN, GRU, etc. Additionally, it improves training efficiency by getting rid of the sequential dependencies from earlier tokens. Transformer uses an encoder-to-decoder architecture, where the model adopts an attention system to forward the entire big picture of the entire sequence to the decoder output, as opposed to sequentially feeding in results.

BERT is pre-trained for two unsupervised tasks:

- A **“masked language model " task**: where 15% of the tokens are randomly masked (i.e. Replaced by "[MASK]" tokens) and the model is trained to predict the masked tokens.
- In the " **Next Sentence Prediction" task** :the model is given her pair of sentences and after the first he is trained to recognize when the second sentence comes. The purpose of this second task is to collect more long-term or actionable information.

As a result, BERT is a model that exclusively employs encoders to incorporate all the features listed above. In fact, Google created two distinct BERT model variants, dubbed BERTBase and BERTLarge. The BERTLarge model is a much larger model made up of 24 transformer blacks, 1024 hidden layers, and 16 attention heads [143]. BERTBase is a basic BERT model made up of 12 transformer blacks, 768 hidden layers, and 12 attention heads.

The BERT framework consists of two steps: Pre-training and fine-tuning. The model is trained during pre-training using unlabeled data from various pre-training tasks. The pre-trained parameters are used to initialize the BERT model, and labeled data from downstream tasks are used to fine-tune each parameter. Each downstream task is initialized with the same pre-trained parameters, but with its own fine-tuned model.

A series of tokens, which can be a concatenation of two sentences, is the contribution to BERT. A distinct [CLS] token is attached to the input sequence, and its final hidden state serves as the accumulated sequential representation. Token-level representations are created using the other tokens' final hidden states. BERT's input layer includes an additional segment embedding in addition to the word embedding and positional embedding used in the original Transformer model to distinguish tokens from the pair of sentences.

The importance of lexical, syntactic, and contextual features has been recognized many times before. Thanks to the introduction of powerful contextualized word embeddings and networks like BERT, we are now able to compute better representations of such features.

The main differences between TF-IDF and BERT are: Unlike BERT, TF-IDF considers the semantic meaning and context of words. Moreover, BERT's architecture includes deep neural networks, which can be more expensive than the TF - IDF, which has no such requirement.

**II.3.4.8. XNET:**

XNET is implemented as a route discovery broker overlay network. The network propagates events based on their content and the subscriptions registered by consumers. Overlay network nodes serve as content-based routers. Every data consumer and producer is linked to a node at the network's edge. These nodes are referred to as producer nodes and consumer nodes (CNs) (PNs). For simplicity of presentation, we assume that CN and PN are separated. Therefore, producers and consumers cannot be directly connected to the same router node. Other nodes have no consumers or producers, so they are called routing or internal nodes. Each router must know its neighbors and the best route to each PN. You also need to know the quantity and location of your PNs. Both assumptions are satisfied by each PN maintaining the spanning tree by sending "announce" messages. Each routing node has a set of interfaces or links that connect it to its immediate neighbors. We assume that each neighbor has only one interface (ignoring redundant links connecting two neighbors). Nodes use reliable point-to-point communication for communication. The interface along the path leading to a particular PN is usually called the upstream interface, and the other interface of the particular PN (along the path to the consumer) is called the downstream interface. Consumers are not associated with any of the node's interfaces because they are connected via a link to a CN that is not part of the overlay network. Furthermore, for simplicity, we assume that the CN is an edge router with a single interface connecting it to the overlay network (this property is always satisfied by introducing a virtual CN at the edge of the overlay). Consumers use CNs to subscribe and unsubscribe. Consumers may not cancel subscriptions that they have not previously subscribed to (CN filters out such requests).

XNET was designed to manipulate XML data, serving as the primary exchange language on the Internet. As long as it's a well-formed XML document, producers are free to create semi-structured events and define custom data types. Subscription's language is used to express consumer interest. Subscriptions allow you to specify predicates on the set of events that are valid for a particular consumer [144].

The following summarizes a typical XNET workflow in [145]: The operator must decide the relevant network configurations and modeling granularity based on the target network problem prior to modeling. In order to represent the complex relationship between various network entities, XNET abstracts network systems as relational graphs according to domain knowledge provided by network experts. The network model is built by XNET using programmable GNN blocks, and the properties of the relationships are used to determine the shape of the aggregate functions. The first two steps concern the expressiveness viewpoint (i.e., configuration modeling). By identifying the differences between states in subsequent time steps, XNet models network state transitions using a recurrent version of the GNN. This step involves the granularity (i.e., time-series) aspect.

## II.4. Conclusion:

This chapter provides an overview of the methods used to perform sentiment analysis. We discussed early approaches and recent advances in using machine learning and deep learning models for sentiment analysis. We have categorized these methods and techniques to show that they are widely used in both approaches and give the best results in sentiment analysis.

# Chapter 3

# Arabic Sentiment Analysis (ASA)

# III. Chapter 3 Arabic Sentiment Analysis (ASA)

## III.1. Introduction

Recently, some attention has started to focus on the potential for Arabic opinion-mining research to reveal more about user opinions. The focus of this study is Arabic opinion mining.

In this chapter, the foundational knowledge that is necessary to comprehend the strategies put forth and employed within the purview of this thesis is reviewed and discussed. We discussed the methods that are frequently employed for Arabic SA. There is a brief explanation of the Arabic language. We offer a summary of the studies and research that are related to Arabic in the form of a literature review. We talked about the ML methods that were applied to our proposed SA systems within the parameters of the thesis.

## III.2. Arabic language:

Arabic is a major language spoken in 22 countries and by over 400 million people. It is a loose term for many existing languages. Arabic is their dialect, the sixth most spoken in the world. Spoken by the majority of the general population, it is among the ten most commonly used languages on the internet. Arabic is written from right to left. Also, there are 28 letters (long vowels are its 3 letters and the rest are consonants). Diacritics and inflectional variants are also used as short vowels, with the exception of one letter that functions as a double consonant sign. Numbers are written left-to-right, making it difficult for Arabic-language editors to process words written in both directions in the same context.

Arabic is also the official language of the countries of the Arab world, which has global significance; unlike the dialects discussed in the West, which are primarily Indo-European. Muslims offer their prayers in Arabic, which is why Arabic has a significant place in Islam. Arabic is a very rich language that belongs to a different language family, particularly the Semitic vernaculars. Arabic is fascinating, and anyone with even a basic understanding of the language. The Middle Eastern language of Arabic has complex morphological, syntactic, and semantic variations that vary depending on where you are. Although Arabic lacks capitalization, the diacritics used above or below a letter can still be used to indicate a difference in pronunciation. Arabic has two forms: formal Arabic, also known as Modern Standard Arabic (MSA), and informal Arabic. Local and colloquial terms are used in different parts of the Arabic-speaking world, while formal Arabic is used in books, newspapers, academic circles, and other forms of formal literature.

The Arabic Language is categorized into three types: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect Arabic (DA) illustrated in Figure 8.

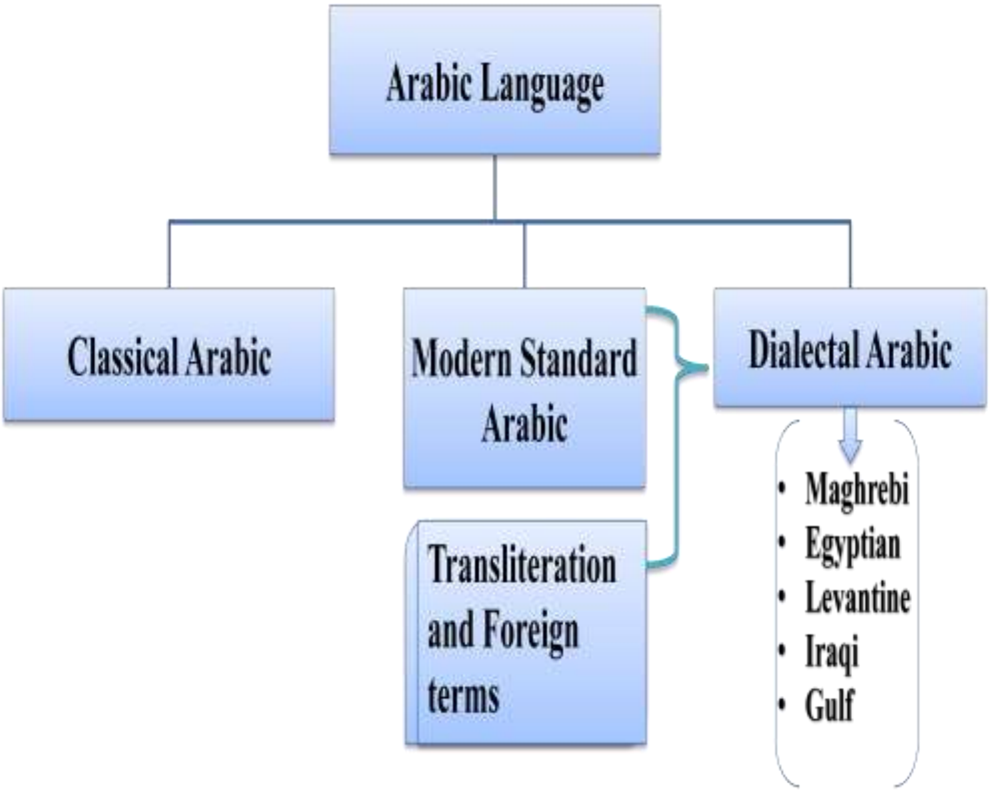


**Figure 8: Arabic language varieties.**

Arabic has many different dialects and is used in informal daily communication, but it is not standardized and not formally taught in schools. Despite the fact that there are numerous dialects, only MSA is a standard form that is officially standardized, regulated, and taught in schools. And consist a vocabulary size greater than 1.5 million words. Classical Arabic used in the Qur'an serves and religious text as the foundation for MSA. The MSA is largely distinct from dialect forms and is not a native tongue of any nation. It differs from the various spoken Arabic dialects, each of which is a regional variant, significantly. Its use in writing, communication in official settings made it necessary. MSA is not a country's native language and is therefore very different from dialect forms. The MSA can be modified to incorporate new words that should be created in light of innovation or science. However, the letter set, spelling, or vocabulary of the composed Arabic script has not changed in at least four millennia. Almost no living dialect can make such a claim. Numerous NLP tools are available for MSA, which has been the subject of in-depth research. Unfortunately, a dialectal writing style that hasn't been extensively studied is used for the majority of web content. As far as we are aware, there are no dependable NLP tools available for it.

Vernacular Arabic, also known as "informal Arabic," is a language spoken by Arabs that vary by country and, more specifically, by region within a country. Most Arabic internet users see the use of Arabic as an element of creating high employment out of online networking, which varies by region. The spoken forms of Arabic, or DAs, are referred to as "the true native language forms" and are typically used in casual daily communication. DAs first appeared in written form and were digitally stored as social media grew in popularity. Arabic on social networks uses informal styles that are a combination of regional dialects, such as Egyptian Arabic and Gulf Arabic, alongside MSA, despite being highly individualistic. As a result, online communications (such as micro blogs) represent a rich source of the Arabic language's variable forms, such as MSA, DAs, or a combination of the two, that can be used to build data sets for computational linguistics.

On the other hand, the dialect of Arabic varies according to location. There are five main subgroups of Arabic dialects: Gulf, Egyptian, Levantine, Iraqi, and Maghrebi. This classification is based on geography, with each dialect group being spoken by close neighboring nations. It's important to note that one dialect group contains numerous dialects that appear to be similar but are actually distinct. For instance, the dialects of Emirati and Saudi Arabia are distinct from one another. The dialect groups and associated regions are listed in Table 8.

**Table 8: Arabic Dialect Groups.**

| **Arabic Dialects group** | **Spoken by** |
|---|---|
| **Gulf** | UAE, Saudi Arabia, Kuwait, Bahrain, Qatar, Oman and Yemen |

| | |
|---|---|
| **Egyptian** | Egypt and Sudan |
| **Levantine** | Levant region (Jordan, Lebanon, Palestine and Syria) |
| **Iraqi** | Iraq |
| **Maghrebi** | North Africa countries; Tunisia, Morocco, Algeria, Libya, and Mauritania |

Given the significance of the Arabic language, there is considerable interest in researching sentiment analysis techniques for Arabic-language social media text data. Due to the large number of Arabic speakers worldwide, and the active Arab Internet users who are constantly posting, sharing, and interacting with social media, this interest has arisen. Consequently, there is an increasing amount of Arabic content online. In fact, Twitter is one of the most widely used social networking sites among Arabs, and Arabic ranks as the fourth most popular language online.

There are numerous linguistic distinctions between the MSA and regional dialects. Some of these distinctions are at the level of short vowels, which are omitted in Arabic text and thus do not appear in writing form. Having said that, many differences do emerge textually:

- The morphology of MSA is more diverse than that of dialects. For example, MSA provides a dual form in addition to the plural and singular forms, whereas most dialects do not. Furthermore, MSA has two plural forms, one masculine and one feminine, whereas dialects frequently do not.
- Dialects do not have grammatical capitalization, but MSA has a complex case system. In the MSA, which is most often expressed with diacritics, but seldom written explicitly, the accusative is expressed with a diacritic plus the suffix (+A), a notable exception (as with objects and adverbs).
- In the dialects, there are no direct translations for some function words in MSA.
- The vocabulary itself has lexical differences, particularly for non-nouns.
- Even when the trilateral root is preserved, there are differences in verb conjugation.

## III.3. Challenges with Arabic Language Sentiment Analysis:

In Arabic, the word for the request is variation. The word that needs to be underlined and placed at the beginning of the sentence is something we can decide on our own. Generally speaking, the syntactic analyzer parses the information tokens from the lexical analyzer and makes an effort to identify the sentence structure using Arabic linguistic usage rules. Because Arabic sentences have a moderately free word request, there are syntactic ambiguities that call for careful examination of all potential grammatical rules as well as the relationships between constituents. The difficulties of the Arabic dialect are discussed in the subsections below, along with their associated computational problems at the orthographic, morphological, and syntactic levels. As they all assist in making the words seem important and well-placed as well as in helping to clarify the sentence, there are covers between these levels in computerizing the method of investigating Arabic sentences. Numerous researchers talked about various difficulties with the Arabic sentiment analysis task. The complexity of the Arabic language is largely to blame for these difficulties, which include:

- An additional characteristic of Arabic is its morphological difficulty. The vocabulary is significantly expanded by the inventive use of roots and morphological specimens, which also contributes to the morphological richness of the language. 10,000 free roots were obtained from 85% of the words collected through tri-requesting roots. Arabic expresses the majority of its syntax and related information at the word level because it is a morphologically rich language. In contrast, English uses far fewer words to convey the same amount of information. The three different forms of a verb in English will result in data sparsity in Arabic, whereas the three terms are likely to appear in the text. A word's Arabic root form can result in. Arabic has high intonations in morphology as a result of its extensive word stress and derivational structure. One interesting example is the **root (كتب )** which is same for many unrelated derive words such as **(books – كتبُ), (office - مكتب),** and **(write – يكتب).** Furthermore, Words in Arabic, a language with templating morphology, contain roots and combine with joins. Even though some Arabic words only consist of one word, they can be translated into full sentences in other languages. For example, the phrase **"Then they will eat it"** can be presented in one Arabic word **" فسيأكلونها".** However, the Arabic language's inflectional structure also presents difficulties. The affixes are added to the root to form the word structure (word = prefix (es) + lemma + suffix (es)). Conjunctions, articles, and prepositions are acceptable as prefixes, and objects or possessive anaphora are frequently used as suffixes. One word can have zero or more affixes thanks to the combination of prefixes and suffixes shown in Figure 9.

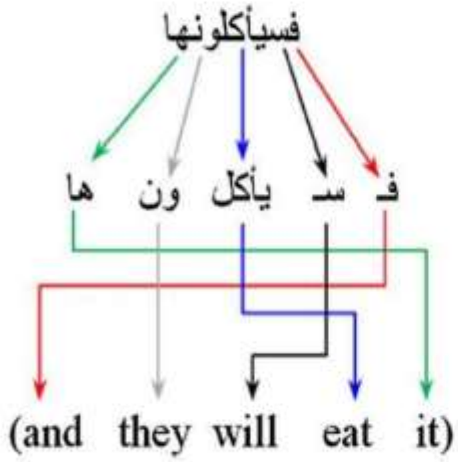


**Figure 9: Arabic word composition.**

- The flexible arrangement of free sentences is another reason why learning Arabic is difficult. In Arabic, the same sentence can be written as a noun sentence (beginning with a subject and ending with a verb) or as a verbal sentence (beginning with a verb and followed by a subject). This is not the case in the English language.
- The Arabic text is different from other languages is that capitalization is not used in Arabic. This makes it difficult for the computer to distinguish between names and other words, especially when two words are written identically. For instance, the word **"hope"** is a common girl's name **"أمل"** but it is also a commonly used sentimental word. The boy's name **"سعيد"**, which is also a popular sentimental word that means **"happy"** is another illustration. This is due to the fact that many Arabs favor using uplifting adjectives as names [146].
- Arabic lacks any distinctive signs, which makes acknowledging a Named Entity more difficult. English shows that a word or a group of words is a name when compared to the specific Orthographic markers, particularly upper case. Since Arabic lacks capital letters, this trademark addresses a major barrier to the crucial task of named entity recognition when compared to other dialects, where capital letters serve as a crucial feature in identifying formal people,

places, or things. A fact that makes it more difficult to identify named entities. Named entity recognition systems are essential for analyzing Arabic texts and distinguishing between entity names and sentiment words. A large percentage of Arabic names are preceded by favorable adjectives. For instance, the first name **"جميلة"** is pronounced similarly to the adjective **"جميلة"** which means **"Beautiful."**

- The variations of the letter shape: the state of a letter inside the orthographic examples of the composed words can change based whether it is linked to a preceding and going to result letter or just linked to a previous letter. For instance, relying on whether they appear earlier, in the middle of, or at the end of a word, the states of the letters **"ف " and "f," i.e.,"ف"/"ـفـ"/"فـ"** change. Arabic spelling includes a collection of orthographic images called phonetic symbols that convey the expected pronunciation of words. This will make clear the meaning and meaning of the word. In the Arabic script, phonetic symbols are used to represent short vowels. These are placed either above or below, the letters to ensure correct pronunciation and make the meaning of the word clearer. In order to distinguish between double consonants and short consonants, Arabic uses signs like the ***fatha, dama, and kasra***. For example the undiacritized word **"شعر "** may mean **(شِعْرٌ poetry), (شَعْرٌ hair) or (شَعَرَ to feel).** The vast majority of MSA texts lack short vowels. This is true because competent speakers can understand a text without the use of diacritical marks. However, diacritical marks are frequently utilized in religious texts, children's books, and books for Arabic learners (the Quran). The majority of texts lack diacritical marks, which causes lexical ambiguity issues for computational systems. All Arabic texts lack punctuation to make it easier for native Arabic speakers to read. By using the context in which the word has appeared and the reader's knowledge, it is simple to determine whether a word is ambiguous. However, this point is one of the main challenges, according to NLP researchers, especially in the tasks of Machine Translation (MT) and Word Sense Disambiguation (WSD).
- Negative words can reverse the meaning of a sentence. So if the sentence was positive before the negative word was added, it becomes negative. In NLP, negative words are considered stop words and removed from sentences as a pre-processing step. Sometimes sentences with negative words are used to

express positive sentiments. For example: *" الفشل لا يعني النهاية بل يعني بداية جديدة "* that means *" **Failure does not mean the end, it means a new beginning.** "*

- Depending on whether the verb is singular, plural, or has a feminine or masculine subject, the verb can be written in various ways. For example: " ***هي تنام" ("she sleeps")*** and ***" هو ينام " ("he sleeps.").***
- Figurative language is one of the most difficult problems that natural language processing has to deal with, especially Arabic language (metaphor and metonymy). By its very nature, the Arabic language makes use of linguistic devices to convey more complex meanings, including metaphors, analogies, ambiguity, irony, sarcasm, euphemisms, hyperbole, context shift, false assertions, oxymoron's, paradoxes, and rhetorical questions. We have not found much research on the subject of Arabic figurative emotion recognition in exaggeration, understatement, rhetorical questions, metaphors, similes and other literary ingenuities.
    - ✓ The use of an overstated tone to convey feelings and opinions is known as hyperbole ***(صيغة المبالغة ).*** It is frequently employed to leave a lasting impression or highlight a point. It has standard weights, which are the most commonly used For example : ***(فَعِلّ : مَرِح :Fun) , (فَعيلّ:كريم:generous) ,(فَعُولّ: طموح : ambitious), ( : مِفْعالّ معطاء : giver) and (فَعّال:صوّامّ : Fasting).***
    - ✓ Euphemism ***(الكناية)*** softens the meaning of the original statement by omitting certain details of an expression or opinion that is supposed to be unpleasant [147].
- Arabs who speak multiple languages, bilingual Arabic speakers, particularly the younger generations, frequently mix up to three languages when speaking or texting. They are bilingual or trilingual, which explains why. Consequently, the sentence might contain some foreign words and be in Arabic. Arabic characters are typically used to represent foreign words. Furthermore, some English words learned by Arabs do not have an Arabic equivalent. For example the following words ***(mention – منشن) ,( Block - بلوك) and (Retweet – ريتويت)*** .( These words are called "loanwords". In [148] discusses other issues in research, such as abbreviations, compound words, misspelled words, abbreviations, dialects (slang), neologisms, concatenation, word extension, and idiomatic expressions. The compound of the words ***(هذا)*** and ***(الوقت )*** into

***(هالوقت)*** is an example of a compound word problem. An example of the elongation is the word ***(النعيم)*** which becomes ***(النعيييييم)*** after elongated.

In addition to the challenges presented by the Arabic language, there are also challenges in using other data. Numerous researchers have discussed these difficulties.

## III.4. Arabic Sentiment Analysis Types:

Sentiment analysis is similar to text classification in that it relies on identifying a document's topic (ex: sport, news, movies, etc.) categorizing subjective phrases into positive, negative, natural, and mixed categories are central to sentiment analysis. Text classification approaches serve as the primary means of sentiment analysis. Phrase, sentence and document levels are just a few of the classification levels at which sentiment analysis is used. To find the polarity, various strategies have been applied at various levels. The literature has employed a variety of methodologies to analyze English sentiment at various levels, but the majority of work on Arabic sentiment analysis has been done at the document and sentence levels, as will be covered in the next section.

### III.4.1. Document Level:

Arabic language sentiment analysis at the document level: The task of document sentiment analysis can be viewed as a classification problem that divides sentiment into two, three, or four categories (positive, negative, natural, and mixed). Since the issue is similar to text classification, all applicable techniques can be applied. These methods rely on identifying sentiment words that denote a favorable or unfavorable opinion. The simplest method in this situation is to represent a word as a bag of words; all other relationships between words are disregarded.

The first attempt to identify sentiment in Arabic text at the document level used supervised learning [68] . The aim of her research was to create effective mood analyzes in different languages. Both Arabic and English were tried. They chose features for both Arabic and English using Support Vector Machines (SVM) and Entropy Weighted Genetic Algorithm (EWGA). A small dataset from two extremist forums was used for the test. There were two forms, one in Arabic and the other in English. In the Arabic web forum, they achieved 91% accuracy, and in the English forum, 90%. They came to the conclusion that accuracy was improved by using both stylistic and syntactic features.

### III.4.2. Sentence Level:

Rather than identifying the sentiment of an entire document, sentence-level analysis distinguishes between subjective and objective sentences and determines the integrity and polarity (positive or negative) of sentiment for each individual subjective sentence. The classification of subjectivity is outside the scope of this proposal as it is not required for aspect extraction. A list of Arabic subjectivity classification works can be found at [149].

Through a series of experiments on sentiment analysis at the sentence level, Abdul-Majeed and his associates produced a corpus of 2855 sentences from the Penn Arabic Tree Bank [150]. Two native Arabic speakers with college degrees annotated those sentences. Each sentence's Objective and Subjective components were noted (Positive, Negative, and Neutral). They used a variety of features in their research, including linguistic features, morphological features specific to Arabic, and genre-specific features. The amount of stemming necessary for such a system was also investigated, and it was discovered that stem setting outperformed other lemmatization techniques. At the sentence level, the system categorizes subjectivity and sentiment Newswire data written in MSA. Following an SVM classifier for sentiment, the system used an SVM classifier for subjectivity. Using the unique, domain, and adjective features, they report accuracy of 95.52%.

### III.4.3. Aspect Level:

At the aspect level, the emotional aspects of the text are identified and separated before the polarities (positive, negative, neutral) are defined. This process includes several subtasks for performing sentiment analysis at a deeper level than the document or sentence level. There are numerous studies that examine opinions at this level in English, including those that extract, categorize, summarize, and extract opinion holders or opinion targets. Many of these tasks are made more difficult by the trustworthy Arabic NLP tools that are widely accessible. There are a few early Arabic works available at this level.

### III.4.4. Conceptual Level:

The semantic analysis of the sentence is the main focus of the conceptual level. It investigates how conceptual sentiment information about emotions is inferred. Table 9 displays the mapping between the proposed levels of processing for Arabic

Language Sentiment Analysis and the current level of traditional sentiment analysis (ALSA) [147].

**Table 9: Arabic Language Sentiment Analysis levels vs. Opinion levels.**

| Arabic Language Sentiment Analysis Levels | Opinion Levels | Explanation/Features |
|---|---|---|
| **Phonetics** | Aspect level | Consonants, Vowels, Syllables |
| **Morphology** | Aspect level | State (indefinite, definite, syntactic), gender (masculine, feminine), number (singular, plural), mood (indicative, subjective), voice (active, passive), person (first, second, third), part of speech (POS) , phonetic symbol, aspect |
| **Syntax** | Sentence level | POS, n-grams of words, lexemes, Bag of Words (BOW), bag of lexemes. Syntactic dependency |
| **Lexicology** | Aspect + sentence + concept | A lexical and learning-based approach to mining concepts from opinions - extracting concept tags from visual content and text metadata - corpus annotation, lemmatization, POS tagging and parsing |
| **Semantics** | Concept level + document level | Semantic features - emotions (surprise, fear, disgust, sadness, happiness, anger) - positive vocabulary, negative vocabulary, neutral vocabulary |

| | | |
|---|---|---|
| **Figurative** | - | Parables, euphemisms, exaggerations, context changes, false claims, oxymorons/paradoxes, rhetorical questions |

## III.5. Arabic sentiment resources :

### III.5.1. Corpora foundations :

A corpus is a collection of texts, speeches, or other examples. It is typically written in a machine-readable format and is regarded as being more or less representative of a language. These days, computer corpora may contain millions of running words that have characteristics that can be examined using tagging methods. Putting labels on words and other formations and classifying them according to those labels is the process of tagging. The corpus is crucial to training the sentiment classification system in sentiment analysis. It contains a staggering amount of words, phrases, sentences, and paragraphs that express emotions. Collection, annotation, and analysis are the three main steps in creating a corpus for opinion mining and sentiment analysis. It is uncommon to compare Arabic sentiment corpora, lexicons, and datasets to those of other natural languages like English. In order to train machine learning classifiers, Arabic corpus annotation for sentiment analysis involves classifying features with the appropriate meta-data. Arabic annotations are applicable on two levels (sentence level and word level). The Arabic annotated corpus needs to demonstrate how effectively it is annotated. Several native speakers can annotate manually, or it can be done automatically or with the help of the public. It is obvious that a high-quality annotated Arabic corpus is needed for sentiment analysis in order to improve classifiers and resolve research problems while taking into account the six levels of Arabic sentiment analysis. The most recent Arabic language annotated corpus for sentiment analysis was reported in a review [151].

The majority of corpora used for sentiment analysis are gathered from websites and social media platforms. These platforms offer perceptions into how people feel about various entities and the parts of those entities, such as people, groups, or things. Web scraping and crawling are popular collection methodologies, as are calls to web APIs like those from Twitter, Facebook, and Google Reader. From the standpoint of data classification and analysis, a scheme definition is necessary for the annotation of the gathered data. This step is difficult for sentiment analysis because there is no established model or theory. The usual approach is to categorize the polarity of the emotions into positive or negative and categorize the emotions into anger, disgust, fear, sadness, surprise, joy and love. Document, paragraph and sentence level annotations can be done manually through crowd sourcing or automatically based on vocabulary and emoticon data.

Subjectivity has a great influence on the annotation work. In order to measure and adjust the inter-annotator disagreement, it is advised to use multiple annotators. To train and test machine learning statistical tools for sentiment classification, annotated corpora are helpful. The effectiveness of the tools is highly influenced by the quantity and quality of the data. By contrast the classification results of the annotated data with the results of human annotation, the reliability of the annotated data are typically assessed. The analysis and use of a corpus may reveal the limitations of the annotated method or data sampling, which can then be improved upon or expanded upon by gathering more suitable data.

The analysis and exploitation of Arabic corpora, particularly dialectical ones, add another layer of complexity to the annotation process. Due to the numerous dialects and foreign terms, manual annotation is relatively expensive and time-consuming. In order to include all potential dialects, the corpus should be annotated by people from various Arabic-speaking nations. An alternative method for annotating Arabic corpora is lexicon-based labeling. However, the lack of Arabic vocabulary for sentiment is still a problem. Annotations based on favorable and unfavorable emoticons would be most helpful in this situation.

### III.5.2. Arabic Sentiment Analysis Designed Corpora:

Since it is used to train and assess classifiers, a corpus is a crucial component of precise sentiment analysis. The development and evolution of ASA have been aided in recent years by the efforts of numerous researchers who have worked to produce more easily accessible Arabic resources, such as corpora. However, it remains difficult to find an Arabic corpus that can be used for subjectivity and sentiment analysis tasks. The available corpora used in ASA are displayed in Table 10. Only the most frequently cited works were chosen, and their resources URLs were provided. We rate each corpus according to its quantity, dialect coverage, fuzziness considerations, and accessibility. Resources are arranged in decreasing size order. Similar to this, the fuzzy entry is examined if the data has been labeled by multiple annotators.

**Table 10: A corpus of Arabic texts for sentiment analysis.**

| Corpus | Size | Dialects | Fuzziness | Public |
|---|---|---|---|---|
| **BRAD [152]** | 500k | ✓ | ∅ | ✓ |
| **HARD [153]** | 370k | ✓ | ∅ | ✓ |
| **LABR [154]** | 63k | ∅ | ∅ | ✓ |
| **Tunisian Corpus [155]** | 17K | ✓ | ∅ | ✓ |
| **AWATIF [150]** | 11K | ∅ | ✓ | ∅ |
| **ASTD [156]** | 10K | ✓ | ✓ | ✓ |
| **AraSenTi-Tweet [157]** | 17K | ✓ | ✓ | ✓ |
| **MASC [158]** | 9K | ✓ | ∅ | ✓ |
| **ArSentD-LEV [159]** | 4K | ✓ | ✓ | ✓ |
| **OCA [160]** | 500 | ∅ | ∅ | ✓ |

One of the earliest publications in Arabic sentiment corpus construction is ***Opinion Corpus for Arabic (OCA).***In [160] made an effort to construct it by manually adding a parallel English version known as EVOCA.500 reviews total, 50 percent negative and 50 percent positive, make up the corpus. They removed irrelevant comments, Arabizi, reviews written in multiple languages, misspelled words, and reviews that were outside the scope of the pre-processing. The corpus, along with uni-grams, bi-grams, and trigrams, is freely available for research. Given the ASA's limited resources and lack of Arabic parsers at the time, attempting to build the OCA was a breakthrough that allowed researchers to create more advanced corpora and prototypes. I've been advised that the OCA's main limitations are its small size, lack of impartiality and objective verification, and coveted domain restrictions.

The most recent corpus presented by [153] is the ***Hotel Arabic Reviews Dataset (HARD).*** More than 370,000 reviews in the MSA with some dialectical content make up the rich dataset known as HARD. 46,968 positive and 47,084 negative reviews make up the 94,052 reviews in the balanced subset of HARD. Based on its rating, each review in the dataset is annotated as positive, negative, or neutral. Both the balanced dataset and the unbalanced, complete set of HARD are freely available to researchers. Despite the size of HARD, there is little dialectical content because the vast majority of dialectal reviews are in Gulf dialects.

A multi-genre MSA corpus labeled for subjectivity and sentiment analysis called ***AWATIF*** was published in 2012 [150]. 10,729 sentences total, including 2,855 sentences from Newswire articles, 5,342 sentences from Wikipedia talk pages, and 2,532 threaded conversations from online forums, make up the corpus. The authors followed simple guidelines for the annotation in addition to linguistically motivated and genre-specific guidelines. They looked at how linguistic expertise affected the caliber of annotations. They also took into account the task's ambiguity in terms of subjectivity and sentiment. Furthermore, they therefore tried to give the annotators the proper preparation for the job. Their dataset, however, only focuses on MSA, which is uncommon when posting reviews on the majority of websites and social media platforms. Furthermore, none of their resources are open to the public.

The ASA big data scale has recently been introduced. A ***Large-scale Arabic Book Review (LABR)*** with 63k records, only for a particular domain, was presented as the first corpus in [154]. Each entry represents a reader-written book review. Books are rated by readers on a scale of 1 to 5 in the LABR. The ratings of 1 and 2 were thought to be negative, 4 and 5 to be positive, and 3 to be neutral by the authors. Their corpus is not diverse enough in terms of Arabizi and Arabic dialects. LABR contains numerous reviews, but it only addresses MSA.

The largest ***Book Reviews in Arabic Dataset (BRAD)*** for sentiment analysis and other applications are presented in [152] as a result of LABR. There are nearly 510K book reviews on BRAD. However, there are only 150K reviews in the balanced dataset. The majority of the dialectal content in the corpus is from the Egyptian dialect.

In this study, we create a new ***classical Arabic sentiment analysis dataset CASAD***. The books of the most famous Arabic authors have been read and each sentence has been manually extracted and marked. Data on human development, history, fairies, novels, medicine, etc. were randomly collected from different books and manually downloaded from the internet. These books were compiled from the 63,257 book reviews found in the Large-Scale Arabic Book Review dataset [154] [161], which were obtained from www.goodreads.com. In total, 9709 paragraphs were created from the data that was gathered. Three Ph.D., holds Arabic-speaking human experts from various faculties were asked to comment on the text. To assess the expertise of the experts, statistical methods (validity and reliability) were applied using SPSS. A Cronbach's alpha reliability coefficient value is recommended.

The ***Arabic Sentiment Tweets Dataset (ASTD)*** was also presented by the authors of LABR [156] ; 10K Arabic tweets were manually categorized into objective, subjectively positive, subjectively negative, and subjectively mixed categories. Both the dataset and the experiments used are available to the public. These sources, however, are only available in Egyptian.

A brand-new Arabic language reference corpus called the ***Multi-domain Arabic Sentiment Corpus (MASC)*** was provided by the authors in [158]. The scientific community could access it freely thanks to their efforts. 8,860 reviews from various domains and Arabic dialects are included in the corpus. The reviews were manually gathered from a variety of websites, including Twitter, Facebook, Google Play, the Jeeran and Qaym websites. The corpus includes evaluations from 15 different fields. The size of the corpus, however, is comparatively small to cover a variety of dialects in 15 different domains, given the diversity of dialectal Arabic.

Some corpora have been created for particular dialects, as opposed to earlier corpora of multiple dialects. The ***Tunisian Sentiment Analysis Corpus (TASC),*** Literature focuses mainly on Tunisian dialects was presented by the authors in (156). He collected 17,000 comments on the official Facebook pages of Tunisian radio and TV channels such as Mosaique FM, JawhraFM, Shemes FM, HiwarElttounsi TV and Nessma TV; manually adding positive and negative polarities to Corpus Notes. Unlike most previous works, TASC is limited to Tunisian dialects and does not exclude Arabic texts. These resources are freely accessible for research purposes within the same region [157] the author also presented his AraSenTiTweet, a collection of Saudi Arabian tweets for the ASA. The corpus contains a total of 17,573 tweets, each labeled as positive, negative, mixed, neutral, or indefinite. Three annotators manually annotated the corpus, and to ensure the accuracy of the annotations, kappa statistics were computed.

***ARSENTD-LEV***, an Arabic sentiment dataset composed of Levantine tweets to cover spoken dialects in Jordan, Lebanon, Palestine, and Syria, was presented by the authors in [159]. The corpus contains 4K tweets. The annotation was done by crowd sourcing on the Crowd flower platform. First, commenter's were asked to choose the overall sentiment of the tweet (very negative, negative, neutral, positive, and very positive). The commenter's, then had to select a sentiment indicator to determine whether sentiment was explicitly or implicitly expressed in the tweet and to indicate the topic discussed.

Although there are many sources for ASA, their use and applicability are still restricted by a number of problems. First off, the vast majority of recent works in ASA have not yet made their resources available. Second, they have similar flaws in terms of the size of the dataset, the caliber of the reviews, and the fuzziness factor. Third, excluding dialects and Arabizi, the majority of corpora are only domain-dependent for Newswire and/or Arabic MSA. They are one of the few sources that only mention one or two dialects. Additionally, they do not include the Arabizi text, except TASC [155]. Fourthly, only a small amount of resources took the fuzziness issue into account by increasing the number of annotators or adding extra features. The sentiment classification accuracy suffers as a result of these flaws. To solve these issues, we must create more extensive resources that can accommodate the rich morphology and wide variety of dialects. More importantly, these resources ought to concentrate on their effectiveness and capability in combating spam and fuzziness. There may be a lot of fuzziness due to the differences in words, syntax, and phonetics between MSA and dialects as well as between dialects themselves. Because people frequently use their regional dialects and are very informal, future work should concentrate on developing dialect-based corpora rather than ones based on MSA. This is especially true for the sentiment analysis task.

## III.6.Arabic Sentiment Analysis Approaches:

Research on the Arabic sentiment analysis is ongoing. There are numerous studies and research projects that aim to advance Arabic sentiment analysis, either by introducing fresh methods and tools or by enhancing already existing ones. In other words, some research focuses on developing and refining the model, while other research focuses on data preparation. Sentiment analysis is typically performed using supervised and unsupervised methodologies. To perform Arabic sentiment analysis, algorithms are used in both corpus-based (supervised) and lexicon-based (unsupervised) approaches. The combination of these two can be referred to as the hybrid approach shown in Figure 10.

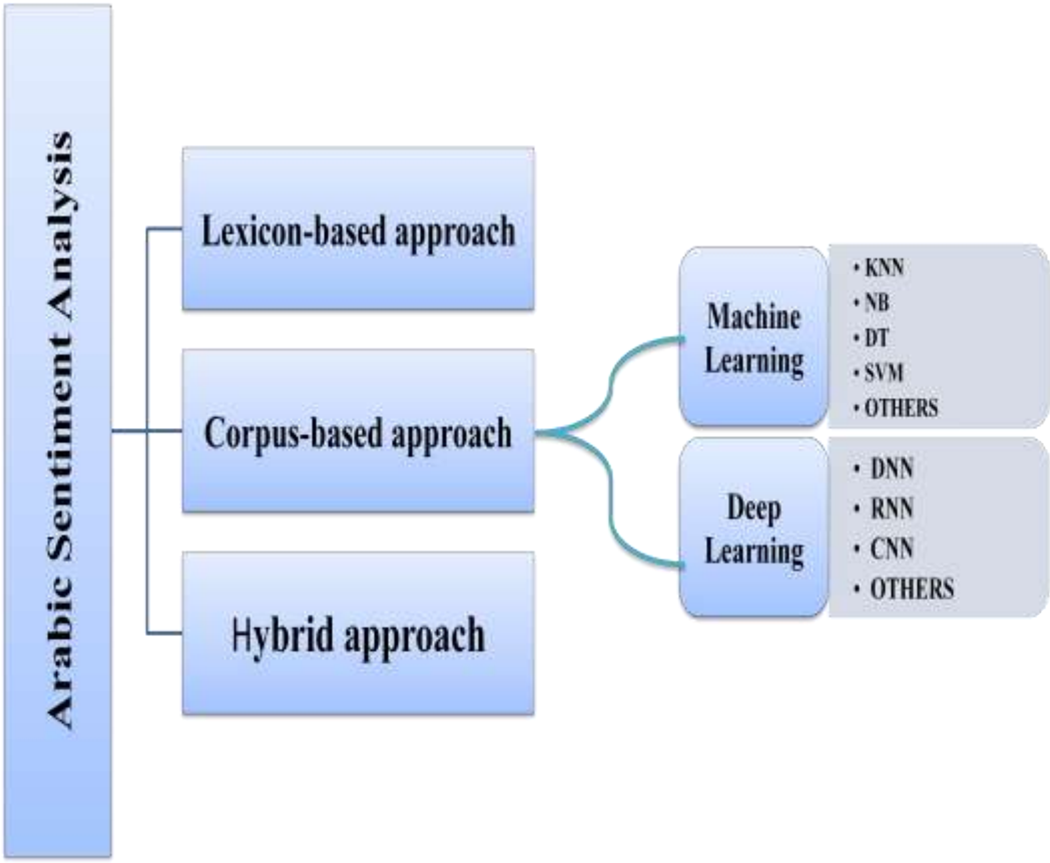


**Figure 10: Arabic sentiment analysis technique.**

### III.6.1. Corpus-based Approach:

Although machine learning is a supervised method that includes many calculations, the key challenge is to identify the right set of skills. The supervised approach, also known as the corpus-based approach, employs various classifiers. Different machine learning classifiers includes Decision Tree, K-Nearest Neighbor, Support Vector Machine, and others. Numerous algorithms are used in machine learning, but the main difficulty is identifying the right feature set. Two approaches the model approach and the relationship-based approach, or a mixture of such methods frequently used in supervised learning techniques. A feature vector made up of one or more feature is used to represent a text that will be subjected to supervised text analysis. Machine learning algorithms and labeled corpora are key components of the corpus-based approach. The general methodology consists of raw text preprocessing, feature removal, training and testing of classification methods in Figure 10 shows the whole process.

The dataset is first gathered and annotated before being split into training and testing datasets for the machine learning method. There are two widely used methods for the annotation process, where each instance in the dataset should be categorized as either positive, negative, or neutral. The first is the manual annotation approach, in which each instance in the dataset is manually classified by two or more annotators. The second method is called crowd sourcing, and it entails classifying the instances in the dataset using a web application created by the authors and a large group of Internet users. The dataset is cleaned in the second stage of developing the sentiment analysis module, where some instances will be eliminated from the dataset, such as duplicate instances and instances with long texts.

### *A. Pre-processing:*

Text preprocessing is a crucial step for any Arabic NLP application due to the Arabic language's rich morphology. It entails eliminating text that is not informative. However, the system might overlook crucial words, if these non-informative parts are defined incorrectly. Tokenization, stop word removal, punctuation removed, upper/lower case conversion, word stemming, spelling check, letter replacement (normalization), and dialect replacement are all used in ASA pre-processing .The pre-processing operations will be applied to each instance in the dataset after the dataset has been cleaned. There have been numerous pre-processing methods employed in these steps; the following points describe these methods and illustrate in Figure 11.

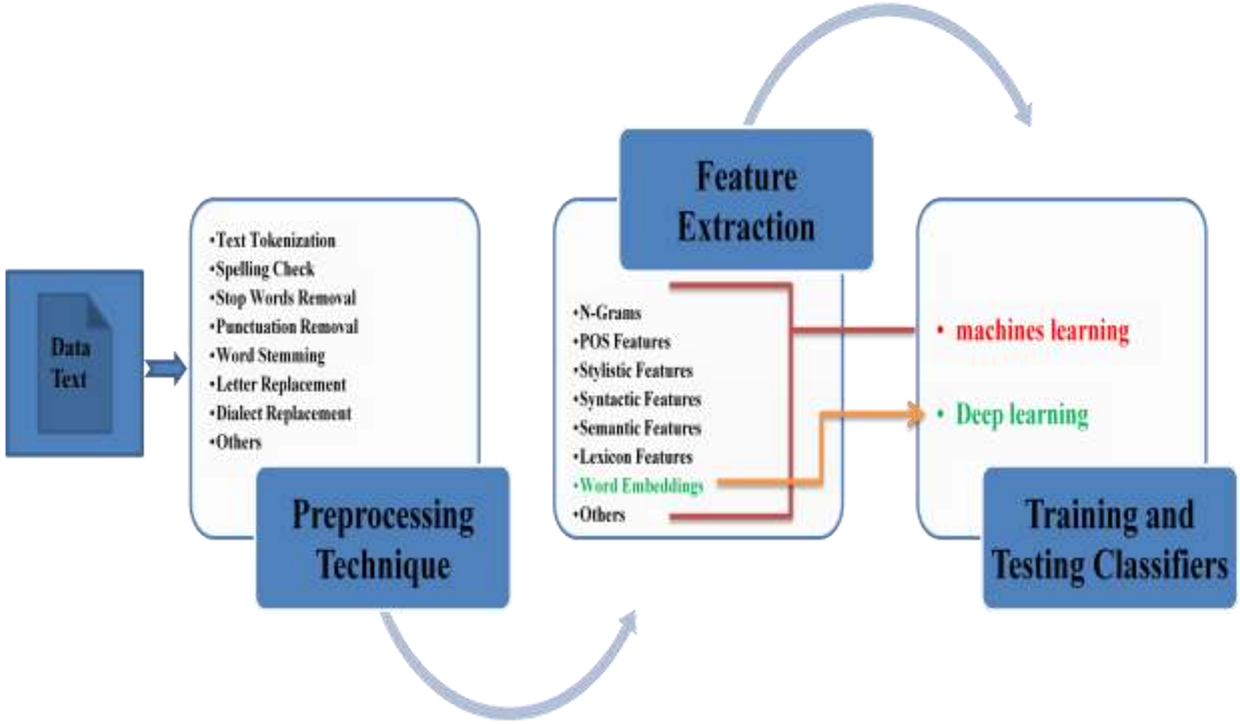


**Figure 11: Machine learning-based methods for ASA.**

- ***Text Tokenization:*** The process of breaking down sentences into words, phrases, symbols, and symbols. Is the process of dividing a sentence into words, phrases, symbols, and characters. For example: *أهلكت -الأمطار- و- الرياح - كلللل -المحاصيل.*
- ***Spelling checks:*** These spelling mistakes can present challenges when attempting to analyze data coming from social media sources, such as extra letters. For example:*أهلكت -الأمطار- و- الرياح - كل –المحاصيل* ***and (محبوووووووب - محبوب).***
- ***Stop words removal:*** Stop words include expressions like ***(في, كان)*** ***'it was'*** and ***'in'*** that are unrelated to the sentiment. Stop words can be eliminated to decrease index space and increase efficiency and response time. Here, we use the terms remove ***(و ),*** which stands for ***"and,"*** and ***(كل ),*** which stands for "***all."*** For example :*أهلكت - الأمطار- الرياح -المحاصيل*
- ***Punctuation removal:*** The majority of punctuation marks, such as full stops and commas, are ineffective at detecting polarity. For example: *أهلكت الأمطار الرياح المحاصيل.*

- ***Word stemming:*** is the process of reducing each word in its most basic form (the root). The word is reduced to its three- or four-letter root by Arabic stemming .For example:***هلك مطر ريح محصول*** .
- ***Letter replacement:*** Some letters in Arabic have various forms. Therefore, some earlier researchers have substituted the default form for each of these letters in place of the various forms. For example: ***(إ, آ, ى = ا), (ة=ه)*** .

***B. Feature extraction:***

The annotated data must be transformed into a feature vector in order to train machine learning classifiers. A combination of particular features will identify a particular class. To create the single feature, a piece of text is transformed into a feature. In the field of sentiment analysis, it is an essential task. The effectiveness of sentiment classification as a whole depends on choosing the appropriate features. The most frequently used features in sentiment analysis studies are term frequency, term uni-grams, parts of speech (POS), and negation.

- ***N-grams:*** A collection of n items extracted from the text. Elements are usually words that make up a sentence. Uni-grams, bi-grams and tri-grams. Each contains one word, two consecutive words, and three consecutive words, respectively, are the most widely used n-grams. N-grams for ASA have been studied extensively.
- ***Part-Of-Speech (POS) tagging:*** The process of tagging words in text based on the nature of the words in the text and their relationship to other words in the text. In English POS tagging, each word is matched to a POS in one of eight grammatical categories. Verbs, nouns, pronouns, adverbs, adjectives, prepositions, conjunctions, and interjections add weight to these words. However, most Arabic POS taggers are dedicated to MSA and some preliminary work has been done for Egyptian dialects. Because of this, many works avoid including its POS tags as features in their classification models. They say its current POS tagger is not suitable for social media texts rich in dialect content.
- ***Stylistic features:*** such as positive and negative emoticons, length distributions, measures of lexical richness, and frequencies of special characters are intended to confirm the presence of specific sentiment

indicators in the text. Increase. Exclamation marks and long question marks are other style elements.

- ***Syntactic features:*** A phrase pattern that indicates emotion, such as a noun followed by a negative adjective. As syntactic features, some researchers have used n-grams of words and POS tags, lemmas, bags of words (BOW), and bags of lemmas. They stated that syntactic features are determined by sentence construction and word assemblage, as well as the use of transitive and intransitive verbs. Transitive verbs are those that are discovered to be associated with object pronouns. For example: ***(سأطير من الفرح: I will fly with joy) means I will be so happy.***
- ***Semantic features:*** are represented by the semantic orientation of the surrounding text. Given the polarity of the sentiment, semantic analysis determines how a group of entities is correlated using various entity concepts.
- ***Lexicon features***: derived from the sentiment lexical include polarity averages and sums. While English has many general lexical items, the Arabic language has a few lexical items that are generated automatically from social media and translated from English lexical items.

The goal of sentiment classification is typically to categorize the text into positive and negative feelings. Due to the handcrafted features, the semantic and syntactic word distributions are well known to be significant obstacles. Word embeddings, on the other hand, get around this issue when deep learning models are used for sentiment classification because they automatically extract features. Word2vec and GloVe are two frequently employed techniques for word semantic distribution that can be used for word embedding for ASA; their explanations are provided in Chapter 2. Only a few recent studies on ASA have made an effort to assess word embeddings in Arabic texts. In this study, we looked into the Arabic sentiment analysis using word2vec.

### *C. Machine learning-based Sentiment classification*

Machine learning algorithms have been trained to conduct ASA. Prediction accuracy is typically used to gauge a person's performance. Although many machine learning classifiers have been used to conduct ASA, only three classifiers SVM, k-nearest neighbor (KNN), and NB consistently demonstrated superior performances. Several pre-processing methods and features were used by some researchers to study these three classifiers. In comparison to the case where there was no pre-processing step, they discovered that the pre-processing step improved the classifier accuracy of Arabic sentiment. They also used word and character n-grams to analyze the performance of these three classifiers. N-grams in levels, words, and characters improved performance, according to the results. They recommended that feature exploration using various classifiers be a requirement for ASA.Many other works targeted only the MSA, Due to its widespread on social media, some dialectical Arabic works have been featured. And they suggested some works for both MSA and Egyptian dialects and colloquial Arabic.

***D. Deep learning-based Sentiment classification***

Machine learning includes deep learning. It is known for learning embedded and abstract representations from raw data with little human intervention. Deep learning has recently been explored in sentiment analysis, with impressive results. Deep learning models efficiently and effectively process large data sets, saving time because no human intervention or feature engineering is required. This model has been used to learn semantic representations of English texts. These representations resulted in a precise sentiment analysis. CNN and long short-term memory (LSTM) networks are two of the most well-known deep learning models.

Sentiment analysis of Arabic content using deep learning research and especially aspect-based sentiment analysis there are few examples in the literature where deep learning research has been applied to sentiment analysis of Arabic content. A CNN model to categorize the polarity of the sentiment in Arabic reviews [162]. Utilizing word embedding features from a corpus of 3.4 billion words, the model was trained. However, there was no research on aspect-based sentiment analysis and the model's main focus was on sentence level sentiment analysis.

### III.6.2. Lexicon-based Approach:

Aspect extraction methods are included in unsupervised methods. According to the highest count, positive and negative lexicons are described. In a document, the word with the highest count determines whether it is a positive or negative word. NLP tools are used in one area of research to help with the extraction. Unsupervised methodology, in contrast to administered methodology, includes illustrations of positive and negative dictionaries that take into account the highest tally. When the data are unlabeled, the lexicon-based method is typically employed. The information is labeled, and its polarity is predicted using a sentiment lexicon. An evaluation of a document's sentiment score (a review) based on the sentiment scores of words or phrases in the lexicon is possible when using a sentiment lexicon. Some researchers claim that when compared to machine learning methods, the lexicon-based method is typically regarded as a weak method. Thus, the lexicon-based method by itself has only been applied in a few works for sentiment analysis in Arabic. In contrast, the majority of studies have combined machine learning with a lexicon-based approach by developing lexicon features. A sentiment lexicon contains words that express emotions and their associated polarity values. There are two methods for creating a sentiment lexicon: dictionary-based and corpus-based.

- **The dictionary-based approach**: works as follows: It starts with a starting set of sentiment words with known positive and negative biases. Use available thesaurus resources and corpora such as WordNet to find synonyms and antonyms for each word in the list. Following the addition of the new word to the seed list, the subsequent iteration begins. When no new words can be discovered, the process is finished. This approach's main drawback is that it fails to identify opinion words with a domain orientation. For instance, saying "The phone speaker is quiet" conveys a negative opinion, whereas saying "The car is quiet" conveys a positive opinion.
- **A corpus-based approach**: can find opinion terms specific to a particular domain and context. This method uses statistical or syntactic patterns and a starting list of opening words of known polarity to search for new opening words of known polarity from a large corpus.
    - ✓ ***For a statistical pattern***: The occurrence frequency of the new sentiment word in a large annotated corpus is used to find it, so if the word appears more frequently in positive documents than in negative documents, it will be added to the word list as a positive word, and if it appears more frequently in negative documents, it will be added as a negative word. In other words, if the word appears more frequently in positive documents, it will be added as a positive word, and if it appears more frequently in negative documents, it will be added as a negative word.

- ✓ ***In the syntactic pattern:*** In the corpus, words with similar opinions are found together. This pattern suggests that if two words are frequently found together in a document, they probably share the same polarity. The words should have the same or opposite polarity value as the known words based on the words connected between the known words. For example, when the word occurs frequently together with other words of known polarity (AND), The word ***(واسعة ": spacious ")*** for instance, in this sentence: ***("The car is comfortable and spacious:السيارة مريحة وواسعة" )*** will have the same polarity value of the word ***(مريحة :comfortable).***

There are two methods described in the literature for creating and growing the Arabic lexicon: manual and automatic. By translating SentiWordNet or SentiStrength content, looking up synonyms for each translated word, and adding the word and its synonyms to the lexicon, the lexicon is manually constructed. In the automatic method, the lexicon is created by beginning with a manually compiled and annotated lexicon (base lexicon) and then expanding it by including synonyms and antonyms. Although the automatic method of lexicon creation takes less time and effort, the manual method is more accurate. The lexicon-based approach is domain independent, and the lexicon is built for all domains, in contrast to the ML approach, which does not work well with data that differs from the training data.

### III.6.3. Hybrid Approach:

The hybrid method combines lexicon-based and machine learning techniques. This strategy has reportedly been found to be superior to machine learning and lexicon-based strategies. The machine learning classifier can be combined with other features, such as sentence-level and linguistic features, as well as some extracted features from the lexicon. Another way to combine the two methods is to eliminate every word that does not belong in the dictionary from all usages, leaving only sentimental words. The sentiment is unaffected by the words that were removed, but keeping them will confuse the classifier and reduce accuracy. Therefore, it is anticipated that the accuracy will increase by eliminating these words. Only a few works for ASA make use of the hybrid methodology. To label text polarity for use in the sentiment classifier, the majority of them used lexicon-based techniques. Also, we note that deep learning models still not explored for a hybrid approach. This strategy is more prevalent in the pertinent literature and is typically acknowledged to perform better than both the lexicon-based method and the machine learning method separately. Typically, the lexicon scores are used as features to feed into the classifier. The research examined dialectical or informal Arabic (DA), modern standard Arabic (MSA), and a combination of MSA and DA lexicons at the word and sentence levels. In terms of machine learning classifiers, naive Bayes and support vector machines were the most popular techniques. Yet other strategies, including entropy and K-nearest neighbor, were also applied.

## III.7. Arabic Sentiment Analysis of Applications:

People frequently want to know what other people think when, among other things, trying to decide what to buy, feeling out the consensus on a subject, or spotting trends. In order to determine policies and campaign strategies, governments and politicians frequently have an interest in mining public opinion. Opinion mining, which aims to automatically extract people's opinions, has consequently become extremely important in business, social media, and politics [163].Sentiment analysis applications have progressed beyond being separate applications that evaluate sentences for subjectivity to being vital entities in critical industries such as politics, healthcare, marketing, finance, services, and education. Sentiment analysis-based applications that target each of the aforementioned industries are constantly emerging. Only a few, however, rely on Arabic sentiment analysis. Following is a summary of the most recent work on Arabic sentiment analysis in each sector.

**Politics:** Sharing your views on social media and blogging platforms can be useful for political purposes, such as alerting political leaders to potential issues and threats, assessing public opinion about specific policies, or compiling popularity indexes that can be used in elections. You will receive valuable information that you can use for your own purposes. And you can know the emotional state of the masses (anger, disgust, happiness). In the creation of such systems, the opinion holder's identification is helpful. In order to track a political figure's political standing, the authors most recently developed a system that, given a political figure, records the correlating opinions introduced on the web and presents a summarized report of that figure [164]. Other pertinent applications included those that concentrated on the impact of sentiment during the 2012 Egyptian presidential elections, proposed an automated tool to identify the political leaning of an Arabic article or comment, and examined user status updates on Facebook posts made during the Tunisian "Arab Spring" period to gain insight into user behavior.

**Healthcare:** Many people share their health-related information and experiences on popular blogs and social media platforms. When people visit healthcare facilities, they discuss their health problems, symptoms, diagnoses, medications they have been given, and experiences. It is frequently critical for patients to learn about the experiences of other patients before making decisions about which healthcare facility to visit or which medication to take; using data from Arabic Twitter to analyze attitudes toward health services [165].

**Marketing:** Sentiment analysis has taken marketing to a new level as social media and the web are increasingly used as platforms for customer interaction. Companies have recognized the value of sentiment analysis for product branding and have invested heavily in recommendation systems and social/sentiment analysis tools. Opinion mining can improve the quality of recommendations made by recommender systems that rely solely on a user-item ranking matrix. In [166] examines the reviews and comments gathered from the Arabic social networking site Yahoo!-Maktoob. They gathered information on a variety of factors that are crucial for marketers, including the length of the reviews, the number of likes and dislikes, the polarity distribution, and the languages used.

In [167] created a social data analytics (SDA) tool that uses the IBM Big Insights platform as its foundation. Through the use of SDA, they are able to recognize user traits like gender, location, name, and hobbies; Build comprehensive user profiles across news and sources. We then associate profiles with expressions of sentiment, buzz, intent, and ownership regarding brands, products, and businesses. It makes it possible for data analysts with little experience in sentiment analysis and information extraction to quickly obtain conclusions from social data. By using extracted Arabic tweets on three different topics: Egyptian news agency, Egyptian government, Saudi employment, and the authors supported Arabic language.

Many companies that provide Arabic sentiment analysis tools, such as, Repustate and LexisNexis, were formed with an interest in providing sentiment analysis for marketing purposes. Some have used sentiment analysis of Arabic tweets to determine how interested people are in a particular brand or topic.

**Finance:** Given the insights it provides on the subject under analysis, sentiment analysis is used as a significant factor in financial decisions. Consider the perceptions provided by sentiment analysis regarding changes in stock price [168]. Sentiment analysis is also used to gauge the level of optimism or pessimism among an individual investor or the general investing public. Stock markets often post positive growth this month as investor sentiment and sentiment is strong, according to research done by the authors in [169] on the impact of the Islamic holy month of Ramadan on Islamic Middle Eastern markets. Using Mubasher products, a top Gulf stock analysis software provider, we also proposed a trading strategy in [42] that is based on sentiment analysis of tweets. Saudi Twitter posts were used to forecast the Saudi stock market. In order to help foreign investors understand the Saudi investor sentiment, the authors looked into the relationship between social media sentiment and the Saudi market index. Additionally, they suggested and advocated the use of a sentiment analysis model based on Arabic tweets to forecast stock market prices for the Al-Marai Dairy Company. Looked into how the sentiment in Arabic tweets affected stock market indicators in Saudi Arabia. They manually analyzed 114K tweets for sentiment before comparing various correlation metrics to Saudi stock price movement.

**Sector of Services:** In [170] a sentiment analysis application for assessing customer service representative (CSR) productivity in real estate call centers is presented. This application uses the Natural Language Toolkit (NLTK). The decision-making process used to create an Arabic system for gauging and measuring productivity is thoroughly covered in the study. A transcription method, feature extraction, training procedure, and analysis are all included in the system; for the purpose of specifically analyzing customer reviews of Sudanese telecommunication products, a prototype sentiment analysis system. Arabic sentiment analysis was used in [171] to approximation customer satisfaction with Saudi Arabian telecommunications providers. Similar to this [172] evaluated customer satisfaction with Jordanian telecom companies using Arabic opinion mining.

**Education:** By examining their tweets, Arabic sentiment analysis has been used to gauge students' satisfaction with and experiences at the university [173]. Some research to understand how students felt about colleges, Arabic opinion mining in order to improve course evaluation moving forward, and presented a system to monitor changes in Arab students' opinions about their courses. Numerous studies are also being done right now to assess the experience of students who sign up for or take online MOOCs.

## III.8. Limitations of Arabic Sentiment Analysis:

The classification process known as "opinion mining" aims to ascertain whether a particular text or document was written to express a favorable or unfavorable opinion about a particular object (e.g., a topic, a product, or a person). Opinion mining generally aims to determine a writer's viewpoint on a specific subject or the overall tone of a document. Among the tasks are sentiment transfer, humor detection, emotion classification, polarity classification, subjectivity detection, and review summarization. Both topic models and conditional random fields are domain-specific and may or may not perform equally well in other domains. The system's performance is constrained by the dearth of trustworthy NLP tools for the Arabic dialect and the difficulty in utilizing Arabic social media text. Although there has been some work on other languages, the majority of sentiment analysis systems are currently created in the English language. The difficulty of learning Arabic is the cause of this restriction on Arabic Sentiment Analysis research. Due to the use of machine translation between the Arabic dialect and English, we show a loss of precision. The Arabic sentiment analysis presents a number of unresolved problems and difficulties, including:

- ❖ One of the Semitic languages, which are a member of the Afro-Asiatic language family, is Arabic. It is written in a cursive style from right to left and is an ancient language that is still spoken today in the Middle East and North Africa. There are no upper- and lower-case consonants like in English, and it has 28 consonants.
- ❖ Compared to other languages like English, Arabic has a very complex morphology.
- ❖ Because Arabic is a highly inflectional and derivational language, monophonic analysis is a very complex and difficult task.
- ❖ Arabic opinion is highly contextual domain-dependent, with words with different polarity categories appearing in different contexts.
- ❖ Where dialectal Arabic resources are scarce, Arabic Internet users prefer dialectal Arabic to MSA. The percentage of spelling errors in these Arabic opinions is high, which adds to the difficulty.

- ❖ There is increased syntactic, semantic, and figurative ambiguity in Arabic due to the language's complexity in terms of spelling, vocabulary, phonetics, and morphology.
- ❖ How to translate figurative language without losing its effective essence.
- ❖ Arabic sentiment analysis has a small community of practitioners.
- ❖ Real-time sentiment analysis; spam detection; morphological errors; poor spelling; unstructured data; and implicit meanings.
- ❖ The ambiguity of the Arabic language has many words with multiple meanings, making sentiment analysis difficult.
- ❖ Figurative expressions that convey emotion.
- ❖ Arabic figurative expressions that convey irony and sarcasm.
- ❖ The identification of similes, metaphors, metonymy, hyperbole, and euphemisms is a difficult task for humans, and it is even more difficult for machines.

Sentiment analysis is a difficult process in and of itself, but it becomes even more difficult when dealing with social media text, which is unstructured, full of spelling errors, and has many peculiarities and conventions. Furthermore, social media text is typically short and abstract, and it is frequently related to another text, such as a response to or elaboration of someone else's post. Unfortunately, conducting sentiment analysis on Arabic social media text makes the problem even more difficult. This is primarily due to the limitations of existing natural language processing tools and resources for Arabic, which were designed specifically for MSA [174]. The dialectal terms and idioms used in social media have been shown to be of a highly dynamic and evolving nature, which presents another challenge for sentiment analysis in Arabic. Subjective creative expressions are frequently created in an instant by social network users, quickly spread, and widely used by other users, transforming them into powerful subjective clauses. The widespread use of transliterated English to convey emotion is another issue. Larger amounts of data are needed for sentiment analysis that is domain-dependent. However, the lack of resources and corpora is one of the biggest problems with the Arabic sentiment analysis. This might be because there aren't many websites where users can post reviews in Arabic. The number of Arabic reviews and datasets is much smaller and more constrained than the total number of Arabic users online. Due to the abundance of corpora in the English language, it is more difficult to compare the sentiment analysis results between Arabic and English. To work toward creating more precise sentiment analysis models and analyses, researchers would greatly benefit from sharing lexicons and knowledge.

## III.9. Conclusion

This chapter reviews concepts and previous work related to this dissertation, has described the basic information about the Arabic language as well as its features that make this language challenges in the Natural Language Processing field. Has defined the Arabic sentiment analysis in general, and explained the process of how to analyze the sentiment in a text. The process includes the features that we can get from the text and the approaches that might be used to make the classification process; and defined some previous works in Arabic sentiment analysis. In addition, it explained some of the issues and gaps that are still open in Arabic sentiment analysis. In this thesis, we do not try to solve challenges related to Arabic syntax. We mainly concentrate on providing solutions related to Arabic sentiment analysis.

# Chapter 4
# Evaluation and Result

# IV. Chapter 4 Evaluation and Result

## IV.1. Introduction:

This chapter deals with the evaluation of the approach framework proposed in this study and the published article. This chapter represents the implementation and demonstration phase of Arabic sentiment analysis. In this chapter presents the primary results analyzed assessing them with respect to other existing NLP tools. It provides a detailed analysis of the results and compares them with the ML result classifier.

## IV.2. Dataset:

Although Arabic is one of the top 10 languages used on the internet, there aren't many websites that offer reviews and feedback in Arabic, so it isn't as rich in content as English [175]. We outline the various steps of the architectural system shown in Figure 12 .Initialize Classic Arabic Sentiment Analysis Dataset (CASAD) first by gathering data from Arabic books. The judges' score is then calculated using Cronbach's alpha after they have evaluated the dataset. Next, various techniques are used to create feature extraction vectors (TFIDF, word embeddings). In the classification phase, the CASAD feature extraction vector is lastly trained using a variety of ML algorithms and cross validated.

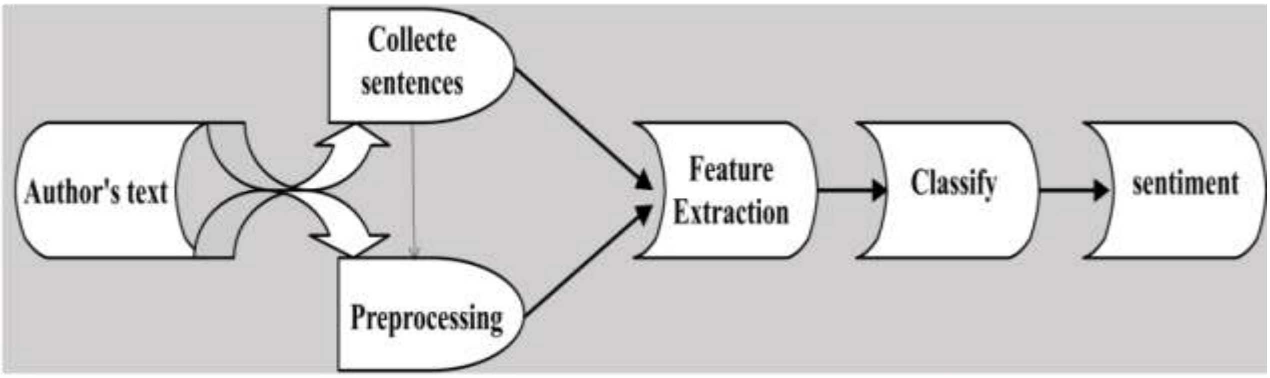


**Figure 12: Arabic Sentiment Analysis System Architecture.**

### IV.2.1. Data Collection:

In this study, we read several Arabic books by the most famous authors and manually extracted and labeled individual sentences.

- Data is randomly collected from many books and the Internet (human development, history, fairy tales, fiction, medicine) and is not repeated. These books were collected from Large Arabic Book Reviews (LABR), a dataset of more than 63,257 his book reviews collected from www.goodreads.com.

- The collected data are divided into 9709 sentences.
- This paragraph is addressed to three Arab experts with doctorates graduate from different faculties and gives opinions on the text.
- SPSS was used to evaluate expert judgment using statistical methods (validity and reliability). It is recommended to accept Cronbach's Alpha reliability coefficient values greater than 0.6. Table 11 shows statistical measures of reliability, Cronbach's alpha is 80.4%.

**Table 11:Reliability statistics**

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| **0.804** | **0.80** | **3** |

- Since this ratio reflects the similarity of the duplicates in Table 12 among the experts, the results are accepted as the values are within the acceptable range.

**Table 12:Inter-item correlation matrix.**

| | Judge 1 | Judge 2 | Judge 3 |
|---|---|---|---|
| **Judge1** | 1,000 | 0.397 | 0.511 |
| **Judge2** | 0.397 | 1,000 | 0.808 |
| **Judge3** | 0.511 | 0.808 | 1,000 |

- In the end, the experts eliminated about 3000 sentences and accepted the rest, including 454 negative sentences, 802 neutral sentences and 7047 positive sentences. Therefore, Figure 13 shows that the collected data represent nearly 10,000 passages. Table 3 shows the CASAD sample taken from the Arabic book as an example.

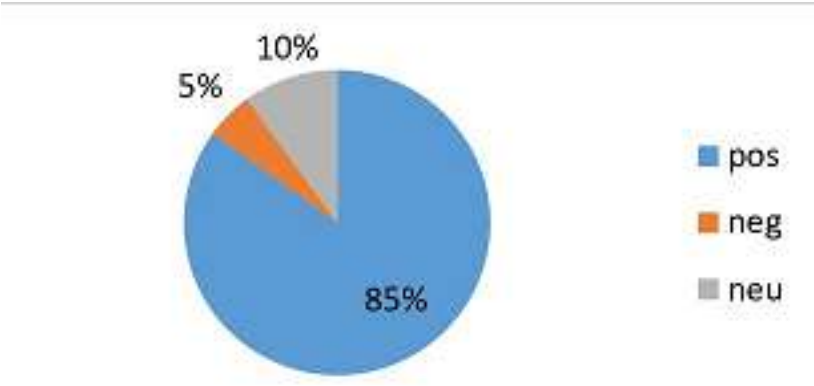


**Figure 13: Classification of book-author ratings.**

### IV.2.2. CASAD Preprocessing:

The book set may have been written sloppily and may be full of errors and abbreviations. As a result, CASAD needs to be cleaned and filtered. Purification, reduction of impurities, and body preparation for better processing are some of the main goals of NLP. In this study, the majority of the preprocessing steps were:

#### IV.2.2.1. Tokenization:

To study the polarity of each document or comment, it is necessary to break the comment into sentences and then into words using NLP preprocessing. Each comment is broken into sentences using NLP processing

The sentence is then broken down into words or terms by tokens. Tokenization can be defined as one of the parsing tasks to distinguish an input character string from another figure; all these tokens are used in the analysis, such as deriving and marking parts of speech on one side, or calculating the probability of word positions in a text using embedding techniques of the word in another place.

#### IV.2.2.2. Stemming:

Stemming is the process of removing suffixes and prefixes from words to reduce redundant tokens in a corpus. It's simply a change from plural to singular, or a verb derived from the gerund form. Other methods exist, such as deriving the roots of pattern words. For example, the root of the words ***علم (أعلام:Flags),( تعليم:Education), (معلم:Teacher),( العلم:Science), (معلوم:Known), (معلومة:Information), (عالم:Scientist).*** In feature selection methods, stems are used. They are the origin of all words with the same morphology. In Arabic, the root or stem is usually made up of three or four letters. Gravitation is important in classification and index/searcher construction because it makes operations less dependent on specific word forms and reduces the potential size of the vocabulary, which must contain all possible forms by definition.

#### IV.2.2.3. Part of Speech Tagging (POS):

Marks words in a sentence that correspond to a particular part of speech based on definition and context. It has been used for various SA tasks such as SentiWordNet and is very useful as it provides linguistic signals about how words are used in the context of a phrase, sentence, or document.

#### IV.2.2.4. Stop Word Filtering:

Basically, stop words are a set of common words in any language. Therefore, stop words are usually removed from the corpus and the proposed approach focuses only on important words (non-stop words). For example the words (“and:و”,”but:لكن ”,”how :كيف ”,”or:أو”, and ”what: ماذا أو ما ”).

## IV.3. TFIDF Feature Extraction:

The first step in this study was to collect a corpus of 8,303 documents and manually classify them into three categories by human experts. Positive, Negative, and Neutral paragraphs. In this part, the features represented by the feature extraction vector are:

- ✓ For each category of the gathered dataset, documents are represented as a vector of words, like succeed. An inverse term of the term frequency and a count vector can be used to denote this success (النجاح) as |V|as in:
  - ➢ log normalization of the term frequent TF in formula (40) as:

$$\mathbf{TF = \log \boldsymbol{f_{t,d}}} \qquad (\mathbf{40})$$

Where $f_{t,d}$ is the raw count of a term $t$ a document $d$.

- ✓ inverse document frequency $IDF$ is a measure of how much information the word provides, i.e., if it is common or rare across all documents in formula (41) as:

$$\boldsymbol{IDF = \log \frac{N}{n_t}} \qquad (\mathbf{41})$$

- ✓ Then, term frequency-inverse document frequency $TFIDF$ I have explained the TF IDF in detail in Chapter 2 are calculated in formula (42 ) as :

$$\boldsymbol{TFIDF(t,d,D) = TF(t,d)IDF(t,D) = \ \ 1 + \log f_{t,d} \ .\mathrm{Log}\frac{N}{n_t}} \quad \textbf{(42)}$$

Where$N$: Total number of documents in the corpus $N$ = | $D$ | $N$ = | $D$ |

|{ $d \in D$: t $\in d$ } || { $d \in D$: $t \in d$ } | : number of documents where the term t appears (i.e., TF($t, d$)≠ 0TF($t, d$) ≠ 0). If the term is not in the corpus, this will lead to a division by zero.It is, therefore, common to adjust the denominator to 1 + |{ $d \in D$ : $t \in d$ }|1 + |{ $d \in D$ : $t \in d$ }|.

## IV.4. Word2vec Feature Extraction:

Text mining and natural language processing (NLP) are both used in sentiment analysis. The research focuses on English SA a lot. There are, however, not many Arabic-language studies on SA. We employ the word embedding or word dispersion approaches to power NLP sentiment analysis tasks in order to get around Arabic's complexity. Word embeddings are techniques that can be used to estimate the probabilities of the location of word domains in a text and the relationships between them. The most important of these techniques is called Word2Vec, which uses deep learning techniques to extract relationships between words. The Word2vec model achieved good result in with SA Arabic language. I have clarified the word embedding briefly in Chapter 2.

It is a neural network with two layers. The corpus of text is represented by the input layer, one of these layers. The output of a set of vectors known as feature vectors for the words in that corpus, on the other hand, is what makes up the outer layer. The Word2vec model, such as the following algorithm, is then used to train the corpus and vocabulary.

1. The first component of the method deals with the discovery of the word representation based on Word2vec model. Given that a corpus D consists of a set of texts, D = {$d_1$, $d_2$, $d_3$,..., $d_n$} and a vocabulary T = {$t_1$, $t_2$, $t_3$,..., $t_m$} consists of unique terms extracted from D. Then, the word representation of the terms they are discovered by using the Skip-gram model of the word2vector to calculate the probability distribution of other terms in context given $t_i$. In particular, $t_i$ is represented by a vector $\vec{v_i}$ that is comprised of probabilistic values of all terms in the vocabulary.

2. This word embedding technique can discover semantic relationships among terms in the corpus. However, the resulting set of vectors for all terms in the corpus is high-dimensional and is inefficient for the classifier in the SA task. As a result, this first component discovers a set of vector : $V_T = \{\overrightarrow{v_1}, \overrightarrow{v_2}, \overrightarrow{v_3}, ..., \overrightarrow{v_m}\}$ representing the set of terms in the vocabulary T .

## IV.5. Machine Learning Training:

In this section, CASAD features are trained to estimate the probability of each feature ($f_n$) in the feature-extraction vector with each category in the dataset. A calculation process was performed to implement this step as follows:

- ✓ When collecting the CASAD dataset, calculate the mean ($C_j$) for each category by dividing the number of each category by the total number of all categories in the collected dataset.
- ✓ Calculate the probability of each word ($w_n$) in each category ($C_j$) in CASAD according to the following equation in formula (43) as :

$$\boldsymbol{P(w_n \mid C_j) = \frac{X_n + 1}{X + |V|}} \qquad (43)$$

Where: $X = \{x_1, x_2, x_3, ..., x_n\}$ represents some features number of words (independent variables) in the given category ($C_j$) in the collected dataset, and $X_n$ is the number of times the word occurs in each category ($C_j$) in the collected dataset.

- ✓ Use multiple ML approaches for dataset evaluation and feature extraction. For example, one of the ML training methods, Naive Bayes, calculates the value of ($V_{NB}$) using the following equation in formula (44):

$$\boldsymbol{(V_{NB})\ \mathrm{agrmax}\ P(C_j) \prod_{X_n \in W} P(X_n \mid C_j)} \qquad (44)$$

- ✓ Finally, we use the confusion matrix of each ML approach to evaluate CASAD in two scenarios: multi (3) classes and binary (2) class.

## IV.6. Experimental Discussion and Analysis Results:

Word2Vec and TFIDF are two feature-extraction techniques we use. The imbalance number of each class is resolved using the 5% of the dataset. Additionally, six machine learning algorithms (LR, LDA, KNN, CART, NB, SVM, adapt, Adapt real) was investigated to assess the classification rate of the feature extraction matrices in two and three classes, as well as in binary (Positive and Negative) and multi-class (Positive, Negative, and Neutral) classes.

The confusion matrix is generated in information retrieval, natural language processing, and classification problems to tabulate the performance of any classifier. This matrix depicts the relationship between correctly predicted and incorrectly predicted author texts. Different Performance evaluation parameters, such as Precision, are a measure of result relevancy, whereas Recall is a measure of how many truly relevant results are returned based on this confusion matrix. Precision and recall are therefore based on understanding and measuring relevance. These numbers are also associated with the F-measure, which is defined as the harmonic mean of precision and recall. Accuracy is the metric used to evaluate the classification model and (Δ) is the feature calculated by dividing the difference between Acc (TFIDF) and Acc (Wor2Vec) were calculated in formula (45):

$$\boldsymbol{\Delta} = \frac{\boldsymbol{Acc(TFIDF) - Acc(Wor2Vec))}}{\boldsymbol{Acc(TFIDF)}} \qquad (\mathbf{45})$$

In this section, we compare manual and automatic classification results. In this comparison, we calculate the precision, recall, F1, precision and (Δ) for the positive, negative and neutral classifiers using the precision and recall determined by the following formulas mentioned in chapter 2. In the confusion matrix, TP (True Positives) represents the number of correctly predicted positive author texts, and FP (False Positives) represents the value of the number of positive author texts that the classifier predicted as negative. Similarly, TN (True Negative) is the number of correctly predicted negative author texts and FN (False Negative) is the number of negative author texts positively predicted by the classifier.

### IV.6.1. Multi class Experiments:

Tables 13 and 14 show the scoring matrices for various machine learning approaches used to score the proposed Arabic dataset, which is described in the following section.

Table 13 displays the score matrix for various machine learning approaches across three grades (Positive, Negative, and Neutral). These methods were used to evaluate the proposed CASAD described in the preceding section.

**Table 13: Comparison of TFIDF and Word2Vec feature extraction in three classes (Positive, Negative, and Neutral).**

| ML | class | TFIDF | | | | Wor2Vec | | | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Pre. | Rec. | F1 | Acc. | Pre. | Rec. | F1 | Acc. | |
| LR | pos | 0.62 | 0.40 | 0.49 | 52.381 | 0.65 | 0.39 | 0.49 | 53.333 | 0.018 |
| | neg | 0.60 | 0.59 | 0.60 | | 0.60 | 0.58 | 0.59 | | |
| | neu | 0.41 | 0.60 | 0.49 | | 0.43 | 0.66 | 0.52 | | |
| | Avg. | 0.55 | 0.52 | 0.52 | | 0.57 | 0.53 | 0.53 | | |
| LDA | pos | 1.00 | 0.03 | 0.05 | 38.095 | 0.00 | 0.00 | 0.00 | 33.81 | -0.112 |
| | neg | 0.35 | 1.00 | 0.52 | | 0.34 | 1.00 | 0.51 | | |
| | neu | 0.88 | 0.11 | 0.20 | | 0.00 | 0.00 | 0.00 | | |
| | Avg. | 0.75 | 0.38 | 0.25 | | 0.11 | 0.34 | 0.17 | | |
| KNN | pos | 0.38 | 0.84 | 0.53 | 40.0 | 0.38 | 0.86 | 0.53 | 39.524 | -0.012 |
| | neg | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | | |
| | neu | 0.46 | 0.31 | 0.37 | | 0.47 | 0.27 | 0.35 | | |
| | Avg. | 0.28 | 0.40 | 0.30 | | 0.28 | 0.40 | 0.30 | | |
| CART | pos | 0.42 | 0.49 | 0.45 | 39.524 | 0.43 | 0.45 | 0.44 | 41.429 | 0.048 |
| | neg | 0.47 | 0.38 | 0.42 | | 0.43 | 0.34 | 0.38 | | |
| | neu | 0.29 | 0.29 | 0.29 | | 0.39 | 0.45 | 0.42 | | |
| | Avg. | 0.40 | 0.40 | 0.39 | | 0.42 | 0.41 | 0.41 | | |
| NB | pos | 0.51 | 0.47 | 0.49 | 52.381 | 0.50 | 0.42 | 0.45 | 49.524 | -0.055 |
| | neg | 0.59 | 0.66 | 0.63 | | 0.54 | 0.63 | 0.58 | | |
| | neu | 0.45 | 0.44 | 0.44 | | 0.43 | 0.44 | 0.43 | | |
| | Avg. | 0.52 | 0.52 | 0.52 | | 0.49 | 0.50 | 0.49 | | |
| SVM | pos | 0.00 | 0.00 | 0.00 | 29.524 | 0.00 | 0.00 | 0.00 | 29.524 | 0.000 |
| | neg | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **neu** | 0.30 | 1.00 | 0.46 | | 0.30 | 1.00 | 0.46 | | |
| | **Avg.** | 0.09 | 0.30 | 0.13 | | 0.09 | 0.30 | 0.13 | | |
| **Adapt** | **pos** | 0.00 | 0.00 | 0.00 | 31.429 | 0.00 | 0.00 | 0.00 | 37.143 | 0.182 |
| | **neg** | 0.80 | 0.06 | 0.11 | | 0.35 | 0.96 | 0.51 | | |
| | **neu** | 0.30 | 1.00 | 0.46 | | 0.62 | 0.16 | 0.26 | | |
| | **Avg.** | 0.36 | 0.31 | 0.17 | | 0.30 | 0.37 | 0.25 | | |
| **Ad_rel** | **pos** | 0.40 | 0.88 | 0.55 | 42.381 | 0.40 | 0.83 | 0.54 | 42.381 | 0.000 |
| | **neg** | 0.47 | 0.21 | 0.29 | | 0.47 | 0.21 | 0.29 | | |
| | **neu** | 0.75 | 0.10 | 0.17 | | 0.59 | 0.16 | 0.25 | | |
| | **Avg.** | 0.53 | 0.42 | 0.35 | | 0.48 | 0.42 | 0.37 | | |

Table 13 shows the evaluation matrix of several machine learning methods. These approaches were applied to evaluate the proposed Arabic dataset in three classes (positive, negative and neutral) with TFIDF feature extraction having an average accuracy of 75% with LDA and average and F1 recoveries were 52% with LR and NB. And with Word2vec feature extraction, they have an average accuracy of 57% and an average and F1 recall of 53% with LR. Accuracy refers to the overall accuracy. As a result, the occurrence rate of TFIDF was almost the same at 52% between LR and NB, but the occurrence rate of Word2vec in LR showed the highest accuracy at 53%. This is because word2vec can extract more hidden relations between texts.

### IV.6.2. Binary-Class Experiments:

Table 14 shows the evaluation matrix of several machine-learning approaches that were applied to evaluate the proposed Arabic dataset in two classes(positive and negative).

**Table 14:Comparison between TFIDF and Word2Vec feature extraction in two classes (Positive and Negative).**

| **ML** | **class** | TFIDF | | | | Wor2Vec | | | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Pre.** | **Rec.** | **F1** | **Acc.** | **Pre.** | **Rec.** | **F1** | **Acc.** | |
| **LR** | **pos** | 0.64 | 0.83 | 0.72 | 71.429 | 0.62 | 0.84 | 0.72 | 70.0 | -0.02 |
| | **neg** | 0.81 | 0.62 | 0.71 | | 0.82 | 0.58 | 0.68 | | |
| | **Avg.** | 0.74 | 0.71 | 0.71 | | 0.73 | 0.70 | 0.70 | | |
| | **pos** | 0.80 | 0.13 | 0.22 | | 0.00 | 0.00 | 0.00 | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **LDA** | **neg** | 0.58 | 0.97 | 0.72 | 59.286 | 0.55 | 1.00 | 0.71 | 55.0 | -0.07 |
| | **Avg.** | 0.68 | 0.59 | 0.50 | | 0.30 | 0.55 | 0.39 | | |
| **KNN** | **pos** | 0.59 | 0.86 | 0.70 | 66.42 | 0.56 | 0.86 | 0.68 | 63.57 | -0.04 |
| | **neg** | 0.81 | 0.51 | 0.62 | | 0.80 | 0.45 | 0.58 | | |
| | **Avg.** | 0.71 | 0.66 | 0.66 | | 0.69 | 0.64 | 0.62 | | |
| **CART** | **pos** | 0.52 | 0.68 | 0.59 | 57.143 | 0.53 | 0.63 | 0.58 | 58.571 | 0.02 |
| | **neg** | 0.65 | 0.48 | 0.55 | | 0.65 | 0.55 | 0.59 | | |
| | **Avg.** | 0.59 | 0.57 | 0.57 | | 0.60 | 0.59 | 0.59 | | |
| **NB** | **pos** | 0.69 | 0.52 | 0.59 | 67.857 | 0.69 | 0.52 | 0.59 | 67.857 | 0.00 |
| | **neg** | 0.67 | 0.81 | 0.73 | | 0.67 | 0.81 | 0.73 | | |
| | **Avg.** | 0.68 | 0.68 | 0.67 | | 0.68 | 0.68 | 0.67 | | |
| **SVM** | **pos** | 0.45 | 1.00 | 0.62 | 45.0 | 0.45 | 1.00 | 0.62 | 45.0 | 0.00 |
| | **neg** | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | | |
| | **Avg.** | 0.20 | 0.45 | 0.28 | | 0.20 | 0.45 | 0.28 | | |
| **Adapt** | **Pos** | 0.46 | 0.83 | 0.59 | 48.571 | 0.46 | 0.83 | 0.59 | 47.857 | -0.01 |
| | **Neg** | 0.59 | 0.21 | 0.31 | | 0.58 | 0.19 | 0.29 | | |
| | **Avg.** | 0.53 | 0.49 | 0.44 | | 0.52 | 0.48 | 0.42 | | |
| **Ad_rel** | **Pos** | 0.46 | 0.83 | 0.59 | 49.286 | 0.45 | 0.83 | 0.58 | 47.143 | -0.04 |
| | **Neg** | 0.61 | 0.22 | 0.32 | | 0.56 | 0.18 | 0.27 | | |
| | **Avg.** | 0.54 | 0.49 | 0.45 | | 0.51 | 0.47 | 0.41 | | |

Table 14 shows the evaluation matrix of several machine learning methods. With TFIDF feature extraction; has an average accuracy of 74% and an average and F1 recall of 71% with LR. And with Word2Vec feature extraction, the average accuracy is 73% and the average and F1 recall is 70% with LR. Accuracy refers to the overall accuracy, with the results indicating that the occurrence of TFIDF gives the highest accuracy of 71.42% with LR and delta only accurate 2.04%. However, NB achieves a significant difference between the TFIDF and Word2Vec ranges of 7.8%.

On the other hand, Tables 13 and 14 shows that our approach makes more sense in extracting polarization in two layers than in multiple layers. This is due to the negative effect of neutral words on ML training, as there are some challenges in omitting ineffective neutral words from the training model.

In the figures 14, 15, 16, and 17 shows the distribution of randomly distributed training, development, and test scores. To solve this problem, we use 10-fold validation to measure the average of the results extracted from the machine learning model. However, the standard deviation of the 10-fold result reflects the stability of the model. Therefore, models (TFIDF or Word2vec) are computed in binary or multi-class polarity with one standard deviation.

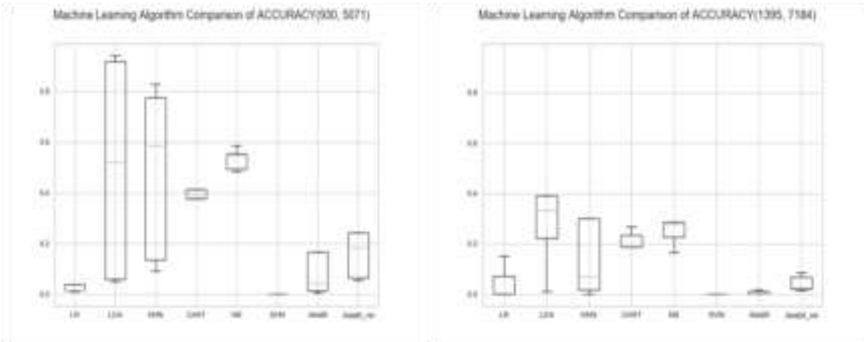


**Figure 14: Machine-learning accuracy.**

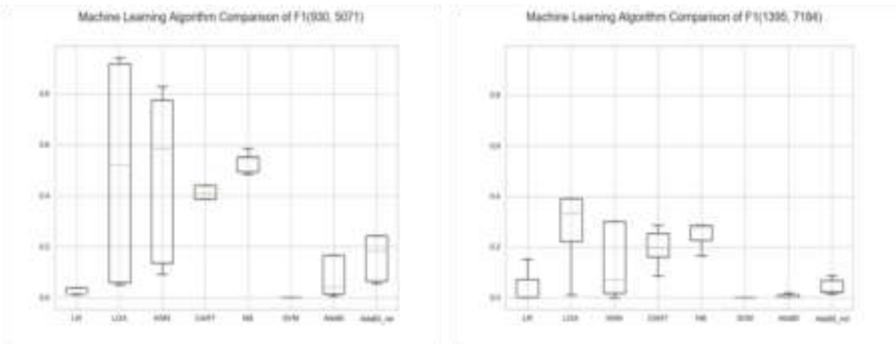


**Figure 15: Machine-learning F-score.**

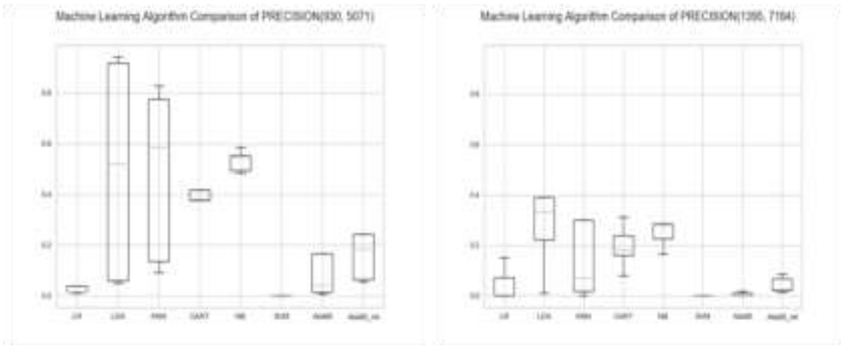


**Figure 16: Machine-learning precision.**

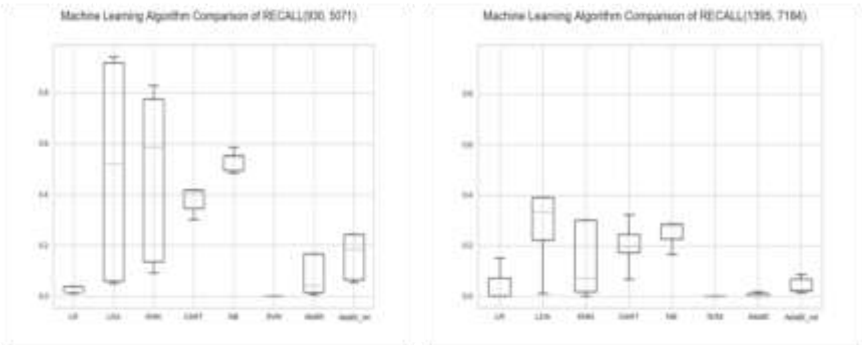


**Figure 17: Machine-learning recall.**

Accordingly, variance helps to find the distribution of data in a population relative to the mean, and standard deviation also helps to know the distribution of data in a population. However, the standard deviation in the numbers clarifies the deviation from the mean. Therefore, the binary classifier outperformed the multi-class classes after using 10-fold cross-validation with standard deviation. On the other hand, LDA obtained the best results compared to other machine learning methods for multi- and binary classification.

## IV.7. Conclusion:

This chapter reviewed a research paper that presented an overview understanding of the Arabic sentiment analysis. We shed light on; the importance of the topic and the challenges of Arabic language sentiment analysis. This chapter reviewed a research paper that presented an overview understanding of the Arabic sentiment analysis. We shed light on; the importance of the topic and the challenges of Arabic language sentiment analysis. In addition, feature extraction of three and two classes (word2vec and TFIDF) of CASAD was evaluated using six ML approaches. Although the binary class results were superior to his three class results, classical

Arabic accuracy requires further study to clarify the ambiguity behind the author's opinion. The accuracy score for Arabic text classification using the most popular ML algorithms such as SVM, LR and Naive Bayes was 71.42% of LR on word2vec for binary classification (Pos and Neg). On the other hand, the polarity extraction approach is more important for his two classes than for multiple classes. This is due to the negative effects of neutral terms in ML training, as there are some challenges in ignoring ineffective neutral terms from training models. Therefore, deep learning for CASAD classification provides more accurate rates than standard feature extraction approaches. Consequently, this work will address in future work.

# General Conclusion

# General Conclusion

Research in the field of sentiment analysis has been very active in recent years due to many challenging research problems and their practical applications. Although this research fills in some gaps, further work is needed to facilitate improvements. This section presents a summary of the thesis, followed by a summary of the research hypotheses, and suggestions for further work are given. The ability to identify sentiment in a language with minimal resources is necessary if we want to build natural language processing systems capable of aggregating, reporting, and responding to sentiments expressed on many genres and applications, including sentiments expressed in written review in high-resource languages but also the emotions expressed towards real-life entities, issues, and events in many parts of the world where hundreds or even thousands of low-resource languages spoken and written.

Although many of the concepts will apply to other languages, especially the Arabic languages, additional Arabic effective vocabularies in other fields could also be developed, which can be used for further research that can expand the boundaries of research in the field of Arab sentiment analysis.

This study focuses only on the Arabic language. This thesis presents resources, techniques, and strategies for identifying target emotions using cross-validation with unlabeled data from a new classical Arabic dataset. Additionally, feature extraction from these datasets is created using equivalent word embedding techniques, which can extract deep relationships that represent the features of the formal Arabic language. In contrast to previously published work in the area, our work covered ground in the problem of cross-validation sentiment analysis.

Through the work presented in the thesis, we aim to encourage future research in the field of Arab sentiment analysis and related applications that can utilize our models and resources. The complexity of Arabic as a target language in sentiment analysis makes these tasks more difficult.

These challenges will encourage researchers involved in the project to develop ideas to address these problems. One possible adaptation mechanism is to use an additional general sentiment corpus along with the training domain. In our case, we can use word2vec to calculate the sentence polarity score and use it as a sentiment analysis feature. This method can provide additional knowledge that can help the classifier learn and make generalizations about sentiment issues in Arabic text.

In addition, recent developments in bidirectional language model-based techniques such as BERT have brought significant successes to a number of lower-level natural language processing tasks, such as question answering, natural language inference , and textual implicative. Using BERT pre-training and fine-tuning our sentiment transfer models is another avenue for future work titled "Semi-Supervised Arabic Sentiment Analysis Prediction Using BERT with Rearrange Sentences". The purpose of this work is to rearrange Arabic sentences based on subjective words such as adjectives and nouns.

In conclusion, we hope that our targeted sentiment and emotion models will be used in applications that ensure the effectiveness of Arabic language.

## References


[1] Knowles, E. "Merriam-Webster Unabridged (Unabridged. Merriam-Webster. com)," *Dictionaries: Journal of the Dictionary Society of North America,* vol. 35, (35), 334--339, 2014.

[2] Quirk, G. Svartvik, G. and Crystal, "A comprehensive grammar of the English language", 397, 1985.

[3] Pang, B.Lee and L, "Opinion mining and sentiment analysis," *Foundations and Trends in information retrieval,* 2, (1--2), 1--135, 2008.

[4] Liu, B."Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies,* "1--167, 2012.

[5] Rehma, Awan, A. and, Butt, M. and Ilyas, "Comparison and evaluation of information retrieval models," 2018.

[6] Hajjaji, Y.Boulila, W.Farah, I.Romdhani, I.Hussain and Amir, "big data and IoT-based applications in smart environments: A systematic review," *Computer Science Review,* 39, 100318, 2021.

[7] Ahmed, Awan, H. Javed, Khan, M. and, Yasin, Shehzad, Muhammad and Hafiz, "Sentiment analysis of online food reviews using big data analytics," *Elementary Education Online*, 20, (2), 827--836, 2021.

[8] Socher, Pennington, R. and, Huang, J. and, Ng, and, Manning, and, D and Christopher, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011.

[9] Popescu, A.Etzioni and Orena, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*, 2007.

[10] Shi,Wang,W.He,and Shaoyi, "Sentiment analysis of Chinese microblogging based on sentiment ontology: a case study of ‘7.23 Wenzhou Train Collision’," *Connection Science,* 25, (4),161--178, 2013.

[11] Dave, K.Lawrence, S.Pennock and D. M, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,"519--528, 2003.

[12] Khan, K.Baharudin, B.Khan, A.Ullah and Ashraf, "Mining opinion components from unstructured reviews: A review," *Journal of King Saud University-*

*Computer and Information Sciences*, 26,(3), 258--275, 2014.

[13] Himmat, M.Salim and Naomie, "Survey on product review sentiment classification and analysis challenges," in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, 2014.

[14] Atteveldt, W.Velden, der, Boukes, M. and Mark, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Communication Methods and Measures*. 15, (2), 121--140, 2021.

[15] Mohammad and S. M, "{Challenges in sentiment analysis," in *A practical guide to sentiment analysis*, 2017.

[16] Poria, S. Hazarika, D. Majumder, N. .Mihalcea and Rada, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, 2020.

[17] Nazir, Rao, A. and, Wu, Y. and, Sun, L. and Ling, "Issues and challenges of aspect-based sentiment analysis: a comprehensive survey," *IEEE Transactions on Affective Computing*, 2020.

[18] Chakraborty, Bhatia, a. Koyel, Bhattacharyya, S. and, Platos, S. and, Bag, J. and, Hassanien, R. and, Ella and Aboul, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, 97, 106754, 2020.

[19] Filho, Balage, Pardo, P. and Thiago, "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in *Second Joint Conference on Lexical and Computational Semantics ( SEM), 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013.

[20] Kim, S.M. Hovy and Eduard, "Automatic detection of opinion bearing words and sentences," in *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*, 2005.

[21] Gamon, M. Aue,Corston-Oliver, S. Ringger and Eric, "Pulse: Mining customer opinions from free text," in *international symposium on intelligent data analysis*, 2005.

[22] Arulmurugan, Sabarmathi, R. and, Anandakumar, K. and HJCC, "Classification of sentence level sentiment analysis using cloud machine learning techniques," *Cluster Computing*, 22, (1), 1199--1209, 2019.

**References**


[23] Wiebe, Bruce, J. and, O'Hara, R.Thomas, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999.

[24] Wilson, Wiebe, T. and, Hoffmann, J. and Paul, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005.

[25] Narayanan, Liu, R. and, Choudhary, B. and Alok, "Sentiment analysis of conditional sentences," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009.

[26] Tang, Tan, H. and, Cheng, S. and and Xueqi, "A survey on sentiment detection of reviews," *Expert Systems with Applications,* 36,(7), 10760--10773, 2009.

[27] Pang, B. Lee and Lillian, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *arXiv preprint cs/0409058*, 2004.

[28] Liu and B others, "Sentiment analysis and subjectivity.," *Handbook of natural language processing,* 2, (2010),627--666, 2010.

[29] Steinberger, Brychc, J. and, Konkol, T. and Michal, "Aspect-level sentiment analysis in czech," in *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2014.

[30] Long, C. Zhang, Zhu and Xiaoyan, "A review selection approach for accurate feature rating estimation," in *Coling 2010: Posters*, 2010.

[31] Wang, Lu, H.Zhai, Y. and ChengXiang, "Latent aspect rating analysis without aspect keyword supervision," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

[32] Mate, C. "Product aspect ranking using sentiment analysis: a survey," *International Research Journal of Engineering and Technology,* 3, (01), 126--127, 2015.

[33] Hu, Liu, M. and Bing, "Mining opinion features in customer reviews," in *AAAI*, 2004.

[34] Chen, Z. Mukherjee, A. Liu, B. Hsu, M. Castellanos, M. Ghosh and Riddhiman, "Exploiting domain knowledge in aspect extraction," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

**References**


[35] Feldman and Ronen, "Techniques and applications for sentiment analysis," *Communications of the ACM,* 56, (4), 82--89, 2013.

[36] Schouten, Frasincar, K. and and Flavius, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering,* 28, (3), 813--830, 2015.

[37] Phan, Tran, H. Nguyen, V. and, Hwang, and Dosam, "Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model," *IEEE Access,* 2022.

[38] Han, Y.Liu, M. Jing and Weipeng, "Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfe," *IEEE Access,* 8, 21314--21325, 2020.

[39] Wang, Chen, Y. and, Shen, Q. and, Hou, J. and, Ahmed, B. and, Li, M.and Zhanhuai, "Aspect-level sentiment analysis based on gradual machine learning," *Knowledge-Based Systems*, 212, 106509, 2021.

[40] Turney, D and Peter, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," 2002.

[41] Bharti, Vachha, S. and, Pradhan, Babu, Kumar and Sanjay, "Sarcastic sentiment detection in tweets streamed in real time: a big data approach," *Digital Communications and Networks,* 2, (3), 108--121, 2016.

[42] Al-Rubaiee, Qiu, H. and, Li, R. and Dayou, "Identifying Mubasher software products through sentiment analysis of Arabic tweets," in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, 2016.

[43] Haddi and Emma, "Sentiment analysis: text, pre-processing, reader views and cross domains," 2015.

[44] Zhang, Zhang, S. and, Yang, C. and Qiang, "Data preparation for data mining," *Applied artificial intelligence,* 17, (5-6), 375--381, 2003.

[45] Paltoglou and Georgios, "Sentiment analysis in social media," in *Online Collective Action*, 2014.

[46] Weitzel, Prati, L. and, Aguiar, R.and, Freire and Raul, "The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques?," in *Sentiment analysis and ontology engineering*, 2016.

[47] Qazi, Raj, A. and, Hardaker, R.and, Standing, G. and Craig, "A systematic literature review on opinion types and sentiment analysis techniques: Tasks and

challenges," *Internet Research*, 2017.

[48] Raez, Martinez-Camara, A. and, Martin-Valdivia, E. and, Urena-Lopez, M. T. and, Alfonso and L, "Ranked wordnet graph for sentiment polarity classification in twitter," *Computer Speech \& Language,* 28,(1),93--107, 2014.

[49] Esuli, Sebastiani, A. and Fabrizio, "Determining the semantic orientation of terms through gloss classification," 617--624, 2005.

[50] Medhat, W.Hassan, A.Korashy and Hoda, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal,* 5,(4),1093--1113}, 2014.

[51] Pang, B. Lee, Vaithyanathan ,and Shivakumar, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs0205070,* 2002.

[52] Aue, Gamon, A.and Michael, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of recent advances in natural language processing (RANLP)*, 2005.

[53] Blitzer, Dredze, J. and, Pereira, M,and Fernando, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007.

[54] Hatzivassiloglou, V.McKeown and Kathleen, "Predicting the semantic orientation of adjectives," in *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, 1997.

[55] Hu, X. Tang, J, Liu and Huan, "Unsupervised sentiment analysis with emotional signals," in *Proceedings of the 22nd international conference on World Wide Web*, 2013.

[56] "Are SentiWordNet scores suited for multi-domain sentiment classification?," in *2009 Fourth International Conference on Digital Information Management*, 2009.

[57] Khan, M. Durrani, M. Ali, Inayat, I Khalid, Khan and K. Habib, "Sentiment analysis and the complex natural language," *Complex Adaptive Systems Modeling,* 4,(1),1--19, 2016.

[58] Mowlaei, Abadeh, M. and, Keshavarz, M, and Hamidreza, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," *Expert Systems with*

*Applications*, 148,13234, 2020.

[59] Prabowo, Thelwall, R. and Mike, "Sentiment analysis: A combined approach," *Journal of Informetrics,* 3, (20),143--157, 2009.

[60] Tan, Na, Theng, J.Chang, Y. and and Kuiyu, "Sentence-level sentiment polarity classification using a linguistic approach," in *International Conference on Asian Digital Libraries,* 2011.

[61] Mitkov and Ruslan, "Computer-Aided Language Processing.," in *FLAIRS Conference*, 2006.

[62] Dasgupta, S. Ng and Vincent, "Topic-wise, sentiment-wise, or otherwise? Identifying the hidden dimension for unsupervised text classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* 2009.

[63] Mevskele, Frasincar, D. and and Flavius, "ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model," *Information Processing \& Management,* 57, (3), 102211, 2020.

[64] Asif, Ishtiaq, M. and, Ahmad, Aljuaid, H. and, Shah, H. "Sentiment analysis of extremism in social media from textual information," *Telematics and Informatics,* 48,101345, 2020.

[65] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge university press, 2020.

[66] Zad, Heidari, and, Jones, Uzuner,and and Ozlem, "A survey on concept-level sentiment analysis techniques of textual data," in *2021 IEEE World AI IoT Congress (AIIoT)*, 2021.

[67] Glaser, Dixit, Green and D, "Studying hate crime with the internet: What makes racists advocate racial violence?," *Journal of Social Issues*, 58, (1) 177--193, 2002.

[68] Abbasi, A. Chen, Salem and Arab, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM transactions on information systems (TOIS),*26,(3), 1--34, 2008.

[69] Talwar, A. Kumar and Yogesh, "Machine Learning: An artificial intelligence methodology," *International Journal of Engineering and Computer Science,* 2, (12), 3400--3404, 2013.

**References**

[70] Agarwal, Mittal, B. and Namita, "Machine learning approach for sentiment analysis," in *Prominent feature extraction for sentiment analysis*, 2016.

[71] Mitchell, Mitchell, M and Tom, "Machine learning," 1, (9), 1997.

[72] Nasteski and Vladimir, "An overview of the supervised machine learning methods," *Horizons. b,* (4), 51--62, 2017.

[73] Jiang, T. Gradus, Rosellini . "Supervised machine learning: a brief primer," *Behavior Therapy*, 51,(5), 675--687, 2020.

[74] Ribeiro, Araujo, F. Gonccalves, M. and, Gonccalves, P. "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science*, 5, (1),1--29, 2016.

[75] Moraes, Valiati, R. and, Neto, J. "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications,* 40,( 2), 621--633, 2013.

[76] Amancio, Comin, D. and, Casanova, H. and, Travieso, D. and, Bruno, G. and, Rodrigues. Fontoura and Luciano, "A systematic comparison of supervised classifiers," *PloS one,* 9, ( 4), e94137, 2014.

[77] Chen, Hsieh, R. and Chung-Hsun, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Systems with Applications,* 31, (2), 427--435, 2006.

[78] Ramasamy, Kadry, S. and and Sangsoon, "Selection of optimal hyper-parameter values of support vector machine for sentiment analysis tasks using nature-inspired optimization methods," *Bulletin of Electrical Engineering and Informatics*, 10,(1), 290--298, 2021.

[79] Naz, Sharan, S.Malik, and Nidhi, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018.

[80] Zainuddin, Selamat, N. and Ali, "Sentiment analysis using support vector machine," in *2014 international conference on computer, communications, and control technology (I4CT)*, 2017.

[81] Biau, Devroye, G. and and Luc, "Lectures on the nearest neighbor method,"246, Springer, 2015.

[82] Cover, T. Hart and Peter, "Nearest neighbor pattern classification," *IEEE transactions on information theory*,13,(1),21--27, 1967.

## References


[83] Triguero, Garcia-Gil, Maillo,Luengo, Garcia, Herrera and Francisco, "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 9,( 2), e1289, 2019.

[84] Tan and Songbo, "An effective refinement strategy for KNN text classifier," *Expert Systems with Applications,* 30, (2), 290--298, 2006.

[85] Daeli, Adiwijaya, N.and and Adiwijaya, "Sentiment analysis on movie reviews using Information gain and K-nearest neighbor," *Journal of Data Science and Its Applications,* 3, (1), 1--7, 2020.

[86] Damarta, R and Hidayat, A and Abdullah."The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality," in *Journal of Physics: Conference Series,* 2021.

[87] Muktafin, Pramono, E. and, Kusrini, P. and and Kusrini, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *TELKOMNIKA (Telecommunication Computing Electronics and Control),*19, (1), 146--154, 2021.

[88] Isnain, A. Supriyanto, Kharisma and M, Pajar. "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* 15, (2),121--130, 2021.

[89] Mustaqim, T Umam, K. Muslim and MA, "Twitter text mining for sentiment analysis on government’s response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," in *Journal of Physics: Conference Series*, 2020.

[90] Zhang and Harry, "The optimality of naive Bayes," *Aa,* 1,( 2), 3, 2003.

[91] Lee and Chang-Hwan, "An information-theoretic filter approach for value weighted classification learning in naive Bayes," *Data Knowledge Engineering,* 113,116--128, 2018.

[92] Osisanwo, Akinsola, F. Awodele, J, and, Akinjobi. "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT),* 48, 3),128--138, 2017.

[93] Rish and Irina and others, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001.

## References


[94] Lewis and D, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*, 1998.

[95] Dey, L. Chakraborty, Biswas, A. Bose, B. Tiwari and Sweta, "Sentiment analysis of review datasets using naive bayes and k-nn classifier," *arXiv preprint arXiv:1610.09982*, 2016.

[96] Rodrigues, Chiplunkar, A. and,Niranjan, "A new big data approach for topic classification and sentiment analysis of Twitter data," *Evolutionary Intelligence*, 15, (2), 877--887, 2022.

[97] Kewsuwun, N. Kajornkasirat and Siriwan, "A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier.," *International Journal of Electrical \& Computer Engineering (2088-8708),*. 12, (3), 2022.

[98] Sudiro, Prasetiyowati, R. and, Sibaroni,and Yuliant, "Aspect Based Sentiment Analysis with Combination Feature Extraction LDA and Word2vec," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 2021.

[99] Blei, Ng,and, Jordanand M."Latent dirichlet allocation," *Journal of machine Learning research*,3, (Jan), 993--1022, 2003.

[100] Abelard, Sibaroni, A,and Yuliant, "Multi-aspect sentiment analysis on netflix application using latent dirichlet allocation and support vector machine methods," *JURNAL INFOTEL*, 13, (3),128-133, 2021.

[101] Chen, Zanocco, X. and, Flora,and, Rajagopal, and Ram, "Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation," *Applied Energy*, 318, 119109, 2022.

[102] joung, J.Kim and H. "Automated keyword filtering in latent Dirichlet allocation for identifying product attributes from online reviews," *Journal of Mechanical Design*, 143, (8), 2021.

[103] Gurcan, F. Ozyurt, O.Cagitay and N. Ercil, "Investigation of emerging trends in the e-learning field using latent Dirichlet allocation," *International Review of Research in Open and Distributed Learning*, 22, (2), 1--18, 2021.

[104] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies* 3, (3),281--299, 2011.

[105] Kalantar, B. Pradhan, and B. Naghibi. "Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between

support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN)," *Geomatics, Natural Hazards and Risk,* 9,(1), 49--69, 2018.

[106] Tyagi, Sharma, A. and Naresh, "Sentiment analysis using logistic regression and effective word score heuristic," *International Journal of Engineering and Technology (UAE),* 7, (2.24), 20--23, 2018.

[107] Ahuja, Chug, R. and, Kohli, A, and Pratyush, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science,* 152, 341--348, 2019.

[108] Poornima, A.Priya and K. Sathiya, "A comparative sentiment analysis of sentence embedding using machine learning techniques," in *2020 6th international conference on advanced computing and communication systems (ICACCS)*, 2020.

[109] Moulya, Pragathi, S, and TR, "Mental Health Assist and Diagnosis Conversational Interface using Logistic Regression Model for Emotion and Sentiment Analysis," in *Journal of Physics: Conference Series*, 2022.

[110] Ghosh, S.Hazra, Raj and Abhishek. "A comparative study of different classification techniques for sentiment analysis," in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, 2022.

[111] Sang, Guo, X. and, Wu, Fu, and, Xie. "Intensity and stationarity analysis of land use change based on CART algorithm," *Scientific reports,* 9, (1),1--12, 2019.

[112] Carrizosa, Molero-Rio,Morales, C. and Dolores, "Mathematical optimization in classification and regression trees," *Top,*29, (1),5--33, 2021.

[113] Timofeev and Roman, "Classification and regression trees (CART) theory and applications," *Humboldt University, Berlin,* 54, 2004.

[114] Loh and Wei-Yin, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery,* 1, (1), 14--23, 2011.

[115] Quinlan, Ross and J, "Decision trees as probabilistic classifiers," *Proceedings of the Fourth International Workshop on Machine Learning,*31--37, 1987.

[116] Mahjoobi, J. Etemad-Shahidi and A. "An alternative approach for the prediction of significant wave heights based on classification and regression trees," *Applied Ocean Research,*30, (3),172--177, 2008.

[117] Wang, Shin, G. Yoon and Seong, "An improved text classification method for sentiment classification," *Journal of information and communication convergence engineering,*17, (1), 41--48, 2019.

**References**

[118] Loh and Wei-Yin, "Fifty years of classification and regression trees," *International Statistical Review*, 82, (3), 329--348, 2014.

[119] LeCun, Y. Bengio. "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks,* 3361, (10),1995, 1995.

[120] Liu, X. Song,  Liu, S. Zhang and Yudong, "A review of deep-learning-based medical image segmentation methods," *Sustainability,* 13, (3), 1224, 2021.

[121] Indrakumari, Poongodi, R.  Singh, and Kiran, "Introduction to Deep Learning," in *Advanced Deep Learning for Engineers and Scientists*, 2021.

[122] Zhang, Li, W.Li, H. Liu, Chen, H. and, Ding, Y. and Xuanming, "Application of deep learning algorithms in geotechnical engineering: a short critical review," *Artificial Intelligence Review,*54, (8), 5633--5673, 2021.

[123] Shafay, Ahmad, M. and, Salah, Yaqoob, K. Jayaraman, I. and, Omar,and Mohammed, "Blockchain for deep learning: review and open challenges," *Cluster Computing,* 1--22, 2022.

[124] Kamilaris, A.Prenafeta-Boldu and F. X, "Deep learning in agriculture: A survey," *Computers and electronics in agriculture,* 147,70--90, 2018.

[125] Chen, Ran, J.  and Xukan, "Deep learning with edge computing: A review," *Proceedings of the IEEE,* 107,(8), 1655--1674, 2019.

[126] Jain and A. K, "Data clustering: 50 years beyond K-means," *Pattern recognition letters,* 31,(8), 651--666, 2010.

[127] Takahashi and Toshihiro, "Statistical max pooling with deep learning,"2018.

[128] Ioffe, S.Szegedy and Christian, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015.

[129] Liu, Li, C. and, Fei, Y. and, Li, H. and Ping, "Deep skip-gram networks for text classification," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, 2019.

[130] Pan, Yang, S. and Qiang, "A survey on transfer learning," 22,(10),1345--1359, 2010.

[131] Shinde, Shah, P.and Seema, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing*

*communication control and automation (ICCUBEA)*, 2018.

[132] Young, Hazarika, T. and, Poria, D. Cambria, S. andErik, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*,13, (3),55--75, 2018.

[133] Otter, Medina, D, Kalita . "A survey of the usages of deep learning for natural language processing," *IEEE transactions on neural networks and learning systems,* 32,(2), 604--624, 2020.

[134] Li, H. "Deep learning for natural language processing: advantages and challenges," *National Science Review.* 5, (1), 24--26, 2018.

[135] Mikolov, Sutskever, T. and, Chen, I. and, Corrado, K. and, Dean,and Jeff, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 26, 2013.

[136] Wang, Zhou, S. and, Jiang, W. and Chao, "A survey of word embeddings based on deep learning," *Computing,* 102,(3),717--740, 2020.

[137] Bakarov and Amir, "A survey of word embeddings evaluation methods," *arXiv preprint arXiv:1801.09536,* 2018.

[138] Mikolov, T. Chen, K. Corrado, G Dean and Jeffrey, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[139] Lau, J. Baldwin and Timothy, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368,* 2016.

[140] Mendsaikhan, Hasegawa, Yamaguchi, Shimada, Bataa, H. "Identification of cybersecurity specific content using different language models," *Journal of Information Processing,* 28, 623--632, 2020.

[141] Pennington, J. Socher, R. Manning and C. "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* 2014.

[142] Bojanowski, Grave, P. Joulin, and, Mikolov, and Tomas, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics,* 5, 135--146, 2017.

[143] Devlin, J. Chang, M,Toutanova and Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

## References


[144] Chand, Felber, R. and Pascal, "Scalable distribution of XML content with XNet," *IEEE Transactions on Parallel and Distributed Systems*, 19,(4), 447--461, 2008.

[145] Wang, Hui, M. and, Cui, L. and, Liang, Y. and, Liu, R. and Zhenhua, "xnet: Improving expressiveness and granularity for network modeling with graph neural networks," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022.

[146] Boudad, Faizi, N. and, Thami, R.Chiheb,and Raddouane, "Sentiment analysis in Arabic: A review of the literature," *Ain Shams Engineering Journal*, 9, (40), 2479--2490, 2018.

[147] Alsayat, Elmitwally, A, and Nouh, "A comprehensive study for Arabic sentiment analysis (challenges and applications)," *Egyptian Informatics Journal*, 21,(1),7--12, 2020.

[148] Alruily and Meshrif, "Issues of dialectal saudi twitter corpus.," *Int. Arab J. Inf. Technol.*,17,(3),367--374, 2020.

[149] Korayem, M.Crandall, D.Abdul-Mageed and Muhammad, "Subjectivity and sentiment analysis of arabic: A survey," 128--139, 2012.

[150] Abdul-Mageed, Diab, M. "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.," in *LREC*, 2012.

[151] Almuqren, Alzammam, L. and, Alotaibi, A. and, Cristea, S. and, Alhumoud, A. and Sarah, "A review on corpus annotation for Arabic sentiment analysis," in *International Conference on Social Computing and Social Media*, 2017.

[152] Elnagar, A.Einea and Omar, "Brad 1.0: Book reviews in arabic dataset," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 2016.

[153] Elnagar, A.Khalifa, Y. Einea and Anas, "Hotel Arabic-reviews dataset construction for sentiment analysis applications," in *Intelligent natural language processing: Trends and applications*, 2018.

[154] Aly, Atiya, and, Mohamed and Amir, "Labr: A large scale arabic book reviews dataset," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (2: Short Papers)*, 2013.

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

**References**


[156] Nabil, Aly, M. and, Atiya, M. and Amir, "Astd: Arabic sentiment tweets dataset," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015.

[157] Al-Twairesh, Al-Khalifa, N. and, Al-Salman, H. and, Al-Ohali,and Yousef, "Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets," *Procedia Computer Science,* 117, 63--72, 2017.

[158] Al-Moslmi, Albared, T. and, Al-Shabi, M. and, Omar, A. and, Abdulla, "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis," *Journal of information science,*44, (3),345--362, 2018.

[159] Baly, Khaddaj, R. Hajj, A, Shaban, W. and, Bashir and Khaled, "Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets," *arXiv preprint arXiv:1906.01830,* 2019.

[160] Rushdi-Saleh,M. Martn-Valdivia, and Urena L'opez, M, and, Perea-Ortega,and Jose, "OCA: Opinion corpus for Arabic," *Journal of the American Society for Information Science and Technology,*62,(10), 2045--2054, 2011.

[161] Nabil, Aly, Atiya, M. and Amir, "Labr: A large scale arabic sentiment analysis benchmark," *arXiv preprint arXiv:1411.6718,* 2014.

[162] Al-Smadi, Talafha, M. Al-Ayyoub, B. and, Jararweh, M. and Yaser, "Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews," *International Journal of Machine Learning and Cybernetics*,10,(8),2163--2175, 2019.

[163] Badaro, Baly, G. Hajj, R, El-Hajj, H.Shaban, W. and, Habash, K. "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, 2019.

[164] Najar, D. and and Slim, "Opinion mining and sentiment analysis for Arabic on-line texts: application on the political domain," *International Journal of Speech Technology,*20,575--585, 2017.

[165] Alayba, Palade,England, V. and, Iqbal, M. and Rahat, "Arabic language sentiment analysis on health services," in *2017 1st international workshop on arabic script analysis and recognition (asar),* 2017.

[166] Abdulla, Al-Ayyoub, N. and, Al-Kabi, M. Naji and Mohammed, "An extended analytical study of arabic sentiments," *International Journal of Big Data*

*Intelligence 1*,1, (1-2),103--113, 2014.

[167] Wang, H. Hanafy, A. Bahgat, M.Noeman, S. Bommireddipalli , "A system for extracting sentiment from large-scale Arabic social data," in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 2015.

[168] Li, Xie, X. Chen, H. Wang, Deng, J. and Xiaotie, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems,* 69,14--23, 2014.

[169] Al-Hajieh, Redhead, K, and Timothy, "Investor sentiment and calendar anomaly effects: A case study of the impact of Ramadan on Islamic Middle Eastern markets," *Research in International Business and Finance*,25, (3), 345--356, 2011.

[170] Ahmed, Toral, A. and, Shaalan, S. and Khaled, "Agent productivity measurement in call center using machine learning," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2,* 2017.

[171] Qamar, Alsuhibany, Sohail and Syed, "Sentiment classification of twitter data belonging to saudi arabian telecommunication companies," *International Journal of Advanced Computer Science and Applications,* 8, 1.

[172] Najadat, Al-Abdi, H. and, Sayaheen, A. and Yasmeen, "Model-based sentiment analysis of customer satisfaction for the Jordanian telecommunication companies," in *2018 9th International Conference on Information and Communication Systems (ICICS)*, 2018.

[173] AL-Rubaiee, Qiu, H. Alomar, R. and, Li, K. and Dayou, "Sentiment analysis of Arabic tweets in e-learning," 2016.

[174] Al-Harbi, Emam, W. and Ahmed, "Effect of Saudi dialect preprocessing on Arabic sentiment analysis," *International Journal of Advanced Computer Technology,* 4,(6),91--99, 2015.

[175] Aoumeur Nour, E. Ziyong L. Alsari. E. "Improving the polarity of text through word2vec embedding for primary classical arabic sentiment analysis," *Neural processing letters*, 1--16, 2023.