Enhanced Emotion Analysis Model Using Machine Learning in Saudi Dialect: COVID-19 Vaccination Case Study

by

ABDULRAHMAN OSAMA HELMY MUSTAFA

A thesis submitted for the requirements of the degree of Master of Science in Information Technology

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhul-Qi'dah 1445 H - May 2024 G

Enhanced Emotion Analysis Model Using Machine Learning in Saudi Dialect: COVID-19 Vaccination Case Study

by

ABDULRAHMAN OSAMA HELMY MUSTAFA

A thesis submitted for the requirements of the degree of Master of Science in Information Technology

Advisor

Prof. Tarig Mohamed Ahmed

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhul-Qi'dah 1445 H - May 2024 G

وما أوتيتم من العلم إلا قليلا

١٤٤٠هـ

Enhanced Emotion Analysis Model Using Machine Learning in Saudi Dialect: COVID-19 Vaccination Case Study

by

ABDULRAHMAN OSAMA HELMY MUSTAFA

A thesis submitted for the requirements of the degree of Master of Science in Information Technology

Name	Rank	Field	Signature
Advisor			
Tarig Mohamed Ahmed	Professor	Information Technology	
Internal Examiner			
Iftikhar Ahmad	Associate Professor	Information Technology	
External Examiner			
Zahid Ullah Sultan	Associate Professor	Information Systems	

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Dhul-Qi'dah 1445 H - May 2024 G

Copyright

Dedication

I wholeheartedly dedicate this thesis to my beloved father, my first and most revered teacher. To my dearest mother, whose prayers have incessantly embraced my name and who has been a constant source of encouragement throughout this challenging endeavor. A heartfelt dedication to my life companion, my cherished wife, who has stood by my side and shared in my life's journey. To the "light" of my life, my daughter, who journeyed with us from conception until she joyously uttered my name.

Acknowledgments

In the Name of Allah, the Most Gracious, the Most Merciful. All praise and gratitude are due to Allah Almighty, the Most Compassionate and Merciful. Without His gracious assistance, completing this work would have been impossible.

I extend my heartfelt thanks to those who offered invaluable assistance throughout the duration of this thesis. I would like to express my sincere thanks to:

My supervisor, ***Professor Tarig Mohamed Ahmed***, provided invaluable guidance, supervision, and constructive feedback, leading to the successful completion of this work. Working on this thesis under his mentorship has been a source of great satisfaction.

My family for their unwavering support, my dear wife, ***Elaf*** , and my beloved daughter, ***Nour***, for their understanding and encouragement. Their love and encouragement provided the foundation on which this academic endeavor was built.

The administration at my company for their support, flexibility, and encouragement in allowing me the time and resources to pursue and complete this research.

This thesis is a testament to these individuals' collective efforts and collaboration, and I am profoundly grateful for their impact on my academic journey.

Abstract

Sentiment Analysis (SA) and Emotion Analysis (EA) are practical areas of research aimed to auto-detect and recognize the sentiment expressed in a text and identify the underlying opinion towards a specific topic. Although they are often considered interchangeable terms, they have slight differences. The primary purpose of SA is to find the polarity expressed in a text by distinguishing between positive, negative, and neutral opinions. EA is concerned with detecting more emotion categories, such as happiness, anger, sadness, and fear. EA allows the analysis to extract more accurate and detailed results suitable to the applied field.

This work delves into EA within the Saudi Arabian dialect, focusing on emotions related to COVID-19 vaccination campaigns. Our endeavor addresses the need for more research on developing an effective EA machine-learning model for Saudi dialect texts, particularly within the healthcare and vaccinations domain, exacerbated by the absence of EA labeled-tweets corpus. Using a systematic approach, a dataset of 33,373 tweets is collected, annotated, and preprocessed. Thirty-six machine learning experiments encompassing Support Vector Machine, Logistic Regression, Decision Tree models, three stemming techniques, and four feature extraction methods enhance the understanding of public sentiment surrounding COVID-19 vaccination campaigns. Our Logistic Regression model achieved 74.95% accuracy.

Findings reveal a predominantly positive sentiment, particularly happiness, among Saudi citizens. This research contributes valuable insights for healthcare communication, public sentiment monitoring, and decision-making while suggesting future directions for improving model performance and exploring broader linguistic and dialectal applications.

Key Word: *Data mining, natural language processing, sentiment analysis, emotion analysis, machine learning, support vector machine, logistic regression, decision tree, covid-19*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Microblogging has gained significant popularity in the last decade as a means of communication and information sharing, mainly through social media platforms, since it is accessible anywhere and anytime. One of the most prominent microblogging platforms is Twitter, which allows users to post short text called "Tweet" in real-time to express their thoughts, opinions, and emotions. Twitter has become a widely used platform globally for various purposes [71], including news updates, social interactions, and discussions on various topics. The abundance of user-generated content on Twitter presents unique opportunities for studying human behavior, emotions, and opinions, making it an exciting and relevant data source for researchers in various fields. Twitter has become one of the top social media platforms in Saudi Arabia, with more than 15.5 million active users [24], which

makes it a data treasure for researchers in the region to gauge people's opinions on a specific topic or service provided.

Natural Language Processing (NLP) techniques, including Sentiment Analysis (SA) and Emotion Analysis (EA), have emerged as powerful techniques for analyzing and understanding the vast amount of textual data generated in any format. Sentiment analysis involves determining a text's sentiment or emotional tone. In contrast, emotion analysis goes beyond sentiment analysis and aims to identify specific emotions expressed in a text, such as happiness, anger, or sadness. These techniques have been widely applied in various domains, including social media analytics [36], customer feedback analysis [46], and market research, to gain insights into people's opinions, emotions, and behaviors.

1.2 Research Gap

The research gap can be attributed to several significant factors within the context of emotion analysis in the Saudi dialect. Firstly, Arabic, known for its rich morphology and many dialects [16], poses inherent complexities for emotion analysis, especially in social media where informal language, including dialects and slang, prevails. This linguistic diversity adds a layer of difficulty in accurately interpreting emotions, necessitating a deeper understanding of the cultural context [24]. Secondly, the lack of a multi-emotion class Saudi dialect labeled tweets corpus, essential for training and validating emotion analysis models, has further contributed to this research gap. Thirdly, diacritical marks (Tashkeel), which signify short vowels and phonetic

features, introduce ambiguity as the same word can hold different meanings in various contexts [16]. Lastly, the Saudi dialect's deviation from the writing norms of Modern Standard Arabic (MSA) [7] exacerbates these challenges. Consequently, these combined linguistic and cultural intricacies have left a noticeable gap in research regarding applying machine learning algorithms for emotion analysis in the Saudi dialect, particularly in the context of COVID-19 vaccination discussions in Saudi Arabia.

Therefore, the main objective of our work is to fill this research gap by developing a machine-learning model that can accurately classify Saudi dialect tweets into seven emotion categories: happiness, fear, disgust, anger, surprise, optimism, and sadness. By achieving this aim, our research will contribute to the advancement of emotion analysis in the Arabic language, specifically in the context of the Saudi dialect, and provide valuable insights into the emotions expressed towards COVID-19 vaccinations in Saudi Arabia. The results will have practical applications in informing public health policies and communication strategies related to vaccination campaigns in the country.

1.3 Problem Statement

The lack of research applying Emotion Analysis (EA) to the Arabic language, specifically in dialects like the Saudi dialect, poses challenges due to its unique linguistic properties, lack of standardization, and absence of a prominent Saudi-labeled corpus for EA. Existing studies have provided limited insights using narrow

classifications, making a more comprehensive range of classifications necessary for more valuable and effective results. Embracing EA offers a more comprehensive understanding of sentiments beyond binary or ternary classifications, particularly in contexts like COVID-19 vaccination campaigns.

1.4 Motivations and Objectives

This section will discuss the motivations and objectives behind our research. As we have all seen, the COVID-19 pandemic has significantly impacted societies worldwide, and vaccination campaigns have become a vital component of global efforts to contain the spread of the virus. Social media platforms like Twitter have been essential in shaping public opinions and emotions toward COVID-19 vaccinations. Therefore, understanding and analyzing public sentiments and emotions toward vaccination campaigns has become increasingly important.

1.4.1 Motivations

In this context, this research involves two primary motivations. The first concentrates on implementing an EA model in the Saudi dialect, explicitly assuming that the EA will majorly influence various fields such as business, healthcare, education, government, and technology. The second is to understand better the general attitudes towards COVID-19 vaccination campaigns held in Saudi Arabia by producing graphical statistics extracted from the textual data.

1.4.2 Objectives

Our main objective is to produce a machine-learning emotion analysis model that can effectively classify the Saudi dialect tweets related to COVID-19 vaccination in Saudi Arabia into seven emotion categories. To achieve this, we have to complete the following sub-objectives:

1. Provide a machine-learning model of Emotion Analysis to classify emotions and feelings expressed in Tweets in the Saudi dialect.
2. Improve the accuracy of the results of the current studies that deal with EA in Saudi dialect using the Machine Learning approach.
3. Determine the general attitudes toward COVID-19 vaccination campaigns in Saudi Arabia.
4. Evaluate the accuracy of the results of the different algorithms used in the proposed model through a comprehensive experiment.
5. Produce a Saudi dialect labeled corpus in the healthcare and COVID-19 vaccination domain.

1.5 Expected Findings

The main expected output of this study is a machine-learning model that can classify Saudi dialect tweets into seven emotion categories. In addition, a published Saudi dialect labeled corpus is to be used in emotion analysis in related domains such

as healthcare, politics, and government. We will also visually highlight Saudi Arabian citizens' general attitudes about COVID-19 vaccination campaigns. Also, the detailed statistics about each emotion category and their corresponding text data should help the concerned parties make calculated, accurate, and appropriate future decisions. Furthermore, other relevant studies can utilize the model and the corpus.

1.6 Outline

In this section, we will provide an overview of the structure and content of this thesis.

Chapter Two, titled "Background and Literature Review" explores the relevant background information related to data mining, sentiment analysis, emotion analysis, and machine learning. Additionally, it discusses the current and previous research related to this topic.

Chapter Three, titled "Material and Method" details our research methodology, including the rationale behind our approach, the detailed design of our model, and the tools and technologies employed in our study. Moreover, it outlines the dataset collection process, ensuring transparency in our research approach. Then it describes the practical execution of our research. It covers essential stages such as data annotation, preprocessing, feature extraction and selection, and addressing data imbalance.

Chapter Four, titled "Implementation" shows the classification and training tasks of machine learning models.

Chapter Five, titled "Results and Discussion" is devoted to our findings and their comprehensive evaluation. It elucidates the evaluation methods, including confusion matrices and class distribution considerations, and introduces performance metrics such as accuracy, precision, recall, and F1-score. Furthermore, this chapter encompasses both the results of our experiments and in-depth discussions of our findings.

Chapter Six, titled "Conclusion, Limitations, and Future Work" is the culmination of our research. It provides a comprehensive summary of our findings, outlines the research limitations we encountered, and suggests potential avenues for future research.

Finally, our comprehensive thesis concludes with a section on references, ensuring due credit is given to all pertinent sources and prior works in sentiment and emotion analysis.

Chapter 2

Background and Literature Review

2.1 Background Information

2.1.1 Data Mining (DM)

It is essential to understand the process of discovering patterns and knowledge from extensive datasets. Data mining (DM) involves extracting valuable patterns and knowledge from vast datasets. DM functionalities can be classified into two main categories: predictive and descriptive. Predictive functionality focuses on constructing models that predict unknown or future values based on available data. On the other hand, explanatory functionality aims to discover new information that enhances the understanding and description of existing data patterns and relationships [23]. DM utilizes techniques to uncover meaningful insights [31], as shown in Figure 2.1, including the following techniques:

- **Classification:** used to categorize data collection into distinct groups or classes, thereby enabling accurate prediction and analysis, particularly in large datasets [50].
- **Regression:** employed to predict missing or unavailable numerical data values instead of discrete class labels [31].
- **Clustering:** groups similar data into clusters.
- **Outlier Analysis:** discovers rules that accurately predict robust sequential relationships among various events.
- **Sequential Pattern Discovery:** Identifies patterns that can effectively forecast significant sequential relationships among diverse events.
- **Association Rule Mining:** uses constraints and interestingness measures to ensure the completeness of mining [31].

Extracting valuable insights from raw data follows a sequential progression of steps. This extraction includes understanding the application domain thoroughly [49], preparing data, modeling using suitable techniques, evaluating results [15], and deploying knowledge for practical applications. These steps form a cohesive framework to extract valuable insights from data. This process is visually represented in Figure 2.2.

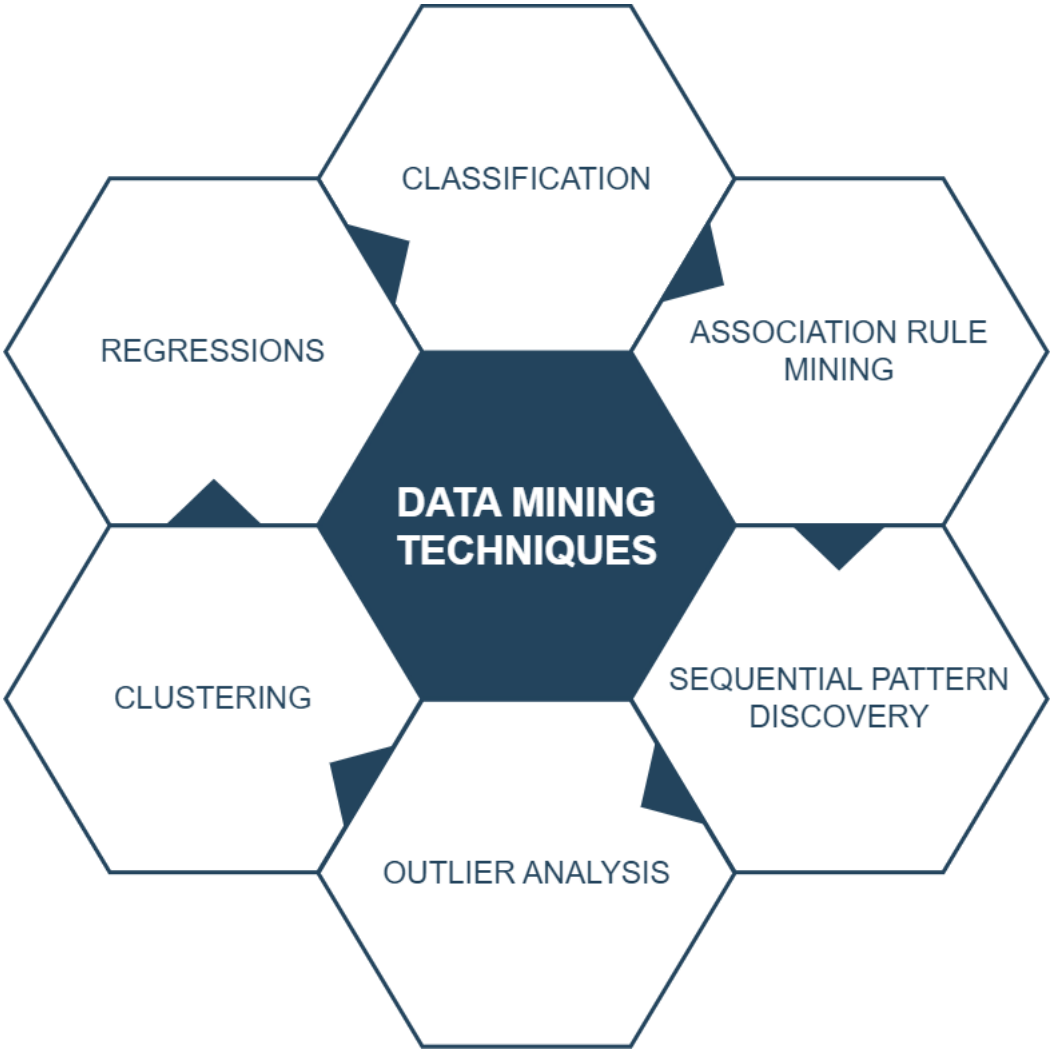

Figure 2.1: Data mining techniques

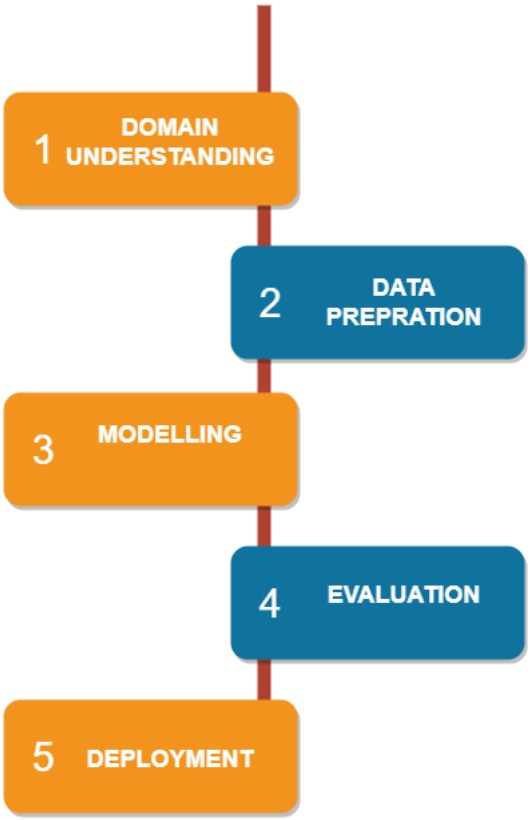

Figure 2.2: Data mining steps

2.1.2 Text Mining (TM)

Text mining is a particular type of DM [35]. Text mining is a specialized field that addresses extracting valuable information and knowledge from unstructured text data [33]. Unstructured texts contain vast amounts of information and cannot be directly processed by machines as they perceive the text as mere sequences of characters. Consequently, specific preprocessing methods and algorithms are necessary to uncover meaningful patterns. Text mining encompasses a range of analysis tasks, including classification, clustering, and association [35]. It is an interdisciplinary field at the intersection of information retrieval, machine learning, statistics, computational linguistics, and data mining. TM techniques apply to various text types, including news articles and academic publications [30], social media text, customer reviews, email and chat transcripts, scientific literature, and legal documents. For instance, text mining can extract information and summarize news articles, analyze sentiment and trends in a social media text, gain insights from customer reviews, understand customer inquiries from email and chat transcripts, and facilitate scientific research by analyzing published literature.

2.1.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence (AI) that focuses on analyzing and comprehending textual data, including both written text and spoken words [34]. By integrating various disciplines, such as computational linguistics, statistical analysis, machine learning, and

deep learning models, NLP enables computers to process and interpret human language effectively. One of the critical challenges in NLP is text classification or text categorization (TC), which involves assigning labels to different text areas, including queries, documents, sentences, and paragraphs. TC has diverse applications, including sentiment analysis, question-answering, news categorization, and user intent-based classification. Text data can be sourced from various channels, such as emails, chats, social media, web data, customer reviews, and feedback [40]. Emotion analysis and sentiment analysis are two of the essential techniques in NLP, that allow researchers to gain deep insights into the emotions and attitudes expressed in text, contributing to a better understanding of human behavior and decision-making processes.

2.1.4 Sentiment Analysis (SA)

Sentiment Analysis (SA), also referred to as opinion mining [58], is a subfield of natural language processing (NLP) that focuses on discerning the sentiment conveyed in textual data. SA aims to analyze and interpret subjective information, including opinions and attitudes expressed by individuals or groups in written text. It primarily involves determining the polarity of the text, which can be binary (positive or negative) in Binary SA (BSA) or further categorized into positive, negative, and neutral in Ternary SA (TSA) [2] [8] [4]. Using computational techniques and machine learning algorithms, SA enables the automatic recognition of sentiment at various levels, ranging from entire documents to individual words [78] [16].

The significance of SA lies in its capacity to extract valuable insights from large

volumes of text data, such as customer reviews, social media posts, online blog comments, and survey responses. By understanding the sentiment behind these texts, businesses and organizations can better understand public opinion, customer satisfaction, brand perception, and market trends.

SA techniques involve several steps shown in Figure 2.3, including

- Text preprocessing: The cleaning and normalizing of the text data by removing noise and irrelevant information.
- Feature extraction: Represents the text in a numerical format that is suitable for machine learning algorithms.
- Sentiment classification: Employs algorithms like Naive Bayes, Support Vector Machines (SVM), or deep learning models to classify text into positive, negative, or neutral sentiment categories.
- Result interpretation: Analyzes sentiment analysis results to extract meaningful insights and support informed decision-making.

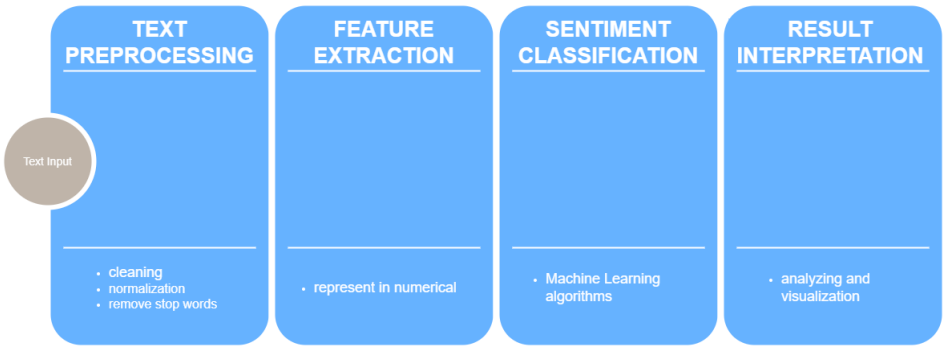

Figure 2.3: Sentiment analysis tasks

2.1.5 Emotion Analysis (EA)

In the context of textual data, EA involves extracting and analyzing emotional states expressed by people in texts [51]. EA is a field within NLP that uses computational techniques and algorithms to identify, classify, and analyze the emotional content expressed in written text. Emotion analysis aims to go beyond simply understanding the literal meaning of words and sentences and uncover individuals' or groups' underlying emotional states and attitudes [8]. It recognizes that human language is a means of conveying information and a powerful medium for expressing emotions, opinions, and subjective experiences. However, emotion analysis is complex due to inherent subjectivity and ambiguity. It requires robust algorithms, domain-specific knowledge, and continuous evaluation and refinement to ensure accurate and reliable results. Researchers and practitioners in the field of NLP continue to explore and develop new techniques to improve the accuracy and granularity of emotion analysis.

SA and EA Overlapping

Sentiment analysis and emotion analysis are effective areas of research aimed to auto-detect and recognize the sentiment expressed in an entire document, sentence, phrase, or word [16] [78]. Although they are often considered interchangeable terms [59], there is a core difference in the outputs of each. The main purpose of SA is to find the polarity expressed in a text by distinguishing between positive, negative, and neutral. At the same time, EA is concerned with detecting more emotion categories such as happiness, anger, sadness, and fear. Utilizing the power of EA allows the analysis to extract more accurate and detailed results that are more suitable to the

field in which it is applied. Since the EA can understand individuals' attitudes and perspectives towards a certain topic, many researchers have been motivated by this capability to utilize it in many real-life areas. For example, it is used in politics to understand public opinion towards a specific election campaign for instance and hence establish their campaigning strategy [48], in brand management and business marketing to measure the opinion of the target audience on a particular product or service or in healthcare where it can be used to predict outbreaks and epidemics [69] [6].

2.1.6 Analysis Approaches

Machine Learning (ML)

The ML approach uses algorithms to quantify and analyze text based on the extracted feature representation [7]. ML algorithms are trained on labeled data, enabling them to learn patterns and relationships within the data. These algorithms can then make predictions or classify new, unseen text based on the learned patterns. In text analysis, ML algorithms can be applied to various tasks, such as sentiment analysis, emotion analysis, and topic classification. ML-based text analysis offers flexibility and adaptability, as it can handle different types of text data and learn from patterns in the data.

Lexicon-based Methods

Another approach to text analysis is the lexicon-based method, which relies on predefined dictionaries or lexicons containing related terms with corresponding sentiment scores. This method assigns sentiment scores to text based on the presence of certain words or phrases from the lexicon. Lexicon-based methods do not involve machine learning techniques but rely on predefined rules and dictionaries to evaluate the text's sentiment or other linguistic features. While lexicon-based methods can be straightforward to implement and interpret, they are limited by the coverage and accuracy of the lexicon used [13] [78] [9].

2.1.7 Machine Learning Algorithms

ML algorithms can be categorized into different types, including supervised, unsupervised, and reinforcement learning, as shown in Figure 2.4. In supervised learning, algorithms are trained on labeled data, where the input data is associated with corresponding output labels or target values. The algorithm learns to map the input data to the correct output based on the provided labels. Unsupervised learning deals with unlabeled data, where the algorithm learns patterns and structures in the data without explicit guidance. Reinforcement learning involves an agent learning to make decisions in an environment to maximize a reward signal [43] [29]. Supervised learning offers versatile solutions for diverse problem domains, including

- **Classification:** categorizing data into a set of known classes. This enables the

algorithm to assign labels to new, unseen data based on the patterns learned from the labeled training examples.

- **Regression:** predicts continuous values based on a given set of input features. For instance, it can be used to estimate the price of a product by considering its factors.

- **Ranking:** ordering data points according to some criteria. For example, ranking search results according to their relevance to a query. This is commonly seen in search engines, where search results are ranked according to their relevance to a user's query [25] [42].

Supervised learning algorithms include a variety of techniques such as Support Vector Machines, Logistic Regression, Decision Trees, Random Forests, Neural Networks, and many more. Each algorithm has its strengths, and the choice of algorithm depends on factors such as the data's nature, the problem's complexity, and the desired performance.

Support Vector Machine (SVM)

Support Vector Machines (SVM) is a widely used supervised learning algorithm in machine learning. It was introduced initially by Vapnik as a kernel-based model for classification and regression tasks. SVMs have gained significant popularity and have become one of the most widely used classification methods in recent years, primarily due to their strong theoretical foundations. While SVMs were initially designed for binary classification, they have been successfully extended to handle

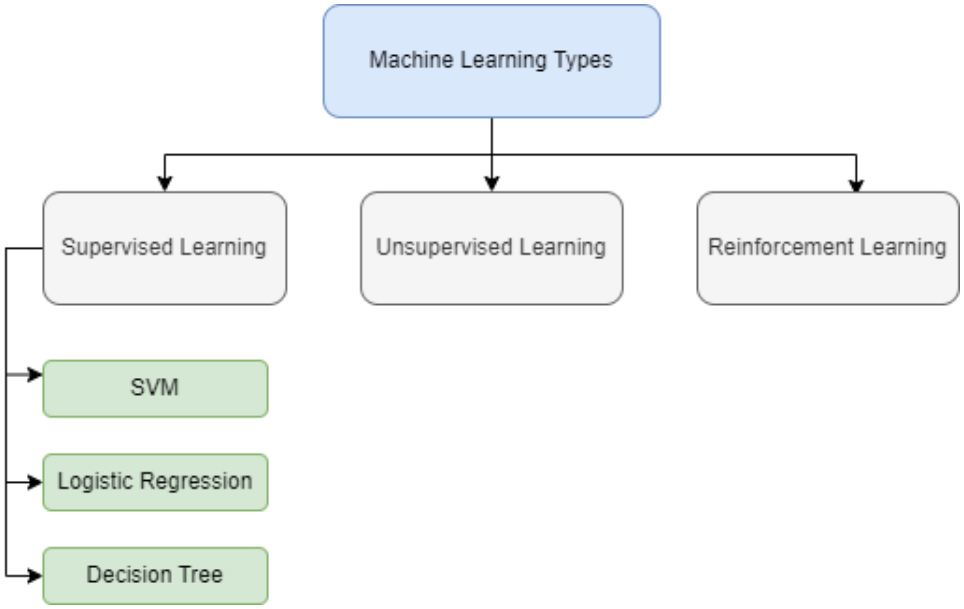

Figure 2.4: ML main types

multi-class classification tasks [20] [52].

SVM aims to find the best hyperplane that separates data points belonging to different classes with the most significant possible margin. Initially developed for binary classification, SVM has been extended to handle multi-class classification by computing the hyperplane between each class and the rest through various techniques, such as one-vs-rest and one-vs-one approaches. In the one-vs-rest approach, a separate SVM model is trained for each class, considering it as the positive class and the rest as the negative class; subsequently, the SVM algorithm selects the most probable classifier as the final result, as shown in Figure 2.5. In the one-on-one approach, SVM models are trained by comparing each class against every other class using multiple classifiers, and the outcome is made based on a voting scheme with the most probable class [52]. One of the critical strengths of SVM is its ability

to handle both linearly separable and non-linearly separable data. The hyperplane in SVM is defined by an equation that separates the data points in the feature space. For linearly separable data, the equation of the hyperplane can be represented as form 2.1 [17]:

$$g(x) = w^T x + b \tag{2.1}$$

Here, w is the weight vector perpendicular to the hyperplane, x is the input data vector, and b is the bias term. The sign of the expression $w^T x + b$ determines which side of the hyperplane a data point lies on. Points with a positive value are classified as one class, while points with a negative value are classified as the other.

The norm of a vector x, denoted as $||x||$, represents its length. For a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the Euclidean norm formula calculates the norm as form 2.2 [27]:

$$||X|| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \tag{2.2}$$

This formula computes the square root of the sum of the squares of each component of the vector, measuring the vector's magnitude or distance from the origin in n-dimensional space.

SVM aims to find the optimal hyperplane that maximizes the margin between the support vectors (data points closest to the hyperplane). This optimization problem is formulated as a convex quadratic programming problem, where the objective is to minimize the norm of the weight vector $||w||$ while satisfying the constraint that

all data points are correctly classified according to their labels, as shown in 2.3 form [27]:

$$\min_{w,b} \frac{1}{2}||w||^2 \tag{2.3}$$

To handle non-linearly separable data, SVM utilizes the kernel trick. The kernel function allows SVM to implicitly transform the input data into a higher-dimensional feature space, where linear separation becomes possible. The transformed data is then linearly separable, and the hyperplane is determined in this new feature space. The choice of the kernel function is crucial, as it determines the mapping of the data and affects the performance of the SVM algorithm. Using different kernel functions, SVM can capture complex patterns and relationships in the data. Commonly used kernel functions include the linear kernel, polynomial kernel, radial basis function (RBF) kernel, and HyperTangent [44]. SVM is well-suited for text classification due to its generalization capabilities, adherence to the Structural Risk Minimization principle, ability to incorporate prior knowledge, and superior performance compared to other methods such as k-nearest-neighbors (kNN) [70].

Logistic Regression (LR)

Logistic Regression is another widely used supervised learning algorithm in machine learning. It is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. Text classification is used to recognize a vector containing variables, evaluate the coefficients for each input variable, and predict the text class as a word vector. The logistic regression model measures the statistical significance of each independent variable concerning probability and is

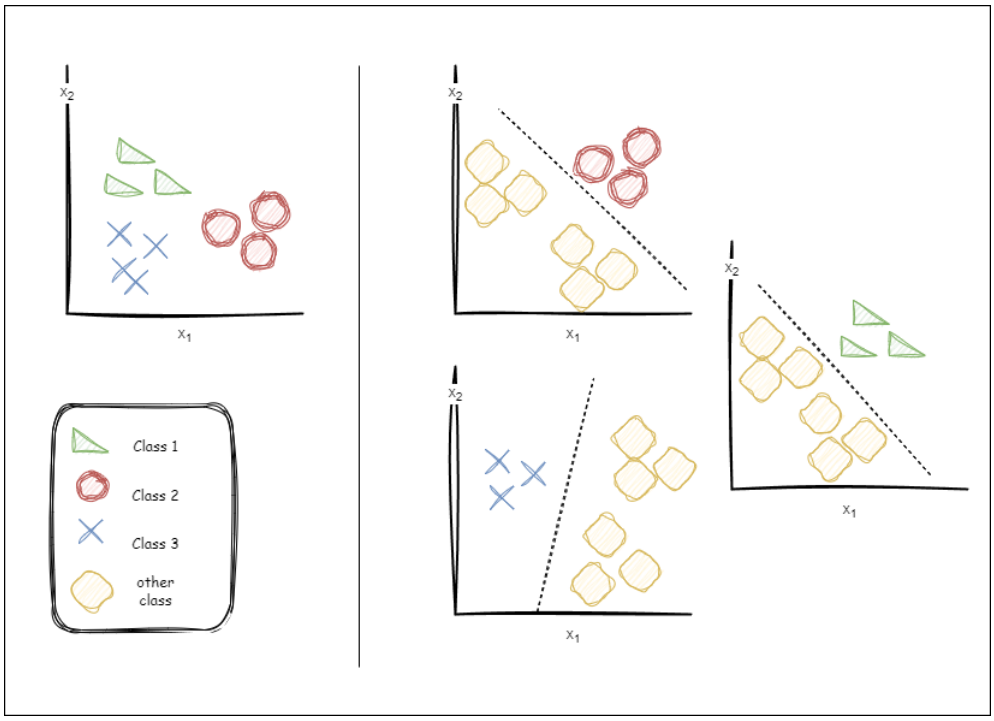

Figure 2.5: SVM multi-class classification -one-vs-rest approach

a powerful way of modeling binomial outcomes. It is commonly used in tackling text categorization problems and has advantages over other algorithms, such as computing the probability value rather than calculating a score [68] [26]. It is called "logistic" because it uses the logistic function, also known as the sigmoid function, to model the relationship between the variables and the probability of the outcome occurring. The logistic function ensures that the predicted probabilities are always between 0 and 1. The equation of the logistic regression model as in 2.5 and 2.4:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(w^T X + b)}} \tag{2.4}$$

$$P(Y = 0|X) = 1 - P(Y = 1|X) \tag{2.5}$$

where *P(y=1|x)* is the probability of the positive class given the input features *x, w* is the weight vector, *b* is the bias term, and *e* is the base of the natural logarithm. To train a logistic regression model, it needs the parameters (weights) *w* and *b* are estimated using a process called maximum likelihood estimation. The objective is to find the values of *w* and *b* that maximize the likelihood of the observed data. Logistic regression offers several advantages. It is computationally efficient, easy to implement, and provides interpretable results regarding the estimated coefficients. It can handle both numerical and categorical input features through appropriate encoding techniques. In addition to binary classification, logistic regression can be extended to handle multi-class classification using various strategies, such as one-vs-rest or softmax regression. In the one-vs-rest approach, a separate logistic regression model is trained for each class, treating it as the positive class and the rest as the negative class. The final prediction is based on the highest probability

obtained among all the models [68].

Decision Tree (DT)

A decision tree is a machine-learning algorithm for classification and regression tasks. It operates by recursively partitioning the data based on the values of input features, ultimately leading to a decision regarding the target variable. The algorithm constructs a tree-like structure representing a sequence of decisions and their potential consequences. Each internal node of the tree corresponds to a test on a specific attribute, while each branch represents the outcome of the test. The tree's leaf nodes correspond to class labels or numerical values, as shown in Figure 2.6 [21].

Decision trees are popular in machine learning due to their interpretability, simplicity, and ability to handle various data types, including categorical and numerical variables [53]. They find applications in various domains, including image processing, clinical practice, and financial analysis. Their inherent structure allows for an intuitive understanding of the decision-making process, making them valuable tools for extracting insights from data.

2.1.8 COVID-19 and Vaccination

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has profoundly impacted societies worldwide [19]. It was first identified when the initial case was reported in Wuhan, China, on December 19, 2019 [77]. The World Health

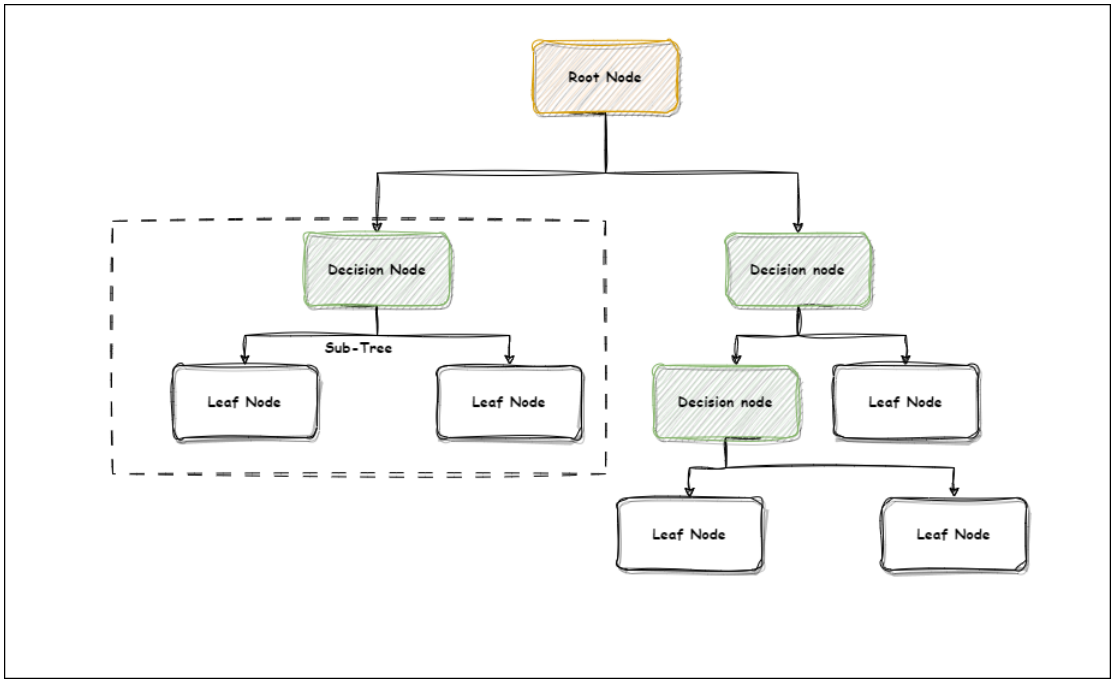

Figure 2.6: Representation of decision tree

Organization (WHO) officially declared the global COVID-19 pandemic on March 11, 2020 [76]. This declaration marked a turning point in the international response to the virus, leading to widespread public health measures to curb its spread.

Saudi Arabia recorded its first confirmed case of COVID-19 on March 2, 2020 [37] [61] [77,78]. The country swiftly implemented various measures to combat the virus's transmission, including lockdowns and travel restrictions.

The introduction of COVID-19 vaccines marked a pivotal moment in the fight against the pandemic. Saudi Arabia approved the Pfizer-BioNTech vaccine on December 10, 2020 [63]. Registration for the vaccine in Saudi Arabia began on December 15, 2020 [62]. As vaccination efforts progressed, restrictions evolved, with announcements such as allowing only vaccinated individuals to enter certain buildings starting from

August 1, 2021.

The vaccine rollout continued to advance, with the commencement of the second vaccine dose administration on June 23, 2021 [64]. As the situation improved, Saudi Arabia took steps to return to normalcy, including lifting many precautionary measures on May 3, 2022 [65].

2.2 Literature Review

The sentiment analysis and emotion analysis fields are highly concerned with understanding and extracting opinions from a text. The analysis process includes Natural Language Processing (NLP) techniques [56] and computational linguistics [78]. According to Assiri et al. [16] and Liu [41], sentiment analysis can be applied by adopting Machine Learning (ML) algorithms, a lexicon-based approach, or using a hybrid approach as in some research [55]. In our literature review, we only reviewed research that has adopted SA or EA for the Arabic language and Saudi dialect specifically; other languages are beyond the scope of this research. We have divided the literature review into Sentiment Analysis in the Arabic Language and Emotion Analysis.

2.2.1 Sentiment Analysis

Several papers dealt with sentiment analysis in Arabic, either using Modern Standard Arabic (MSA) or different Arabic dialects. We tried to cover the most recent papers

on our topic that adopted ML, Lexicon-based, and hybrid approaches.

Using Machine Learning

A contribution has been made in the health domain to gain better control of the COVID-19 pandemic, Aljameel et al. [12] developed a machine-learning model to measure individuals' awareness of preventive measures during quarantine in Saudi Arabia. Their model depends on a dataset collected during the curfew period in Saudi Arabia and then processed built on three machine learning classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbors (KNN). To improve the classification accuracy in the Arabic dataset related to the awareness of the COVID-19 epidemic, they aimed to find a suitable combination of feature extraction-classifier; thus, they used N-Gram and Term Frequency-Inverse Documents Frequency (TF-IDF) NLP models. Results showed that combining TF-IDF with the SVM classifier achieved the highest accuracy of 85%. However, although the study was for Saudi Arabia regions, it was not focused on the Saudi dialect. In another study conducted in the Entertainment field in Saudi Arabia, Al Sari et al. [60] applied five machine learning algorithms: Multilayer Perception (MLP), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Voting. Their main goal was to understand better SA's impressions of the first cruise trip in Saudi Arabia. Thus, they collected their dataset from three social networks (Twitter, Instagram, and Snapchat). The results showed that applying NB and MLP algorithms in the Twitter dataset achieved 90% accuracy. However, the datasets were collected within two months and did not focus only on the Saudi dialects;

the study also focused on studying the polarity of the text (positive, negative, and neutral) only. Another study conducted by Alhuri et al. [11] achieved an F1-score of 81% using a Gated Recurrent Unit (GRU), which is part of the Recurrent Neural Network (RNN) of ML. They created a model capable of getting insights into public reactions from Arabic tweets towards the COVID-19 pandemic. However, it only used two classifications (positive or negative) with one evaluation technique and was also not focused on the Saudi dialect. Because of the low performance of the well-known ML classifiers when compared to the Deep Learning (DL) algorithms in SA, Alahmary et al. [7] applied two DL algorithms: Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) to the Saudi Dialect Twitter Corpus (SDTC) collected from Twitter in different domains. They achieved an accuracy of 94% using the Bi-LSTM, which outperformed both the DL LSTM, which achieved 92% accuracy and the ML SVM, which achieved 86.4%. However, improving the preprocessing steps could improve the accuracy; the classification was dependent only on the three label classes of the text polarity. AlYami and AlZaidy [14] conducted a supervised machine-learning analysis on Arabic text retrieved from Twitter in Saudi and Egyptian dialects. They preprocessed the data in three steps to allow Arabic dialect identification (ADI), then trained the ML model to implement four ML algorithms (SVM, RF, NB, and LR). They achieved 87% and 86% accuracy using LR and SVM models in Egyptian dialects, respectively. However, the dataset was limited to 600 tweets, not annotated manually, no intensive preprocessing tasks were applied, and aimed only at identifying the dialect.

Using Lexicon-based

To address the lack of lexicon-based algorithms and handle Saudi dialect rather than MSA or other Arabic dialects, Assiri et al. [16] proposed a domain-independent algorithm based on a lexicon-based approach for handling Saudi dialect exclusively. In an attempt to fill this gap, along with the fact that there were no labeled social network data existing for the Saudi dialect, the authors created an annotated set of data with a focus on negotiation and supplication, in addition to creating a large Saudi lexicon before developing the proposed weighted lexicon-based algorithm for handling Saudi dialect tweets. However, the proposed algorithm was evaluated against a non-Saudi dataset, and the classification was based only on the polarity of the text (positive, negative, and neutral). For the same purpose, Al-Thubaity, Alharbi, Alqahtani, and Aljandal [4] manually extracted Saudi tweets to create a Saudi dialect sentiment lexicon named "SauDiSenti." This lexicon contained 4431 words and phrases annotated manually by Saudi persons. This lexicon contained a combination of the Saudi dialect and MSA. Because of the lack of a Saudi lexicon to compare their lexicon against, the authors compared their lexicon against a larger Arabic dictionary [5]. However, the results may not be accurate since their testing dataset was generated by the same annotators who annotated the "SauDiSenti." Also, no accuracy or pre-processing steps were provided for a considerably small lexicon. In addition, they focused only on the polarity of (positive, negative, and neutral) text. For the same purpose of increasing the accuracy performance of SA in the Saudi dialect and since the formal-dependent lexicons are inefficient in SA as they cannot capture the colloquial dialects, Al-Ghaith [3] aimed to achieve that

goal differently by concentrating on employing the preprocessing tasks on the Saudi dialect lexicon itself rather than applying them on the dataset as the case as in most of the studies. They also produced two algorithms for dealing with the prefixes and suffixes of words and achieved an accuracy of 81%. However, they depended on an English lexicon to create their original lexicon.

Using Hybrid Approach

Very few published studies have utilized the Hybrid approach of SA in Arabic - where semantic orientation and ML techniques are combined. Aldayel and Azmi [8] used this approach to improve the F-measure score; they achieved an overall F-measure and accuracy of 84% and 84.01%, respectively. Alhumoud et al. [10] used the same approach of combining two ML algorithms. They applied the SA on 3000 Saudi dialect tweets to prove the efficiency of the hybrid learning approach compared to solo ML.

2.2.2 Emotion Analysis

EA can be described as recognizing distinct human emotions in contrast to Sentiment Analysis, which identifies whether data is positive, negative, or neutral [74]. Because of the lack of a labeled corpus for Saudi dialect that can be used for classifying emotions and polarity behind a text, Al-Thubaity, Alharbi, Alqahtani and Aljandal [4] introduced the Saudi Dialect Twitter Corpus (SDTC) that contains 5400 tweets of SD and MSA classified for sentiment analysis and emotion analysis annotated by

three raters based on their polarity (positive, negative, and neutral) for the sentiment, and based on Ekman's basic emotions (anger, fear, disgust, sadness, happiness, surprise, no emotion and not sure) for the emotion analysis. However, no ML or lexicon-based approaches have been applied to this corpus to evaluate its efficiency in this study. Another study by A. AlFutamani and H. Al-Baity [9] was the first study in the field of EA in Arabic with a focus on Saudi dialect in Arabic Textual content retrieved from Twitter, mainly in Saudi-based tweets. They built a system that can detect the underlying emotions of Saudi dialect tweets to classify them based on seven emotion categories (happiness, fear, disgust, anger, surprise, optimism, and sadness). They used two ML algorithms (SVM and MNB), achieving 73.39% accuracy in the SVM approach. However, they applied the analysis in different dataset domain sets, and limited preprocessing was applied to the dataset.

Although the positive and negative classifications of SA can be very suitable to be applied in some domains, they may not be suitable for many other domains needing greater detail. Table 2.1 summarizes the reviewed papers.

Table 2.1: Summary of the Literature Review

Cite	Approach	Algorithm(s)	Results	Limitation/Opinion
46	ML	SVM, NB, and KNN	85% accuracy by SVM with TF-IDF	Not focused on Saudi dialect
47	ML	MLP, NB, SVM, RF, and Voting	90% accuracy using NB	- Short period of dataset - Not only in Saudi dialect - Positive, negative, and neutral classification only.
48	DL	GRU and RNN	81% F1-score	- Only two classes (positive, negative). - Only one evaluation method. - Not focused on the Saudi dialect.
4	DL and ML	LSTM, Bi-LSTM, and SVM	94% accuracy using Bi-LSTM	Limited preprocessing tasks. - Three sentiment analysis classes only.
3	Lexicon	Not mentioned	Not mentioned	- The proposed algorithm was evaluated against the non-Saudi dataset. - Based on three sentiment analysis classifications only.
20	Lexicon	Not mentioned	Not mentioned	- No Preprocessing tasks. - Considerably small dataset. - No evaluation tasks were applied.
51	Lexicon	New technique to deal with prefixes and suffixes of words	81% accuracy	- When created, the lexicon depends on the English Lexicon.
19	Hybrid	Not mentioned	84.01% accuracy	Three sentiment analysis classes only.
52	Hybrid	Two ML algorithms	Not mentioned	- No evaluation applied - Three sentiment analysis classifications only.
20	Lexicon	Not mentioned	Not mentioned	- Only created a Saudi dialect corpus for SA and EA - No ML or Lexicon-based algorithms are applied.
28	ML	SVM, and MNB	73.9% accuracy using SVM	Limited preprocessing, accuracy could be improved, and Not in Vaccination of COVID-19
49	ML	SVM, MNB, RF, NB, and LR	87% accuracy using LR and 86% using SVM	- limited dataset - focused on Arabic dialect identification - no evaluation explanation

Chapter 3

Material and Method

3.1 Rationale Reason

Our research addresses the problem of limited emotion analysis in the healthcare domain, specifically in the context of the Saudi dialect. This research aims to develop an ML model that accurately classifies Saudi dialect tweets into the basic emotions defined by Paul Ekman, focusing on vaccinations.

The importance of this problem lies in the significance of understanding public sentiment and opinions surrounding healthcare issues. By analyzing emotions expressed in social media posts, decision-makers, and policy developers can gain valuable insights that inform targeted interventions, effective communication strategies, and informed decision-making in healthcare. Although there are studies have focused on sentiment analysis in Arabic and the Saudi dialect [12] [3] [7] [16]. However, there is a notable research gap in emotion analysis, particularly in healthcare. Positive

and negative sentiment classifications may not provide the detail required in many contexts.

The existing knowledge needs more substantial efforts to develop an ML model specifically for classifying Saudi dialect tweets into the basic emotions in the discipline of vaccinations. Furthermore, there needs to be more available labeled corpora manually annotated with multiple emotion categories in healthcare in the Saudi dialect. Our research fills these gaps by developing an ML model tailored to the Saudi dialect and the healthcare domain. We also create a labeled corpus manually annotated with multiple emotion categories in healthcare. This comprehensive approach enables more detailed and accurate emotion analysis, advancing sentiment and emotion analysis in Saudi Arabia.

Furthermore, our research is a pioneering step for future studies aiming to enhance machine learning models in sentiment analysis. Researchers can build upon our findings and improve the accuracy and effectiveness of sentiment analysis in various domains by adopting our approach and utilizing the provided labeled corpus of public tweets.

3.2 Methodology Overview

To achieve our objectives, we have followed a series of stages for developing an ML model and creating a labeled corpus in the Saudi dialect for emotion analysis in healthcare. These stages include data collection, annotation, preprocessing, feature extraction, classification, and evaluation. Figure 3.1 shows the basic block diagram

of the proposed methodology.

In the data collection stage, relevant vaccinations-related tweets are gathered using specific keywords using Twitter API. The collected tweets are then manually annotated to label the expressed emotions by three Saudi natives, forming a labeled corpus for training and evaluation. Data preprocessing involves filtering irrelevant data, cleaning text from errors and inconsistencies, normalizing text, and tokenizing it for analysis. Next, we extract and select features that effectively capture the emotions expressed in the tweets using Bag-of-Words, N-Gram, and TF-IDF, which serve as inputs for the ML model. The classification stage involves developing and training the ML model using various algorithms (Support Vector Machines, Logistic Regression, and Decision Tree) to classify the tweets into seven emotions. We evaluate the model's performance using appropriate metrics and conduct experiments to validate our methodology. Unseen data is used to test the model, and accuracy, precision, recall, and F1-score are measured.

3.3 Detailed Design

In the detailed design of our methodology, we have carefully considered the techniques and procedures used in each stage to achieve our research objectives. Our design choices ensure the effectiveness and reliability of our approach to emotion analysis within the Saudi dialect, specifically in the healthcare domain.

In the data collection stage, we leverage the Twitter API with academic access and specify relevant keywords to retrieve vaccination-related tweets. This task allows

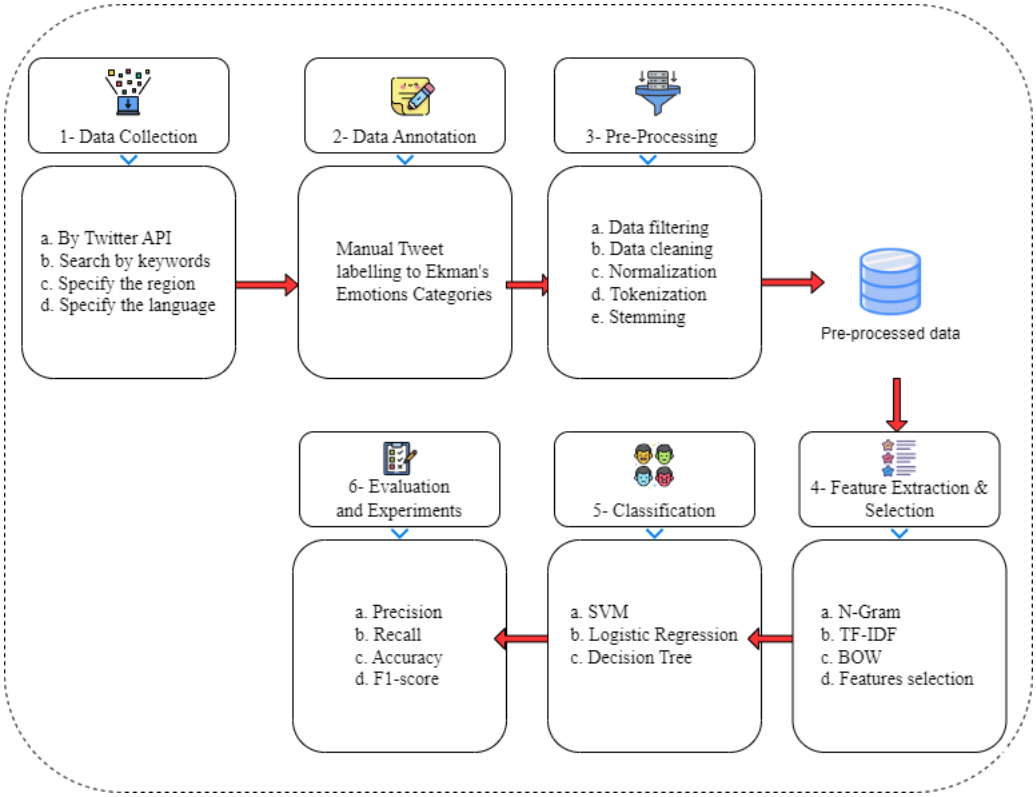

Figure 3.1: Basic block diagram of the proposed methodology

us to collect real-time data that captures a diverse range of tweet emotions. Python snippets consume the Twitter API, enabling multiple consecutive requests with query parameters such as keywords, start time, end time, region, and language.

During the annotation stage, we rely on the expertise of three Saudi natives fluent in the Saudi dialect. They manually label the collected tweets, considering the Saudi dialect's cultural nuances and linguistic intricacies. This meticulous annotation process ensures the creation of a high-quality labeled corpus, which is essential for training and evaluating our ML model. The tweets are classified into seven basic emotions (Happiness, Fear, Disgust, Anger, Surprise, Optimism, Sadness), with an additional eighth class for Neutral/ad/indefinable tweets.

In the data preprocessing stage, we employ various techniques to clean and prepare the data for analysis. Data filtering removes irrelevant or noisy tweets, focusing only on those directly related to vaccinations and healthcare. Data cleaning addresses spelling errors, punctuation inconsistencies, and other textual irregularities. Text normalization standardizes the text format, and tokenization splits the text into individual units for further analysis. Stemming techniques such as Kohlen Stemmer, Porter Stemmer, and SnowBall Stemmer are also applied.

Feature extraction and selection are crucial in capturing the expressed emotions. We utilize techniques like Bag-of-Words, N-Gram, and TF-IDF to extract relevant features from the preprocessed text. These features are carefully selected based on their ability to represent the emotional content of the tweets effectively. They serve as inputs for our ML model, enabling it to learn and make accurate predictions about the emotional state of the tweets.

The classification stage involves the development and training of our ML model. We consider various classification algorithms, including Support Vector Machines, Logistic Regression, and Decision Trees, to classify the tweets into the seven emotions. Training the ML model using the labeled corpus allows it to learn the patterns and relationships between the extracted features and the labeled emotions.

To evaluate the performance of our ML model, we employ appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Unseen data, separate from the training data, is used to assess the generalization capabilities of the model. This evaluation provides insights into the model's effectiveness in accurately classifying emotions in the Saudi dialect.

Following a systematic and rigorous design, our methodology addresses the research gaps in emotion analysis within the Saudi dialect, specifically in the healthcare domain. Our approach ensures the development of an accurate ML model and the creation of a reliable labeled corpus, providing valuable insights for healthcare decision-makers and policy developers.

3.4 Tools and Technologies Used

3.4.1 Twitter API (Academic Research Access)

The Twitter API for Academic Research is a specialized track of the Twitter API that provides academic researchers access to a wider range of features and functionality than the standard Twitter API. This API includes access to the full history of public

tweets, the ability to request a higher monthly Tweet cap, and the ability to filter tweets by language and request data in real-time [73]. The Twitter API is a crucial tool utilized in our data collection stage to gather relevant tweets for our research on emotion analysis in the Saudi dialect within the healthcare domain. With academic access to the Twitter API, we can retrieve real-time vaccination-related tweets, capturing a diverse range of emotions users express. This access provides a valuable data source for training our machine-learning model and conducting comprehensive emotion analysis. By leveraging the Twitter API, we can collect and analyze many tweets, enabling us to gain insights into public sentiment and emotional responses in healthcare discussions.

3.4.2 Python Programming Language and Visual Studio Code

Python programming language is the foundation for leveraging the Twitter API and collecting tweets from Twitter for our data collection stage. After installing Python, we utilized various libraries to facilitate the process. Notably, we employed Pandas and Requests libraries, among others, to make fetching requests to the Twitter API [54].

By leveraging Python and these libraries, we could specify parameters in each API request, such as keywords, start time, end time, and language, to retrieve relevant tweets related to our desired topic. The tweet data was saved in CSV format for further analysis and processing.

We utilized the Visual Studio Code editor to facilitate the development of our Python

code. Its intuitive interface and powerful features enabled us to write efficient and effective Python code for interacting with the Twitter API [45].

The choice of Python as our programming language and the utilization of relevant libraries such as Pandas and Requests allowed us to leverage the capabilities of the Twitter API, retrieve tweets in real-time, and efficiently store them for subsequent analysis. Python's versatility, extensive library support, and user-friendly syntax made it an ideal choice for our data collection process.

3.4.3 Google Sheets

Google Sheets is a web-based spreadsheet application provided by Google. It allows users to create, edit, and share spreadsheets online, providing a collaborative data organization and analysis platform. Our research used Google Sheets for various purposes throughout the data preprocessing and annotation stages [28].

One of the key uses of Google Sheets was to organize and arrange the collected tweets. We utilized its functions and features to perform preprocessing tasks such as cleaning, filtering, and removing duplicates. For example, we leveraged functions like REGEXREPLACE to clean the text by removing unwanted characters or patterns. Additionally, functions like VLOOKUP were employed to search and detect specific criteria in the dataset.

Google Sheets played a significant role in facilitating the manual annotation process. We shared copies of the dataset with the annotators, allowing them to manually annotate each tweet by selecting one of the predefined emotion classes. The Google

Sheets interface simplified this task by providing checkboxes corresponding to each emotion class, making the annotation process more efficient and user-friendly.

After completing the manual annotation, we merged the annotated sheets into one, consolidating the labeled data. This merged sheet allowed us to perform further analysis and calculations. For instance, we utilized Google Sheets' functions to count the tweets that matched each emotion class, providing valuable insights into the distribution of emotions in the dataset.

3.4.4 KNIME Analytics Platform

The KNIME Platform is a comprehensive open-source data analytics and integration tool. It played a pivotal role in our research implementation, serving as a versatile and powerful tool throughout various project stages, from initial data preprocessing to evaluating the ML model [38].

One of the key strengths of the KNIME Analytics Platform lies in its advanced data preprocessing and cleaning capabilities. We utilized this platform to perform further and in-depth data preprocessing tasks beyond what was achieved using Google Sheets. KNIME enabled us to apply a wide range of data transformations, filters, and cleaning techniques to ensure the quality and consistency of the dataset.

Furthermore, the platform facilitated data insights and exploration. We leveraged its intuitive interface, comprehensive visualizations, and statistical tools to understand the dataset better. This allowed us to identify patterns, trends, and outliers, enabling more informed decision-making throughout the research process. The platform also

played a vital role in feature extraction and selection. We leveraged its extensive collection of built-in nodes and components to extract relevant features from the preprocessed data. This enabled us to represent the emotional content of the tweets effectively, providing valuable inputs for the subsequent ML modeling phase.

KNIME Analytics Platform supported us in addressing class imbalance through data oversampling techniques. We utilized its capabilities to balance the distribution of emotion classes, enhancing the performance and generalizability of the ML model. Moreover, the platform enabled the application of various ML algorithms to train and evaluate the model. We utilized its intuitive data flow components to structure and organize the ML workflow, incorporating the necessary preprocessing steps, feature engineering, and model training. Additionally, we leveraged KNIME's evaluation capabilities to assess the performance of the ML model using appropriate metrics.

3.5 Dataset Collection

Data collection is a critical stage of our implementation process. We aim to gather a wide range of tweets related to vaccinations in the Saudi dialect. This dataset will be used to analyze the emotions users express and develop a machine-learning model that can accurately classify these emotions.

To achieve this, we leveraged the Twitter API for Academic Research using Python programming language, as mentioned in sections 3.4.1 and 3.4.2, to access real-time Twitter data. We retrieved tweets aligned with our research objectives using Python libraries (Tweepy [72], Pandas, and Requests).

We focused our data collection on a specific time range from December 15, 2020 (the date of the first announcement for registration of the first COVID-19 vaccination by the Saudi Ministry of Health [62]) to March 10, 2022 (after the decision to lift precautionary and preventive measures related to the pandemic by five days [65]). This time frame encompasses crucial periods related to COVID-19 vaccinations in Saudi Arabia.

To ensure the collection of relevant tweets, we specified specific Arabic keywords in each fetching request, including (COVID-19 vaccine, "لقاح كورونا" - Pfizer, فايزر - Second Dose, الجرعة الثانية - boosting dose, الجرعة التنشيطية - Oxford Vaccine لقاح اكسفورد ..etc.) and other related terms. These keywords targeted tweets specifically focused on the COVID-19 vaccination domain.

Furthermore, we utilized various parameters within the API requests to filter the tweets based on the Saudi region. This allowed us to target tweets primarily written in the Saudi dialect, capturing the language nuances and cultural context specific to Saudi Arabia. We also employed parameters to ensure that only Arabic tweets were collected and to turn off the collection of retweets. These parameters were applied to each fetching request, ensuring the dataset's quality and relevance. Figure 3.2 shows a snapshot of the fetching process.

We prioritized privacy, data security, and ethical considerations throughout the data collection. We adhered to Twitter's terms of service and privacy policy, ensuring compliance with ethical guidelines and data usage restrictions.

Our data collection efforts resulted in the successful retrieval of a total of 34,074 raw

```
Start Date:  2021-10-16T00:00:00.000Z
# of Tweets added from this response:  494
Total # of Tweets added:  9216
-------------------
-------------------
Token:  None
Endpoint Response Code: 200
Next Token:  b26v89c19zqg8o3fpdv7wclzzrx410mgu4exs33kgd6d9
Start Date:  2021-11-01T00:00:00.000Z
# of Tweets added from this response:  461
Total # of Tweets added:  9677
-------------------
-------------------
Token:  None
Endpoint Response Code: 200
Next Token:  b26v89c19zqg8o3fpdy5zm8mo028mzgmgun6xoyuh08l9
Start Date:  2021-11-16T00:00:00.000Z
# of Tweets added from this response:  471
Total # of Tweets added:  10148
-------------------
-------------------
Token:  None
Endpoint Response Code: 200
Next Token:  b26v89c19zqg8o3fpdy8i97qut86g4zpe2eap8tbi8h31
Start Date:  2021-12-01T00:00:00.000Z
# of Tweets added from this response:  466
Total # of Tweets added:  10614
-------------------
-------------------
Token:  None
Endpoint Response Code: 200
Next Token:  b26v89c19zqg8o3fpe17fu35iwq23s1vcljg92zq6v1bx
Start Date:  2021-12-16T00:00:00.000Z
# of Tweets added from this response:  477
Total # of Tweets added:  11091
-------------------
Total number of results:  11091
PS D:\KAU\6- 2022\CPIT- 699 Thesis\Python files\TweetScrapper>
```

Figure 3.2: Snapshot of Python code output during the fetching process

tweets related to vaccinations in Saudi Arabia. Each collected tweet is represented by a row in our dataset, containing the following properties:

- **Tweet:** The actual text of the tweet.
- **Author_ID:** The unique user ID of the Twitter account that posted the tweet.
- **Created_at:** The date and time when the tweet was posted.
- **Geo:** A numeric representation of the location or region associated with the user who posted the tweet.
- **ID:** The unique numeric ID assigned to the tweet.
- **Lang:** The primary language of the tweet.
- **Like_count:** The number of likes received by the tweet.
- **Quote_count:** The number of times the tweet has been quoted by other users in their tweets.
- **Reply_count:** The number of replies received by the tweet.
- **Retweet_count:** The number of retweets of the tweet.
- **Source:** The platform or operating system used to post the tweet.

Table 3.1 shows an instance of a collected tweet. The dataset was saved in comma-separated values *(.CSV)* format, providing a robust foundation for our subsequent stages of data annotation, preprocessing, feature extraction, and classification.

Table 3.1: A raw instance of the collected tweets

Tweet	@_doje_ @FBasme اتنهت حفلة كورونا!! وقطاع السوبرماركت
Author_ID	534798407
Created_at	2020-12-19 17:37:51+00:00
Geo	000799c66e428a87
ID	1.34035×10^{18}
Lang	ar
Like_count	0
Quote_count	0
Reply_count	0
Retweet_count	0
Source	Twitter for Android

3.6 Data Annotation

In the data annotation stage, we manually assign emotions to the collected tweets. However, before manually annotating, we conducted initial preprocessing steps to ensure the highest possible labeling accuracy without compromising the tweet's context or meaning. The preprocessing tasks performed before the annotation process include removing duplicate tweets, eliminating Twitter handles (@), removing URLs, filtering out English text, and removing emojis and non-emoji symbols. The details of each preprocessing task will be explained further in section 3.7.

During the annotation stage, the tweets were categorized into emotion classes aligned with the emotions defined by Paul Ekman: Happiness, Fear, Disgust, Anger, Sadness, and Surprise. These emotions encompass a wide range of human emotional responses and provide a comprehensive framework for capturing different sentiments expressed in the tweets. In addition, a neutral/spam class was included to annotate tweets that lacked any discernible emotional content or contained spam/advertise-

ments. This class allowed us to identify and filter out tweets that did not contribute to the emotional analysis. Also, an optimistic class was introduced as an additional emotion category to account for the Saudi dialect's unique characteristics and the dataset's specific context. This class aimed to capture tweets expressing optimism, hope, or positive outlooks concerning vaccinations. This resulted in eight emotion categories in the end. Table 3.2 provides an overview of the emotion labels used during the annotation stage.

A folder was created in Google Drive, containing three duplicate Google Sheets. Each Sheet included an identical set of 33,373 unique tweets, and alongside each tweet were columns of checkboxes representing the different emotion categories. These duplicate Sheets were then shared with individual annotators who firmly understood the Saudi dialect and its nuances. Figure 3.3 illustrates one of the shared sheets.

To maintain objectivity and ensure consistent annotation, detailed instructions were provided to the annotators. These instructions guided the annotators in assigning the appropriate emotion to each tweet in their assigned Sheet. The instructions emphasized the following key points:

- **Privacy:** The annotators were assured that no personal information would be collected during the annotation process, ensuring their anonymity and confidentiality.
- **Single Emotion Selection:** In cases where a tweet exhibited multiple emotions, the annotators were instructed to select the strongest or most dominant emotion from the provided emotion categories.

Table 3.2: The emotions used in the annotation stage

#	English Emotion Class	Arabic Emotion Class	Explanation
1	Happiness	سعادة	When the tweet expresses feelings of joy, happiness, or delight
2	Fear	خوف	When the tweet expresses feelings of fear
3	Disgust	اشمئزاز	when the tweet expresses disgust, disgust, or disgust with the tweeter
4	Anger	غضب	When the tweet expresses angry feelings
5	Surprise	تفاجؤ	When the tweet expresses feelings of surprise, astonishment, or wonder
6	Optimism	تفاؤل	When the tweet expresses feelings of optimism and a positive view of the future
7	Sadness	حزن	When the tweet expresses feelings of sadness, brokenness, grief, or depression
8	Neutral	محايد أو إعلان أو لايمكن تحديده	When the content of the tweet does not include any emotions or feelings that cannot be identified from other options or includes only advertising hashtags without any emotions

- **Accuracy:** The annotators were encouraged to make precise and accurate judgments in choosing the emotion that best represented the content of each tweet. This attention to detail would improve the overall quality of the emotion analysis results.
- **Time:** To manage the annotation process effectively, each annotator was given a maximum of two months to complete the annotation of tweets. An estimated daily average number of 530 tweets was provided to maintain a consistent workflow and prevent overload. This approach aimed to ensure accurate and reliable annotations while allowing annotators enough time to analyze and assign emotions to each tweet without feeling rushed.

By providing clear instructions, highlighting these important considerations, and clearly explaining the meaning of each emotion class, the annotation process was carried out systematically and reliably, ensuring the integrity of the collected data and subsequent analysis.

Upon completion of the annotation period, which took approximately two months, the results from all annotators were collected and merged into a consolidated dataset. The output of the annotation process shows that there are 23689 tweets are fully class-matched between the three raters out of the 33373 tweets, and 9684 are not fully matched between the three raters. Table 3.3 shows the counts of tweets annotated in each emotion class by each annotator. Table 3.4 shows examples of the annotated tweets by each rater.

The data annotation stage allowed us to assign specific emotions to each tweet, providing a labeled dataset that serves as the foundation for the subsequent stages

of our implementation process, such as data preprocessing, feature extraction, and classification.

#	التغريدة	سعادة	خوف	اشمئزاز	غضب	تفاجؤ	تفاؤل	حزن	محايد/اعلان/لا يمكن تحديده
1	(اسعد إنسانة اليوم) يا زين الكلام اللي على الطبيعة كانت شيخة الحربي اول مواطنة تتطعم بلقاح فايزر تعبر بفرحتها عن فرحتنا قبلها	☑	☐	☐	☐	☐	☐	☐	☐
2	السلام عليكم من فضلكم اين الاماكن ال اقدر اسوي فيها فحص كورونا المجاني قبل السفر لمصر ؟	☐	☐	☐	☐	☐	☐	☐	☐
3	لاكن هل ترى اي عوامل إيجابية مثل لقاح كورونا تمنع اي نزول	☐	☐	☐	☐	☐	☐	☐	☐
4	من جد ... وباء مؤامره لقاح مؤامره... شلون يعني ... اللي راحو مؤامره .... في حد يمرض ثم يموت ثم يقولون مؤامره... هاتو القاح اليوم قبل بكره ... كفايه بؤس وتخلطم...	☐	☐	☐	☐	☐	☐	☐	☐
5	"الشمري" لـ"سبق": لقاح "كورونا" آمن ولا تصدقوا الشائعات..وهذا موعد انتهاء الجائحة	☐	☐	☐	☐	☐	☐	☐	☐
6	المملكة حولت أزمة «كورونا» إلى فرصة لتسريع «التحول الرقمي»	☐	☐	☐	☐	☐	☐	☐	☐
7	السلام عليكم لدي استفسار موظف انتدب من الرياض الى جده وخلال فترة انتدابه تم تعليق الحضور لمقرات العمل بسبب كورونا واستمر في المدينة المنتدب لها وقام بمهام عمله عن بعد حتى انتهاء انتدابه هل النظام يجيز الصرف له عن الفترة الواقعة خلال تعليق الحضور لمقرات العمل وشكرا	☐	☐	☐	☐	☐	☐	☐	☐
8	الفيديو بعنوان «ممرضة تموت بعد أخذها لقاح فايزر» غير صحيح، وهو للحظة إغماء تعرضت لها ممرضة أمريكية، وهو رد فعل يمكن أن يحدث مع أي لقاح أو حقنة.	☐	☐	☐	☐	☐	☐	☐	☐

Figure 3.3: One of the shared annotation sheets

3.7 Preprocessing

In the preprocessing stage, we divided the tasks into two sub-stages: one conducted before the annotation stage and the other after. Before the annotation stage, the preprocessing tasks were performed using REGEX formulas in Google Sheets, employing specific patterns to achieve the following:

- **Removing duplicate tweets:** to ensure that our dataset only contained unique entries, we identified and eliminated any duplicate tweets. This eliminated

Table 3.3: The counts of the annotated tweets in each emotion class by three annotators

Emotion Class	Rater 1		Rater 2		Rater 3	
	count	ratio	count	ratio	count	ratio
Happiness	4277	12.82%	2414	7.23%	4604	13.79%
Fear	1277	3.82%	576	1.73%	1214	3.63%
Disgust	821	2.47%	179	0.53%	840	2.51%
Anger	1358	4.06%	1723	5.16%	1499	4.49%
Surprise	756	2.27%	19	0.06%	419	1.25%
Optimism	1169	3.50%	1177	3.53%	1069	3.20%
Sadness	1723	5.16%	794	2.38%	1313	3.93%
Neutral	21992	65.90%	26491	79.38%	22415	67.16%

701 rows from the total of 34,074 to remain 33373 unique tweets. This step was crucial for maintaining data integrity and avoiding bias in our analysis.

- **Eliminating Twitter handles (@):** We removed Twitter handles from the tweets during preprocessing. By excluding the Twitter handles, we focused on the core content of the tweets, enabling us to analyze the emotions expressed without any influences.

- **Removing URLs:** We filtered out any URLs present in the tweets. This step helped us eliminate any external links or references that could potentially skew the emotional context of the tweets.

- **Filtering out English text:** Our analysis was specifically focused on tweets in the Arabic language. Therefore, we filtered out any English text that might have been present in the dataset. This ensured that our subsequent analysis was accurate and aligned with our research objectives.

- **Removing emojis and non-emoji symbols:** Emojis and non-emoji symbols,

Table 3.4: Examples of annotated tweets by different raters

Tweet	class by rater #1	class by rater #2	class by rater #3
مدينة الضباب تغلق ابوابها ابتداء من الغد ديسمبر بسبب السلاله الجديده المتحوره من يروس كورونا اغلاق تام للبوتيكات والمطاعم والصالونات وجميع المحلات الغير اساسيه فقط الصيدليات والسوبر ماركات بتكون فاتحه	5 (surprise)	8 (neutral/spam)	1 (happiness)
الحمدلله اخذت اول جرعة من اللقاح اللهم انفع بها خذ الخطوة	1 (happiness)	1 (happiness)	1 (happiness)
تعبت نفسيا والسبب ان كورونا للحين مو راضي يخلص	7 (sadness)	8 (neutral/spam)	2 (fearness)
اذا شركة واحدة اللي بتصنع اللقاح بيكون احتكار والشركات الطبية كل سنة تزيد توسعها علشان الدواء مثل له اكثر من شركة مصنعة هو نفس الحال مع كورونا	8 (neutral/spam)	8 (neutral/spam)	8 (neutral/spam)

while adding visual elements to the tweets, do not contribute to the textual content that conveys emotions. Hence, we removed these elements to maintain the focus on the textual information.

After the annotation stage, we proceeded with further deep preprocessing using the KNIME platform. This powerful data analytics tool allowed us to perform various preprocessing tasks to refine and clean the tweet data. The following steps were applied to the total of only class-matched tweets from the three raters (23689 tweets):

- **Remove punctuation:** Punctuation marks were removed from the tweet text. This step helped us eliminate unnecessary noise and focus solely on the words and their emotional significance. Punctuations include

  `(!@#\$%^&*)(_+=-><][/: ’×÷»«)`

- **Removing numbers:** Numeric characters were eliminated from the tweets. Since numbers do not contribute directly to the emotional content, their removal simplified the text and enhanced the accuracy of our subsequent analysis.

- **Removing double spaces and new lines:** Consecutive spaces and new lines were removed to ensure consistency in the text format.

- **Remove Arabic diacritics and character normalization:** Arabic diacritics, such as vowel marks, were removed from the text. Additionally, characters were normalized to ensure consistency and standardization across the dataset. This step eliminated any variations in the text that could potentially impact

the accuracy of our emotion analysis. Figure 3.4 shows an example of a tweet before and after this step.

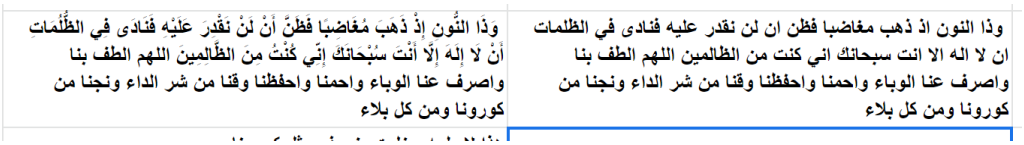

Figure 3.4: An example of a tweet before and after removing Arabic diacritics

- **Normalization:** Further normalization techniques were applied to the tweet text. This process aimed to standardize the data and convert characters to their original shape for all characters. Using KNIME's replacer node, the normalization of characters has been done as follows:
  - For the (ا) letter, only one character is allowed in a sequence. For example, (يا سلاام) is not allowed and is replaced with يا سلام.
  - For the (ب) letter, only one or two characters are allowed in a sequence. If three letters were in a sequence, they were replaced into one character only. For example, (السبانخ) and (بسبب) are allowed, but (ببيطاطس) is replaced with بطاطس.
  - For the (ت) letter, only one or two characters are allowed in a sequence. If three letters were in a sequence, they were replaced into one character only. For example, (تفاعل) and (تتفاعل) are allowed, but (تتتفاعل) is replaced with تفاعل

– For (ظ ع و خ) letters, only one character is allowed in a sequence. If two letters were in a sequence, they were replaced into one character only. Figure 3.5 shows an example of a tweet before and after this step.

واخيرا لقيت موعد لقاح كورونا والله ما اخليه ذا بطلعة الروح لقيته

واخيييييييرا لقيت موعد لقاح كورونا والله ما اخليه ذا بطلعة الروح لقيته

Figure 3.5: An example of a tweet before and after normalization

- **Remove duplicates again:** After the preprocessing steps, we conducted another round of duplicate removal to ensure a clean dataset without any redundant entries; 1140 duplicates were detected and eliminated in this step. Figure 3.6 shows the number of tweets at each stage throughout the implementation.

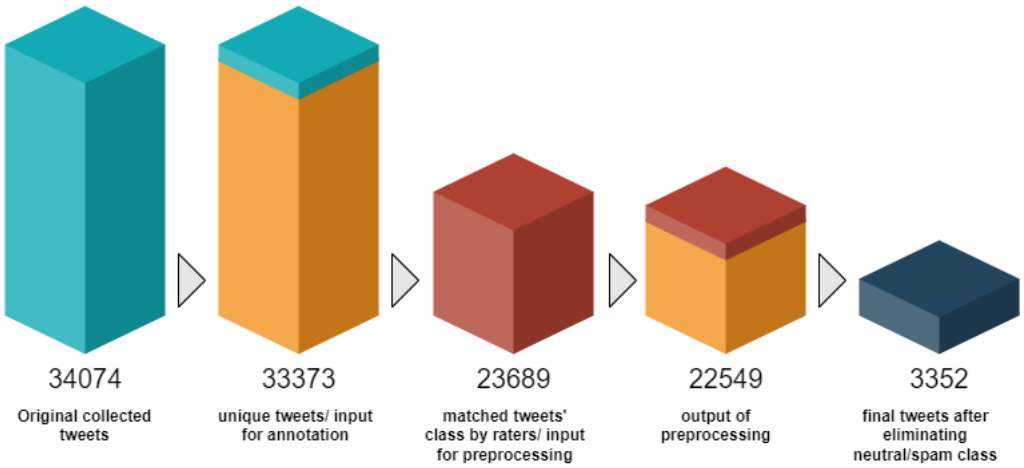

Figure 3.6: Counts of tweets at each stage of the implementation

- **Stemming:** Three different stemming techniques were applied to reduce words to their root form (Snowball Stemmer, Porter Stemmer, and Kuhlen

Stemmer [1] [22] [39]. This process helped consolidate similar word variations, enabling more accurate analysis and interpretation of emotions expressed in the tweets. The machine learning models are trained independently later on each of the outputs of these stemming techniques to compare between them.

- **Removing stopwords:** These are common words that do not have significant meaning in a given language. We removed stopwords from the text to focus on the meaningful content of the tweets. This was done using the Stop Word Filter node by KNIME and a stopwords dictionary published by [47], which provides a list of predefined stopwords for the Arabic language.

- **Tokenization:** Involves splitting the text into individual tokens or words. This step allows us to analyze the text at a more granular level. We utilized the Arabic tokenizer provided by the NLTK library to tokenize the preprocessed text into individual words.

- **Columns Filter:**We have filtered the dataset columns to focus only on the tweet and label class columns and ignored other features that will not affect the emotion analysis.

Each preprocessing task was carefully implemented using appropriate nodes and techniques within the KNIME platform. This comprehensive preprocessing approach ensured that our data was clean, standardized, and ready for further analysis and feature extraction. The output of this stage is 22549 clean tweets and ready for the features extraction stage. Figure 3.9 illustrates the flow of preprocessing nodes

in the KNIME platform. Figures 3.7 and 3.8 show the Word Cloud for the data set before and after the preprocessing stage.

Figure 3.7: Word Cloud for the original dataset before applying preprocessing tasks

3.8 Feature Extraction and Selection

In the Feature Extraction and Selection stage, we first eliminated the 8th emotion class label, which includes neutral/spam/advertisement text (19197 tweets), then we applied three distinct methods to the rest of (3352 tweets) to represent the

Figure 3.8: Word Cloud for the preprocessed dataset after applying preprocessing tasks

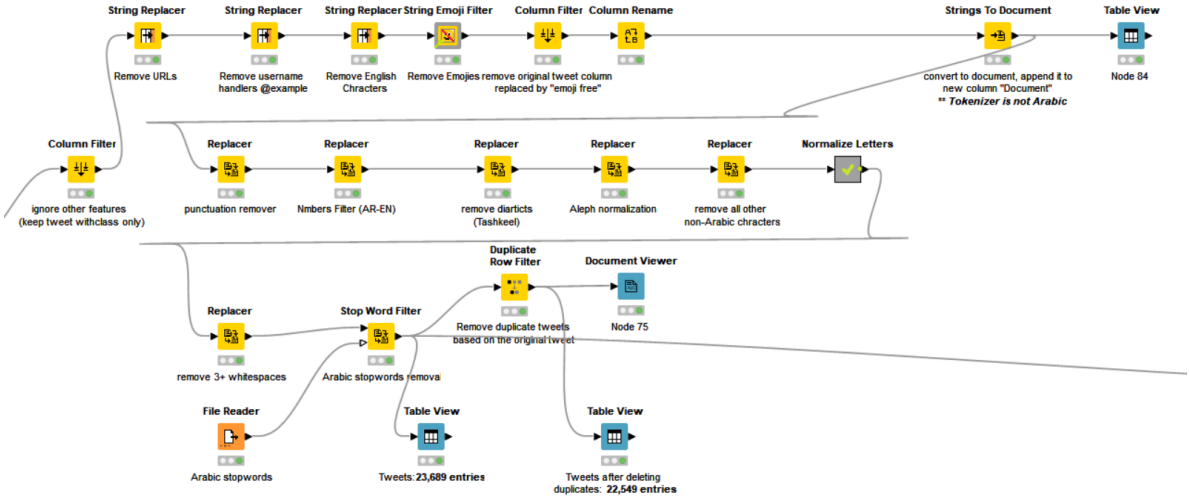

Figure 3.9: The flow of preprocessing nodes in the KNIME platform

preprocessed text data in a format suitable for machine learning models: Bag-of-Words, N-Gram, and TF-IDF.

3.8.1 Bag-of-Words (BoW)

The Bag-of-Words approach is a simple yet powerful technique used to convert text data into numerical vectors. In this method, each word in the preprocessed text is considered as a separate feature, and the frequency of each word is counted. The resulting vector represents the occurrence of each word in the text, disregarding the word order or grammar. We utilized the BoW method to create a feature matrix that captures the frequency of words in each tweet. We have applied two sub-flows of BoW to test which has better model accuracy. Figures 3.10 and 3.11 show the sequence of the two sub-flows. By employing BoW, we obtained a high-dimensional

representation of the text data, which served as input for the machine learning model.

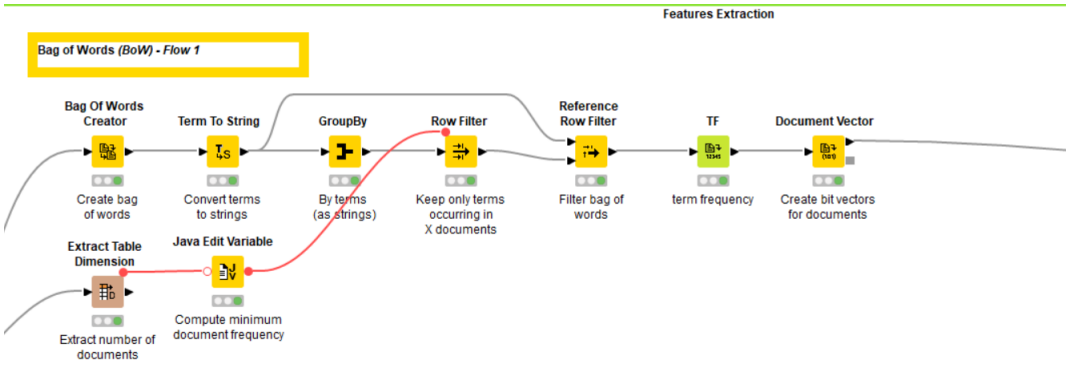

Figure 3.10: The first sequence of BoW flow

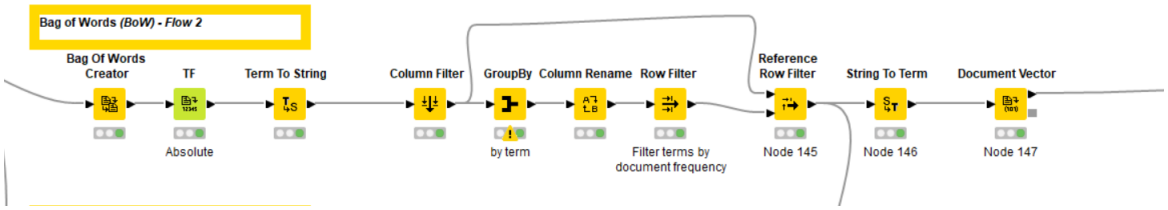

Figure 3.11: The second sequence of BoW flow

3.8.2 N-Gram

The N-Gram approach extends the Bag-of-Words method by considering sequences of N consecutive words as features. In addition to individual words, N-Gram includes word combinations of length *N*. This approach allows for capturing some context and word relationships in the text. We used bi-grams *(N=2)* to represent pairs of consecutive words in each tweet for our implementation. By incorporating N-Gram features alongside BoW, we aimed to capture some contextual information that may provide more nuanced insights into the emotions expressed in the text.

3.8.3 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a widely used technique in natural language processing to quantify the importance of each word in a document relative to a corpus. It considers both the frequency of a word in a specific document (Term Frequency) and the rarity of the word across the entire corpus (Inverse Document Frequency). TF-IDF assigns higher weights to words frequent in a document but rare across the corpus, making them more discriminative. By employing TF-IDF, we aimed to enhance the representation of words based on their significance in each tweet and across the entire dataset.

By employing these three feature extraction methods, we transformed the preprocessed text data into numerical representations that capture the essential information about the emotions expressed in the tweets. These feature matrices served as inputs for the machine learning model, enabling it to learn the patterns and relationships between the features and labeled emotions during training.

3.9 Resolve Data Imbalance (Oversampling)

Addressing class imbalance is imperative to developing an effective emotion analysis model. The inherent imbalance in the class distribution poses challenges in machine learning tasks. In this section, we elucidate the strategies to mitigate this issue, emphasizing data partitioning and oversampling techniques.

Table 3.5: Class distributions of the 3352 tweets before data splitting

Class	Tweets Count
Happiness (1)	1,926
Fear (2)	230
Disgust (3)	71
Anger (4)	482
Surprise (5)	13
Optimism (6)	243
Sadness (7)	387
Total	**3352**

3.9.1 Class Distribution

Before embarking on oversampling, it is pertinent to comprehend the initial class distribution within the dataset. An imbalance is evident in the original dataset, as highlighted by the following class distribution, as shown in Table 3.5 and Figure 3.12.

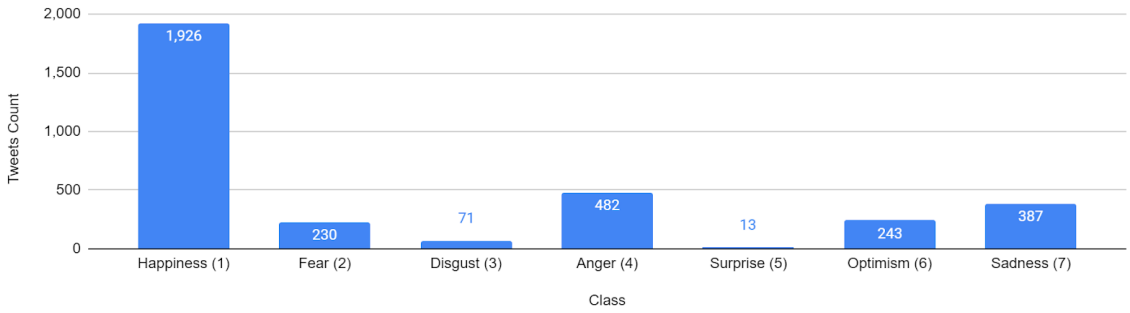

Figure 3.12: Class distributions of the 3352 tweets before data Splitting

The total number of tweets in the dataset was 3,352. Such class imbalance can affect the model's capacity to effectively discern less frequent emotion categories.

Table 3.6: Class distribution of the training dataset contains 2346 tweets after data splitting and before oversampling

Class	Tweets Count
Happiness (1)	1,348
Fear (2)	161
Disgust (3)	50
Anger (4)	337
Surprise (5)	9
Optimism (6)	170
Sadness (7)	271
Total	**2346**

Table 3.7: Class distribution of the testing dataset contains 1006 tweets after data splitting

Class	Tweets Count
Happiness (1)	578
Fear (2)	69
Disgust (3)	21
Anger (4)	145
Surprise (5)	4
Optimism (6)	73
Sadness (7)	116
Total	**1006**

3.9.2 Dataset Splitting

To evaluate the performance of the machine learning models, we employed a 70:30 train-test split ratio. This strategy ensured that 70% of the data was used for training the models, while the remaining 30% was set aside for evaluating their performance. This partitioning strategy serves a dual purpose: to allocate a substantial corpus for the training process and to ensure a robust evaluation phase by reserving a separate testing set. The class distribution of the training dataset and the testing dataset after the splitting is as in tables 3.6, 3.7 and Figures 3.13, 3.14.

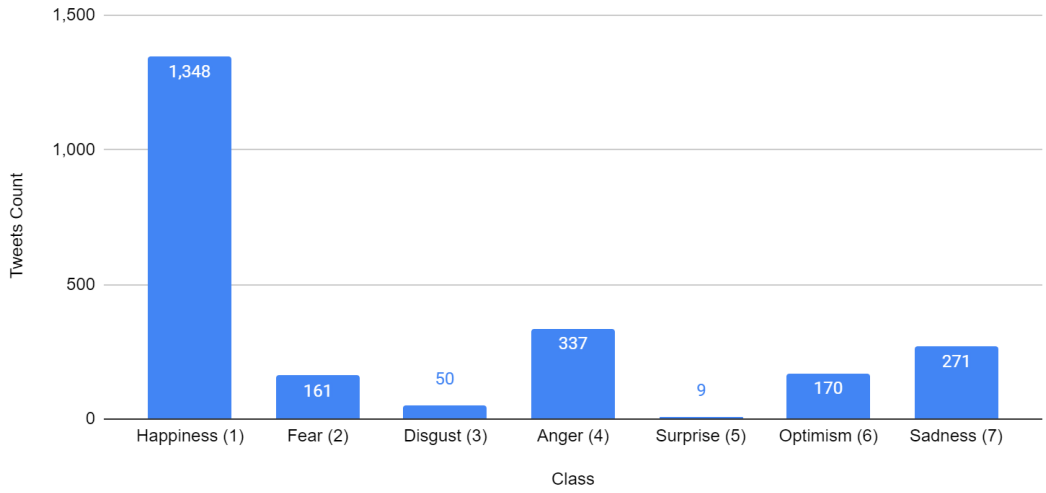

Figure 3.13: Class distribution of the training dataset contains 2346 tweets after data splitting and before oversampling

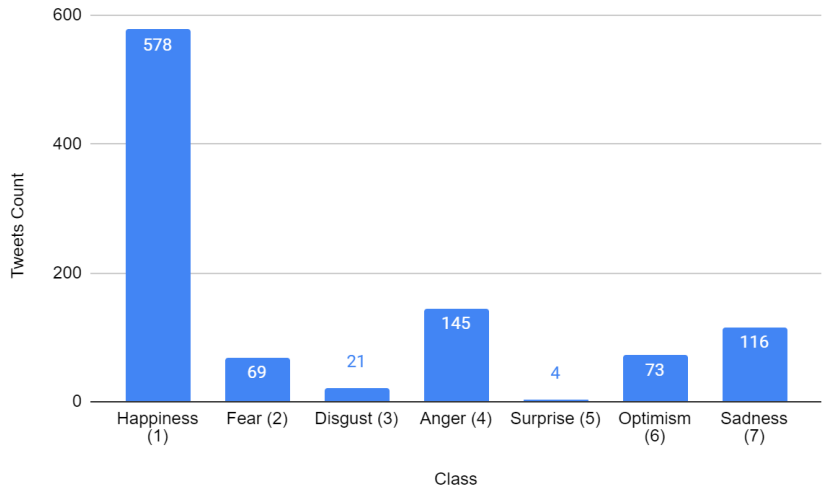

Figure 3.14: Class distribution of the testing dataset contains 1006 tweets after data splitting

Table 3.8: Class distribution of the training dataset contains 9399 tweets after data splitting and after oversampling

Class	Tweets Count
Happiness (1)	1,348
Fear (2)	1288
Disgust (3)	1350
Anger (4)	1348
Surprise (5)	1350
Optimism (6)	1360
Sadness (7)	1355
Total	**9399**

3.9.3 Oversampling

The principal endeavor to rectify the class imbalance within the training data encompassed the strategic application of oversampling techniques. Specifically, oversampling was directed solely at the training data. After the oversampling process, the class distribution in the training dataset was reconfigured as shown in table 3.8 and Figure 3.15.

This resampling procedure aimed to equalize the representation of emotions across classes in the training dataset, thereby enhancing the machine learning models' learning experience. With an augmented corpus, the models are better equipped to learn and generalize across all emotion categories. Conversely, the testing dataset remained unaltered by oversampling. This approach ensures the testing dataset remains representative of real-world scenarios and maintains its original distribution. It enables the comprehensive evaluation of the model's ability to accurately classify emotions, thus facilitating a robust and unbiased assessment of the model's performance. Figures 3.16 and 3.17 show this process implementation in KNME

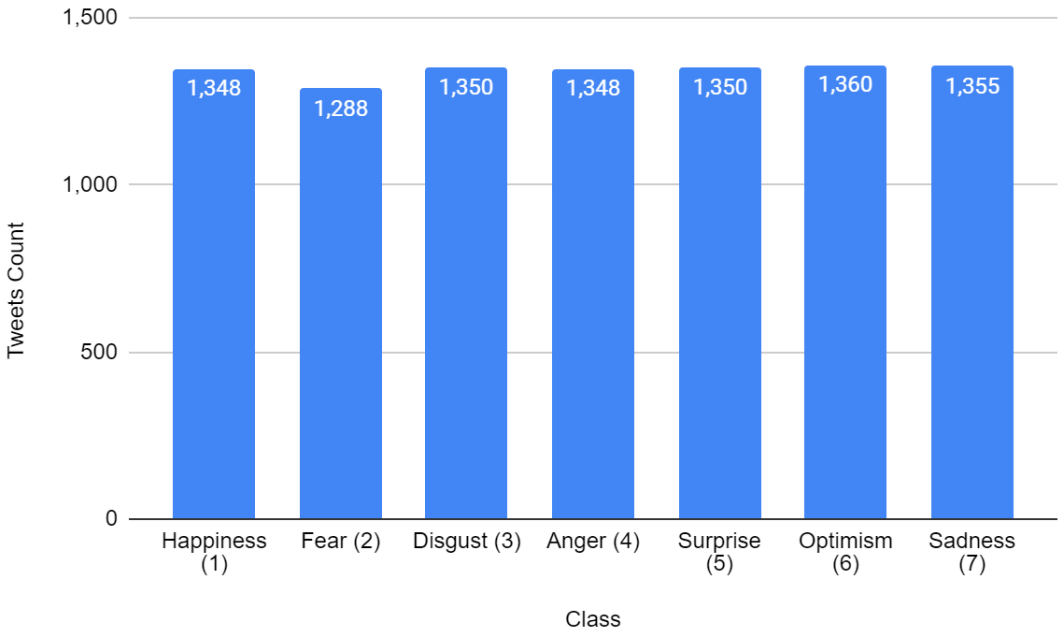

Figure 3.15: Class distribution of the training dataset contains 9399 tweets after data splitting and after oversampling

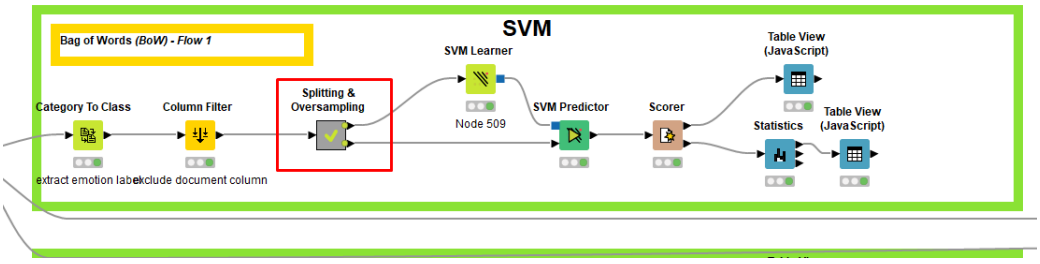

Figure 3.16: Oversampling flow at KNIME platform

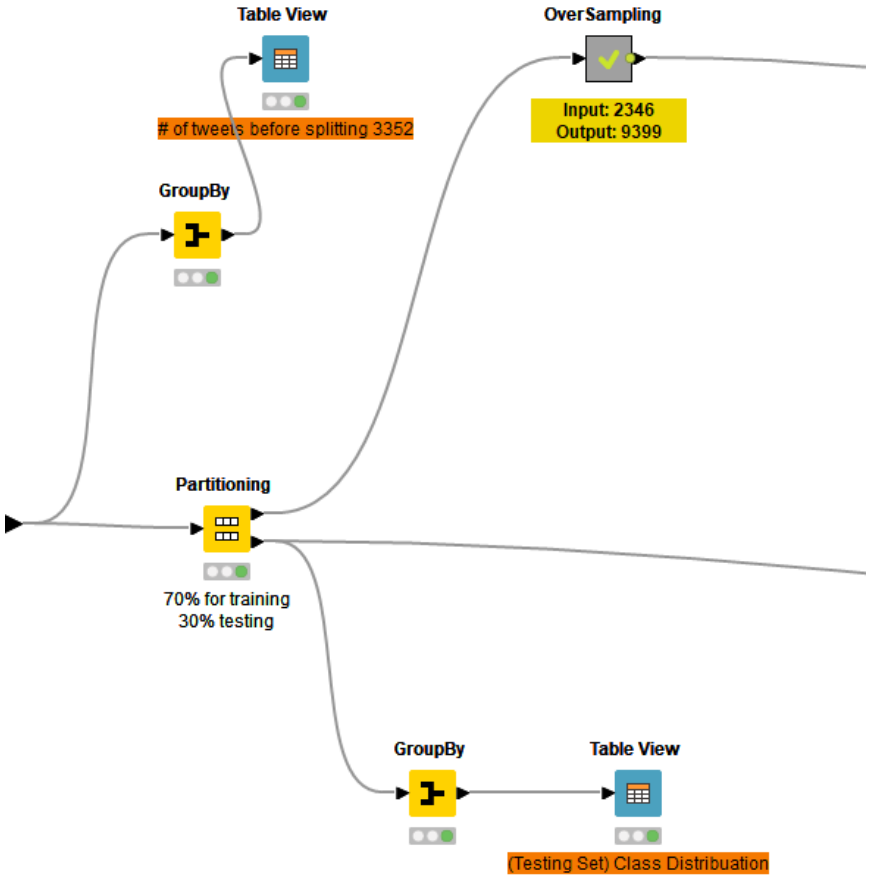

Figure 3.17: Oversampling flow at KNIME platform

Chapter 4

Implementation

4.1 Classification and Models Training

In the Classification stage, we embarked on the crucial task of predicting the emotions expressed in the 9399 tweets using the extracted features. Indeed, the combination of three stemming techniques, four feature extraction methods, and three machine learning algorithms resulted in a total of 36 different models being trained in this stage. Each model represents a unique configuration of the preprocessing and classification pipeline, contributing to the comprehensive evaluation and comparison of various approaches.

We obtained an enriched dataset with numerical representations of the extracted features. However, the emotion labels needed to be converted to numerical classes to train the machine learning models. This transformation involved mapping each emotion category to a unique numerical class, making the dataset suitable for

classification.

Having prepared the data and set up the train-test split, we applied three widely used machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and Decision Tree. Each algorithm underwent training on the training data to learn the underlying patterns and relationships between the extracted features and emotions. After training, the models were tested using the test data to evaluate their predictive capabilities.

In the learner node of each machine learning algorithm, specific configurations and options were carefully selected to optimize the the implementation of performance of the models. For the Support Vector Machine (SVM), we utilized the polynomial kernel with power approximately equal to 1, bias around 1, and gamma set to approximately 1. The overlapping penalty was set within the range of 0.1 to 1, enabling us to control the influence of overlapping data points on the model's decision boundaries. In the case of logistic regression, we opted for the stochastic average gradient (SAG) solver, known for its efficiency [67] and ability to handle large datasets. The maximal number of epochs was set between 10 and 20, ensuring the algorithm converged to the optimal solution while avoiding overfitting. For the Decision Tree algorithm, we set the maximum number of patterns the tree will store to support highlighting to a default value of 10,000. This configuration allowed us to manage the complexity of the tree while still capturing the relevant patterns and relationships in the data. Figures 4.1, 4.2, and 4.3 show the training implementation nodes and model configurations.

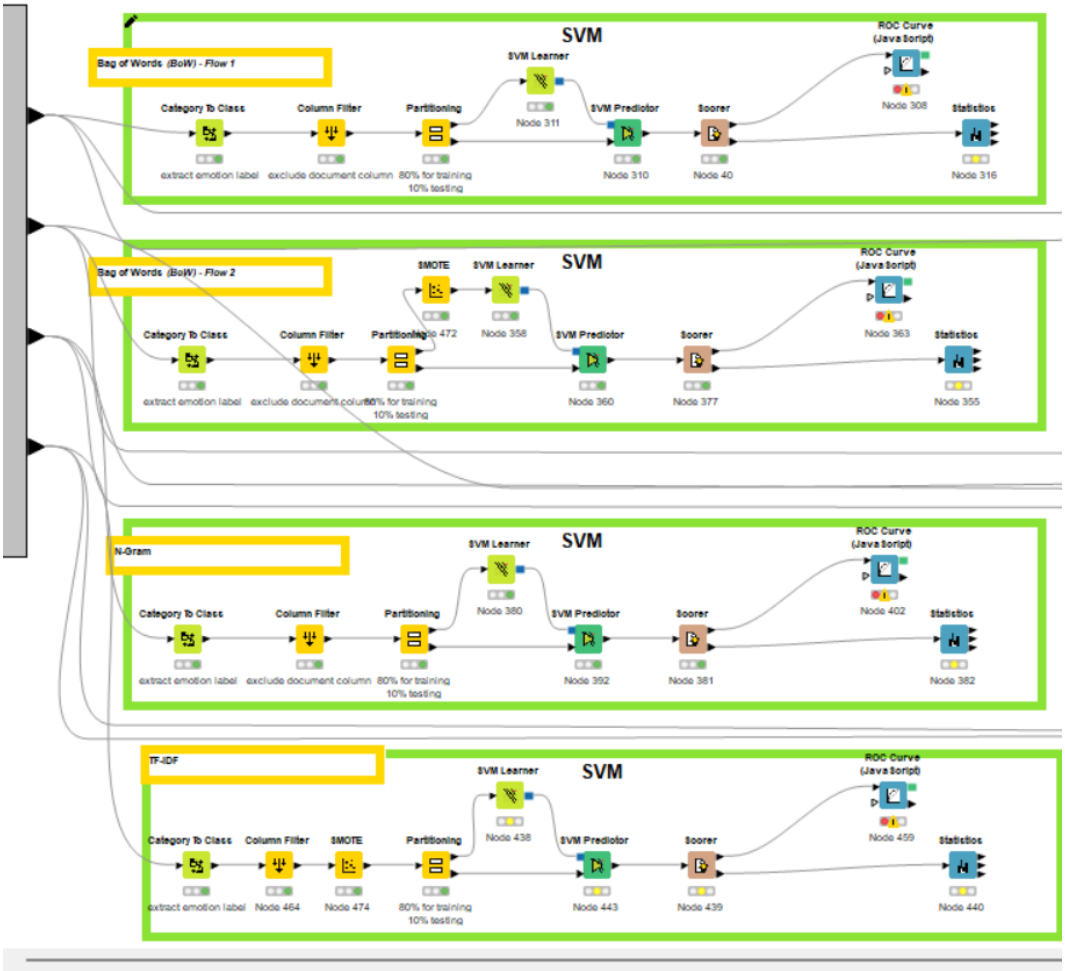

Figure 4.1: SVM model nodes with four features extraction techniques using only one Stemming method (Kuhlen Stemmer)

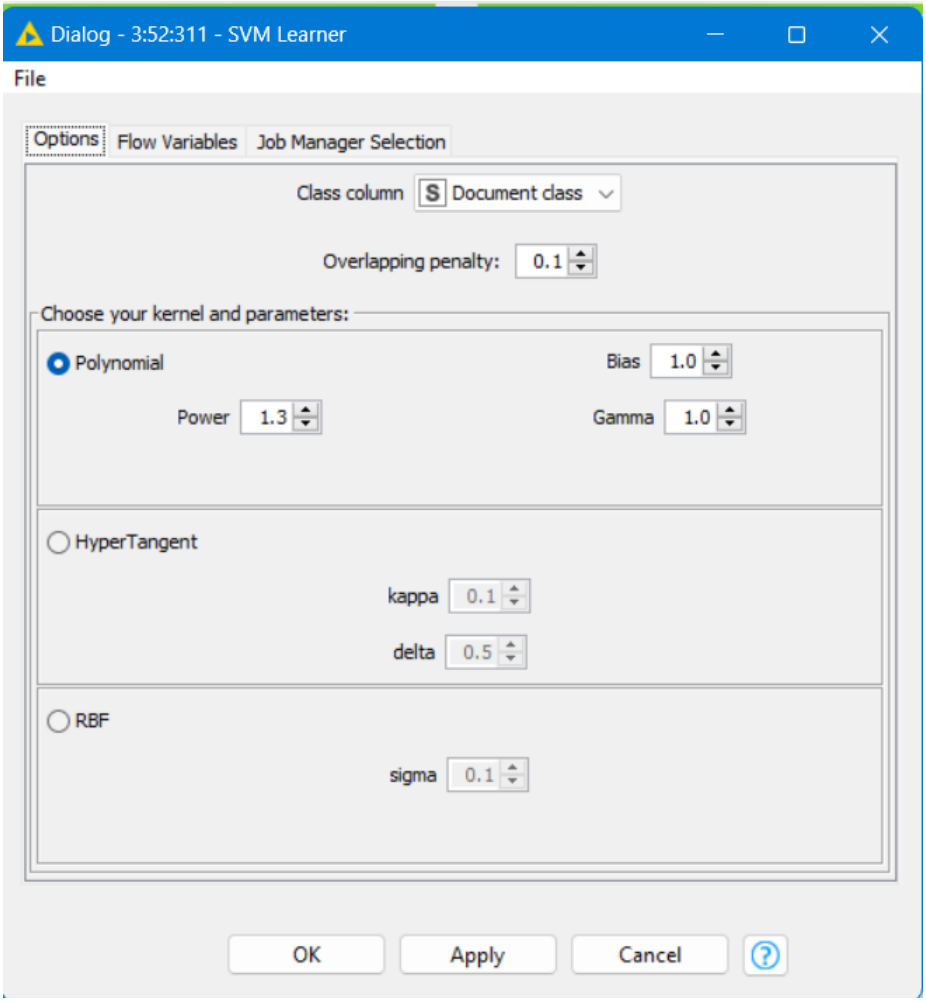

Figure 4.2: SVM learner node's configuration

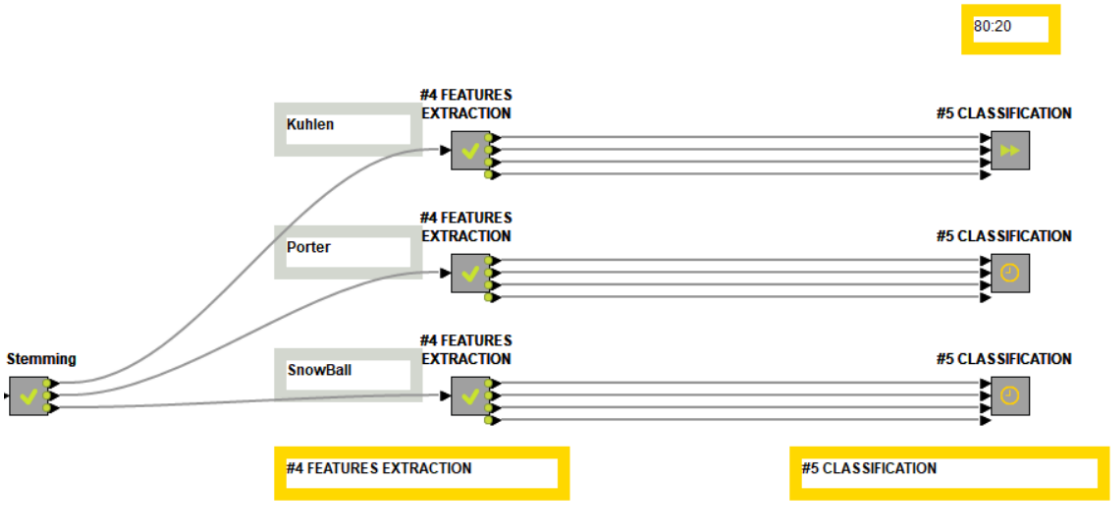

Figure 4.3: General flow of stemming, features extraction, and classification stages

Chapter 5

Results and Discussion

5.1 Evaluation Methods

For the evaluation process, we utilized the KNIME platform, employing the Model Predictor node and the Scorer node. The Model Predictor node takes the test data partition from the partitioning node and the trained model from the learner node as inputs. It then uses the trained model to predict the emotion labels for the given test data. The Scorer node compares the predicted and actual labels, generating a confusion matrix that displays the number of correct and incorrect predictions for each emotion class.

The output of the Scorer node includes various performance metrics such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, overall accuracy, and Cohen's kappa. These metrics provide valuable insights into the classification performance of each model, enabling

us to compare and evaluate their effectiveness in classifying emotions.

Additionally, the Scorer node calculates the statistics based on the confusion matrix, providing further insights into the models' classification capabilities. Figure 5.1 illustrates the whole flow of evaluation of one of the experiments.

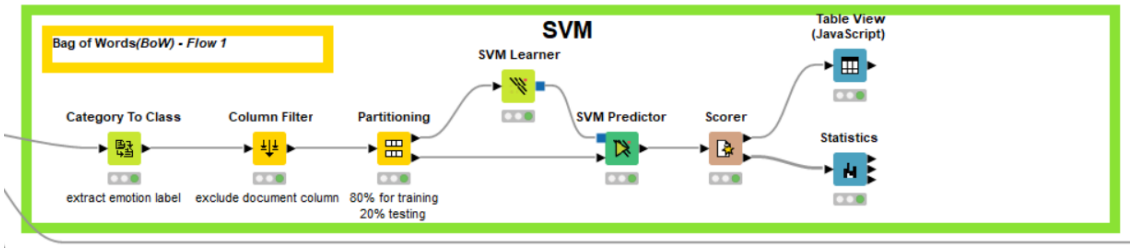

Figure 5.1: One of the SVM model training and evaluation experiments flow

5.1.1 Confusion Matrix

The confusion matrix is a fundamental evaluation tool that provides a tabular representation of the performance of a machine-learning classification model [18]. It gives a comparison between actual and predicted values as it displays the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions for each emotion class, which will be used in the measure metrics as in section 5.2. Furthermore, the Confusion Matrix allows us to assess the model's accuracy and ability to classify emotions correctly. The confusion matrix is a square matrix of size N x N, where N represents the number of emotion classes. In our specific case, we have a 7 x 7 confusion matrix, as there are seven emotion classes: happiness, fear, disgust, anger, sadness, surprise, and optimism. Table 5.1 displays an example of a confusion matrix.

Table 5.1: Confusion Matrix for seven classes

	Predicted Class						
Actual Class	**Happiness**	**Fear**	**Disgust**	**Anger**	**Sadness**	**Suprise**	**Optimism**
Happiness	TP1	FP1	FP2	FP3	FP4	FP5	FP6
Fear	FP7	TP2	FP8	FP9	FP10	FP11	FP12
Disgust	FP13	FP14	TP3	FP15	FP16	FP17	FP18
Anger	FP19	FP20	FP21	TP4	FP22	FP23	FP24
Sadness	FP25	FP26	FP27	FP28	TP5	FP29	FP30
Suprise	FP31	FP32	FP33	FP34	FP35	TP6	FP36
Optimism	FP37	FP38	FP39	FP40	FP41	FP42	TP7

5.1.2 Class Distribution Consideration

As our dataset exhibits an imbalanced class distribution, with some emotion classes having significantly fewer instances than others, we applied stratified cross-validation. Stratified cross-validation ensures that each fold retains the same proportion of instances for each emotion class as the original dataset. This approach is crucial for preventing biased evaluations and ensuring that each emotion class is represented appropriately during model training and testing.

5.1.3 Comparative Analysis

To conduct a comparative analysis, we evaluated a total of 36 different models, considering the combination of three stemming techniques (Kuhlen Stemmer, Porter Stemmer, and Snowball Stemmer), four flows of feature extraction methods (Bag-of-Words, N-Gram, and TF-IDF), and three machine learning algorithms (Support Vector Machine, Logistic Regression, and Decision Tree). By comparing the performance of these models, we can identify the most practical combination of techniques

and algorithms for sentiment analysis in the context of our research.

We ensure a rigorous and comprehensive assessment of our emotion classification models by employing these evaluation methods. Through these evaluation methods, we gain a holistic understanding of the strengths and weaknesses of our models in accurately capturing emotions expressed in the Saudi dialect tweets about vaccinations.

5.2 Evaluation Performance Metrics

In evaluating our emotion classification models, we employed a set of well-established performance metrics to assess their effectiveness in emotion analysis of the Saudi dialect tweets related to vaccinations. Metrics include Accuracy, Precision, Recall, and F1-score [32]. These metrics provide valuable insights into the models' ability to classify emotions accurately and are vital for comparing the performance of different machine learning algorithms, stemming techniques, and feature extraction methods [74].

5.2.1 Accuracy

Accuracy measures the proportion of correctly classified instances over the total number of instances in the dataset. It is a widely used metric for assessing the overall performance of a classification model. In the context of our research, accuracy indicates the model's ability to correctly identify emotions in the tweets,

considering all emotion classes. The mathematical formula for accuracy is [57] $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

where

- TP (True Positives) is the number of correctly predicted instances for a particular emotion class.
- TN (True Negatives) is the number of correctly predicted instances for all classes that are not the particular emotion class.
- FP (False Positives) is the number of instances incorrectly classified as the particular emotion class.
- FN (False Negatives) is the number of instances that should have been classified as the particular emotion class but were not.

5.2.2 Precision

Precision measures the proportion of true positive predictions for a particular emotion class over the total number of instances predicted as that class. It indicates the model's ability to avoid false positive predictions for a given class. In our research, precision provides valuable information about the model's reliability when it classifies a tweet as a specific emotion. The mathematical formula for precision is $Precision = \frac{TP}{TP+FP}$

5.2.3 Recall (Sensitivity)

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions for a particular emotion class over the total number of instances that belong to that class. Recall reflects the model's ability to identify all instances of a specific emotion class correctly. In our study, recall is crucial for understanding how well the model captures the emotions expressed in the tweets. The mathematical formula for the recall is $Recall = \frac{TP}{TP+FN}$

5.2.4 F1-Score

The F1-score is the harmonic mean of precision and recall. It combines both metrics into a single value and is particularly useful when there is an imbalance in class distribution. The F1-score provides a balanced evaluation of the model's performance in terms of both false positives and false negatives. In our research, the F1 score is essential as it allows us to consider the trade-offs between precision and recall in emotion classification [70]. The mathematical formula for the F1-score is $F1\text{-}score = 2 \ = \frac{PrecisionxRecall}{Precision+Recall}$. These metrics are especially relevant to our research given the dataset's imbalanced distribution of emotion classes and the significance of accurately capturing emotions expressed in the Saudi dialect tweets about vaccinations.

5.3 Results and Evaluation

This section presents the outcomes of the 36 experiments conducted during the implementation stage. As mentioned earlier, our input dataset comprised 9399 tweets. To explore various configurations, we employed three different stemming techniques, four distinct feature extraction methods, and three distinct machine learning algorithms, leading to 36 models for training and evaluation.

Tables 5.2, 5.3, and 5.4 comprehensively summarize the results for all 36 models, presenting the performance metrics for each emotion class. These tables offer a basis for comparing each model's accuracy, precision, recall, and F1-score of each model across different feature extraction methods and stemming techniques. Among the configurations, the Logistic Regression model achieved the highest accuracy, reaching 74.95% when combined with N-Gram feature extraction and Snowball stemming. It also achieves the same 74.95% accuracy when combined with the BoW with the Snowball stemmer, followed by 74.75% Logistic Regression when combined with TF-IDF and Snowball Stemmer. The SVM model showcased a close performance with 74.35% accuracy when combined with N-Gram and Snowball.

Moving on to the recall metric, the SVM model achieved a remarkable 91.34% in both experiments, compromising N-Gram and TF-IDF when combined with the Snowball stemmer, displaying a higher ability to identify true positive instances correctly and making them top performers in emotion classification. The Logistic Regression result shows a close percentage of 91.00% with TF-IDF and Snowball stemmer.

Table 5.2: SVM model results based on four features extraction techniques and three stemming methods

SVM		Evaluation Metrics			
Features Extraction Technique	**Stemming Method**	**Accuracy %**	**Recall%**	**Precision%**	**F-Measure%**
BoW-1	**Kuhlen**	64.31	83.04	95.61	88.88
BoW-1	**Snowball**	67.39	85.46	96.10	90.47
BoW-1	**Porter**	64.61	84.94	95.71	90.00
BoW-2	**Kuhlen**	70.67	87.54	93.35	90.35
BoW-2	**Snowball**	74.15	89.96	94.37	92.11
BoW-2	**Porter**	70.57	87.02	93.32	90.06
N-Gram	**Kuhlen**	70.67	87.19	92.98	90.00
N-Gram	**Snowball**	74.35	91.34	93.45	92.38
N-Gram	**Porter**	70.17	87.19	98.98	90.00
TF-IDF	**Kuhlen**	70.17	87.19	92.98	90.00
TF-IDF	**Snowball**	74.35	91.34	93.45	92.38
TF-IDF	**Porter**	70.17	87.19	92.98	90.00

Regarding precision, the SVM model demonstrated an impressive 98.98% precision rate when trained after using the N-Gram technique and Porter stemmer, indicating its capability to limit the number of false positives and ensure the accuracy of positive predictions. The Logistic Regression result shows a close percentage of 97.23% with BoW and Kuhlen Stemmer, followed by another Logistic Regression experiment with Bow and Porter Stemmer achieving 96.60%.

The F-measure results highlighted the Logistic Regression model as the most balanced performer between precision and recall, achieving 92.93% and 92.51%, particularly when combined with TF-IDF and BoW feature extraction techniques with the Snowball stemmer. The SVM model followed closely in third place, achieving 92.38% when N-Gram and Snowball were used.

Table 5.3: Decision Tree model results based on four features extraction techniques and three stemming methods

Decision Tree		Evaluation Metrics			
Features Extraction Technique	**Stemming Method**	**Accuracy %**	**Recall%**	**Precision%**	**F-Measure%**
BoW-1	**Kuhlen**	67.29	83.73	95.46	89.21
BoW-1	**Snowball**	66.10	84.25	92.40	88.14
BoW-1	**Porter**	67.29	83.73	95.46	89.21
BoW-2	**Kuhlen**	66.00	84.42	94.39	89.13
BoW-2	**Snowball**	69.38	88.23	90.58	89.39
BoW-2	**Porter**	66.79	86.85	91.27	89.00
N-Gram	**Kuhlen**	66.79	86.85	91.27	89.00
N-Gram	**Snowball**	69.38	88.23	90.58	89.39
N-Gram	**Porter**	66.79	86.85	91.27	89.00
TF-IDF	**Kuhlen**	65.90	84.25	95.49	89.52
TF-IDF	**Snowball**	67.99	85.81	92.36	88.96
TF-IDF	**Porter**	66.00	84.42	94.39	89.13

Table 5.4: Logistic Regression model results based on four features extraction techniques and three stemming methods

Logistic Regression		Evaluation Metrics			
Features Extraction Technique	**Stemming Method**	**Accuracy %**	**Recall%**	**Precision%**	**F-Measure%**
BoW-1	**Kuhlen**	63.81	79.23	97.23	87.32
BoW-1	**Snowball**	69.28	87.19	94.73	69.28
BoW-1	**Porter**	66.00	83.73	96.60	89.71
BoW-2	**Kuhlen**	72.86	89.79	93.68	91.69
BoW-2	**Snowball**	74.95	90.83	94.25	92.51
BoW-2	**Porter**	73.36	89.61	93.84	91.68
N-Gram	**Kuhlen**	73.06	89.96	94.71	92.28
N-Gram	**Snowball**	74.95	90.31	94.05	74.95
N-Gram	**Porter**	73.26	89.10	94.49	91.71
TF-IDF	**Kuhlen**	72.16	87.88	94.24	90.95
TF-IDF	**Snowball**	74.75	91.00	94.94	92.93
TF-IDF	**Porter**	72.76	88.58	93.60	91.02

5.4 Discussion

The results from our comprehensive experiment illuminate the effectiveness of our proposed Emotion Analysis machine learning model in classifying emotions and feelings expressed in Tweets written in the Saudi dialect. Our study's primary aim was to develop a precise and dependable model that enhances existing study results and aligns with the Saudi dialect's unique linguistic and cultural intricacies. Figures 5.2, 5.3, 5.4, and 5.5 highlight the most accurate, precise recall and F1-score models. The Logistic Regression model, as depicted, surpasses others in terms of Accuracy and F-Measure metrics, signifying a significant achievement. The SVM model displays the highest Recall (sensitivity) and Precision performance. The third model (Decision Tree) is out of competition with the top three. This success validates the fulfillment of objectives 1 and 4 in this work.

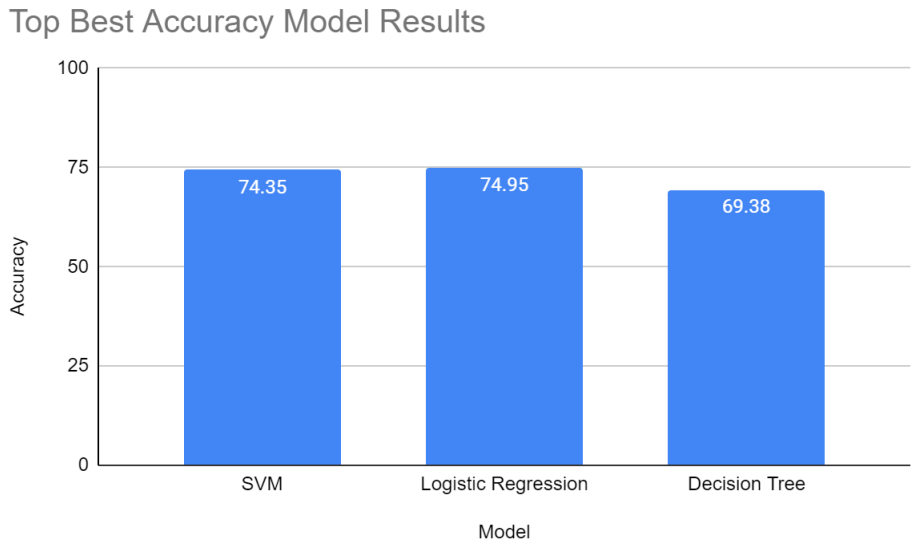

Figure 5.2: The top accuracy results achieved by the three models

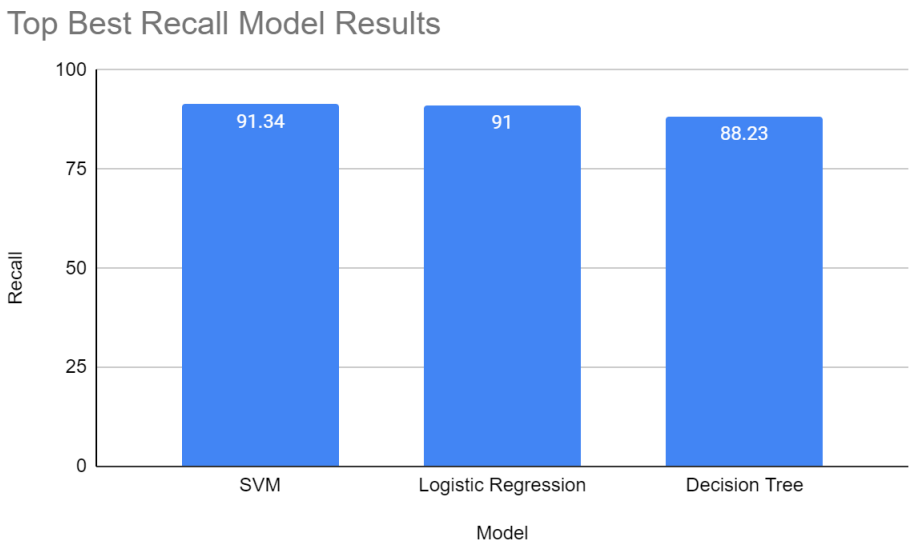

Figure 5.3: The top recall results achieved by the three models

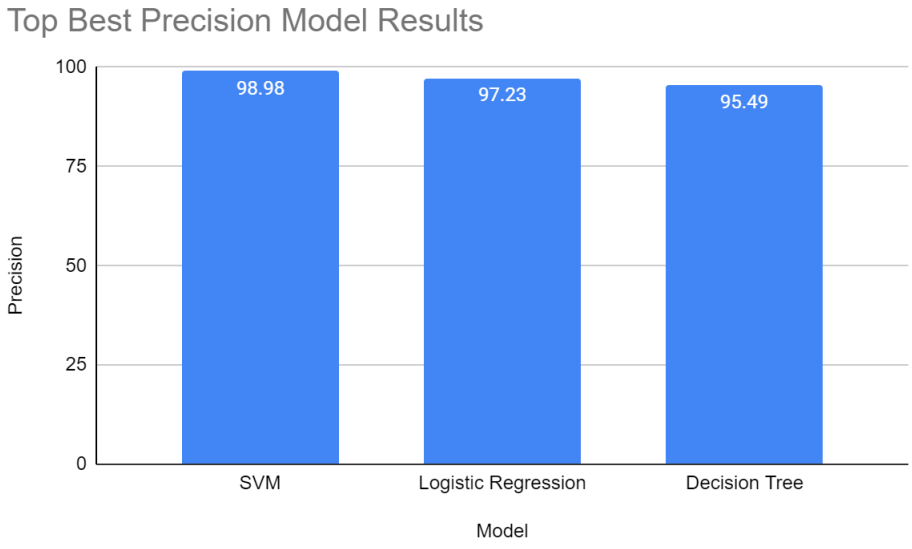

Figure 5.4: The top precision results achieved by the three models

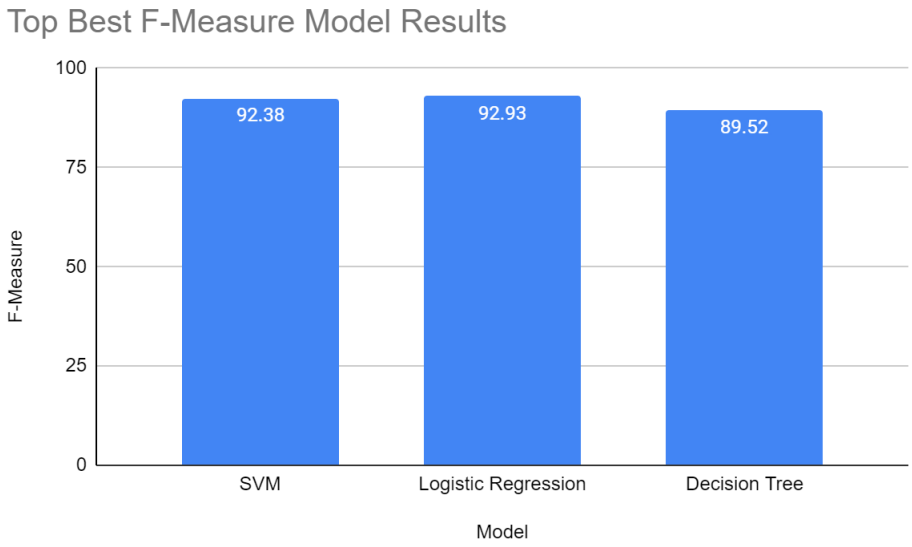

Figure 5.5: The top f-measure results performed by the three models

When compared to existing works in the domain, our progress is remarkable. For instance, considering a study that conducted 73.39% accuracy through an SVM approach, which is 1.56% lower than our Logistic Regression result, and they solely employed two machine learning algorithms (SVM and MNB) [9], our accomplishments are striking. Not only did we surpass this accuracy benchmark, but we also encompassed a diverse array of machine-learning algorithms, leading to a more robust and comprehensive evaluation. This aligns with objective 2 of this work. These achievements can be attributed to the meticulous approach we adopted in our implementation, embracing multiple stemming techniques, feature extraction methods, and machine learning algorithms. By exploring varied configurations, we harnessed the strengths of each element, synergizing them to yield the most accurate outcomes.

The performance of our model was significantly influenced by the careful selection of

stemming techniques and feature extraction methods. Results indicate that specific combinations of N-Gram feature extraction with Snowball stemming, yield the highest accuracy. This underscores the importance of selecting the suitable machine learning algorithm and optimizing preprocessing and feature engineering stages to fully exploit the data's potential.

5.4.1 General Attitudes Towards COVID-19 Vaccinations in Saudi Arabia

The outcomes stemming from our model's performance shed light on the predominant concerns and apprehensions surrounding the COVID-19 vaccination endeavors in Saudi Arabia. The precision of our emotion analysis has armed us with the capability to extract profoundly insightful understandings from the prevailing public sentiment. This advancement has facilitated a richer and more nuanced comprehension of the perspectives held by the citizens, thereby fortifying decision-making in healthcare initiatives by harnessing knowledge-rich insights. The distribution of annotated tweets among diverse emotion classes is visually presented in Table 5.5 and Figure 5.6 below.

These graphical depictions offer an unmistakable portrayal of the emotional landscape within the amassed tweets. It is unmistakably evident that emotions span a broad continuum, with 'Happiness' taking the forefront, trailed by 'Anger,' 'Sadness,' and 'Optimism.' Notably, 'Disgust' is relatively scarce, while 'Surprise' exhibits the lowest frequency. As displayed in Figure 5.7, the word cloud of preprocessed tweets further enriches our understanding.

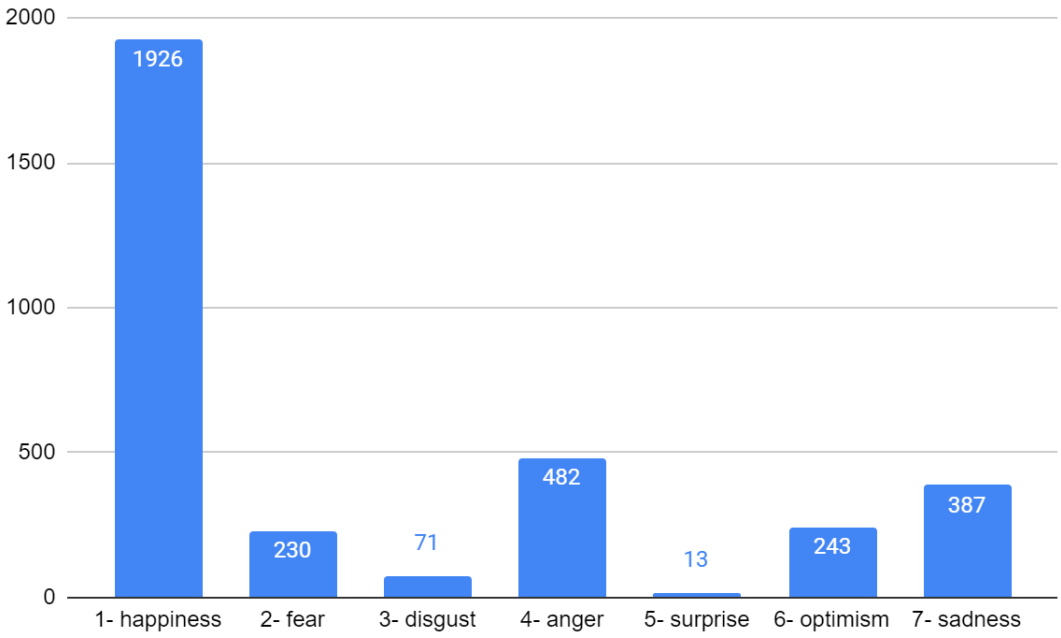

Figure 5.6: The distribution of annotated tweets across various emotion classes

Table 5.5: Number of instances in each emotion class in the final stage of implementation

Class	Number of Instances	%
1- happiness	1926	57.5
2- fear	230	6.9
3- disgust	71	2.1
4- anger	482	14.4
5- surprise	13	0.4
6- optimism	243	7.2
7- sadness	387	11.5

Figure 5.7: Word cloud of preprocessed tweets

Our insights, coupled with the utilization of Latent Dirichlet Allocation (LDA) [75] on the resultant dataset, signify a prevailing positive disposition among the Saudi Arabian populace towards the ongoing vaccination campaign against the COVID-19 pandemic. As illustrated in Figure 5.6, a staggering 57.46% of tweets emanate happiness towards the pronouncements and campaigns by the Health Ministry. This is closely trailed by anger at 14.38%, sadness at 11.55%, and optimism at 7.25%, including prayers and an optimistic outlook for swift relief from this crisis. Fearful emotions account for 6.86%, while disgust and surprise comprise 2.12% and 0.39%, respectively. These findings underscore the populace's contentment with vaccine uptake and availability, particularly the high disparity between the frequencies of positive emotions versus fear and sadness. By this, we have achieved our third objective, which entails providing valuable insight into the general attitudes toward vaccinations.

5.4.2 Saudi Dialect Labeled-Tweets Corpus Availability

In alignment with our objective to produce a Saudi dialect labeled-tweets corpus in the healthcare and COVID-19 vaccination domain, we are pleased to announce the availability of the "saudiEAR" repository on GitHub [66]. This repository contains a comprehensive collection of tweets, including both the original dataset collected and the preprocessed version. By making this corpus publicly accessible, we aim to contribute to the research community and facilitate advancements in sentiment analysis, emotion classification, and related fields. This open dataset enables researchers and practitioners to explore the intricacies of sentiment expression in the

Saudi dialect, particularly in healthcare and COVID-19 vaccination discussions. By releasing this resource, we have successfully achieved our fifth objective, which entails providing a valuable dataset for further research.

Chapter 6

Conclusion, Limitations, and Future Work

6.1 Conclusion

In conclusion, this research endeavor aimed to bridge a substantial gap in the field of Emotion Analysis, with a specific focus on the Saudi Arabian context within the COVID-19 vaccination campaigns. The primary objective was to develop a highly effective machine-learning model to classify Saudi dialect tweets into multiple distinct emotion categories accurately. This pursuit was motivated by the scarcity of comprehensive studies addressing the nuances of emotional analysis within the Saudi dialect. Furthermore, there was a distinct lack of a labeled-tweets corpus suitable for training models in this context.

Employing a robust methodology, we embarked on a multifaceted journey. A to-

tal of 34,074 Arabic tweets were meticulously collected from Twitter, focusing on discussions related to COVID-19 vaccines in the Kingdom of Saudi Arabia in a period of. These tweets were manually annotated by three expert raters fluent in Saudi dialects, ensuring cultural relevance. Each tweet was thoughtfully categorized into one of eight emotion classes: happiness, sadness, disgust, surprise, optimism, anger, fear, and neutral sentiments. The agreed-upon tweets were then subjected to thorough preprocessing. Various cleaning and filtering steps were applied to achieve a refined dataset of 3,352 tweets, which was later significantly expanded through an oversampling process until it reached 9,399 tweets. The subsequent phase of the study involved conducting 36 machine-learning experiments. We employed a trio of main machine learning models, including Support Vector Machines (SVM), Logistic Regression, and Decision Trees. These models were coupled with three distinct stemming techniques, namely Porter, Kuhlen, and Snowball, and four diverse methods for identifying and extracting features, which included Term Frequency-Inverse Document Frequency (TF-IDF), Bag-of-Words (BoW) in its variations (BoW-1 and BoW-2), and N-Gram analysis. This comprehensive approach enabled a deep exploration of the public sentiment surrounding COVID-19 vaccination campaigns in Saudi Arabia.

Our findings yield significant insights on multiple fronts. First and foremost, our refined machine learning model, particularly the Logistic Regression model, achieved a notable accuracy rate of 74.95%. This finding underscores the capability of our approach to classify emotions expressed in Saudi dialect tweets effectively. Notably, the SVM model achieved a recall rate of 91.34%, complemented by an impressive precision rate of 98.98%. The F1-Score metric demonstrated a commendable rate

of 92.93% when employing the Logistic Regression model.

A comparative analysis of our results with existing studies accentuates the progress achieved in our research. Our model's accuracy improved by 1.56%, while metrics such as precision, recall, and F1-Score also exhibited marked enhancements, underscoring the effectiveness of our approach.

The insights gleaned from the dataset unveiled a prevailing positive sentiment toward COVID-19 vaccination campaigns. These results provided valuable insights into the general sentiments and attitudes of the Saudi Arabian population regarding COVID-19 vaccines and their engagement with vaccination initiatives. The data indicated that happiness was the predominant emotion at 57.5%, anger at 14.4%, and sadness at 11.5%. These findings signify the joy people felt at the prospect of a vaccine heralding the end of the pandemic and its associated challenges. It also reflects the populace's loyalty and trust in their government and the decisions made by the Ministry of Health, emphasizing their belief that collective interests are intertwined in the fight against the virus.

Lastly, as a contribution to the scientific research community and the advancement of machine learning and sentiment analysis in the Saudi dialect, we have made available the labeled dataset generated throughout this study. This dataset encompasses 33,373 tweets with their original details, the processed versions, and their assigned emotional classifications. These resources are intended to facilitate further research and exploration in emotion analysis, particularly within the Saudi dialect.

6.2 Research Limitations

In acknowledging our achievements, it is equally vital to recognize the inherent limitations of this research. Firstly, although substantial, the dataset's size may still present constraints when addressing complex and evolving linguistic expressions in social media. Additionally, potential annotation biases introduced by human raters could affect the data quality. As with any study, the specific context of COVID-19 vaccination campaigns may influence the emotional expressions in the collected tweets. This context dependency might restrict the model's applicability to domains or topics. Lastly, it's worth noting that the Decision Tree model exhibited lower performance across the four primary evaluation metrics, underscoring its limited suitability for Emotion Analysis in the Saudi dialect context.

6.3 Future Work

Our study has opened up many possibilities for future research. First, we need to improve the performance of our model by incorporating deep learning methods. More advanced neural network architectures, such as recurrent or transformer-based models, could allow us to make more nuanced sentiment classifications, especially when capturing context-dependent linguistic nuances.

In addition, we can gain deeper insights into the linguistic and contextual cues that underlie the expression of emotions by exploring the interpretability of our model's decisions. Techniques such as attention mechanisms or layer-wise relevance

propagation can help us understand how the model makes its decisions.

Finally, we can broaden the scope of emotion analysis and its implications for healthcare communication, public sentiment monitoring, and decision-making by extending our model's applicability to other dialects and languages within the Arab region.

Bibliography

[1] Agichtein, E. and Gravano, L. [2000], Snowball: extracting relations from large plain-text collections, *in* 'Proceedings of the fifth ACM conference on Digital libraries (DL '00)', Association for Computing Machinery, New York, NY, USA, pp. 85–94.

[2] Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y. and Al-Kabi, M. N. [2019], 'A comprehensive survey of arabic sentiment analysis', *Information Processing & Management* **56**(2), 320–342.

[3] Al-Ghaith, W. [2019], 'Developing lexicon-based algorithms and sentiment lexicon for sentiment analysis of saudi dialect tweets', *International Journal of Advanced Computer Science and Applications* **10**(11).

[4] Al-Thubaity, A., Alharbi, M., Alqahtani, S. and Aljandal, A. [2018], A saudi dialect twitter corpus for sentiment and emotion analysis, *in* '2018 21st Saudi Computer Society National Computer Conference (NCC)', IEEE, p. 1–6.

[5] Al-Twairesh, N., Al-Khalifa, H. and AlSalman, A. [2016], Arasenti: Large-scale twitter-specific arabic sentiment lexicons, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, p. 697–705.

[6] Alabrah, A., Alawadh, H. M., Okon, O. D., Meraj, T. and Rauf, H. T. [2022], 'Gulf countries' citizens' acceptance of covid-19 vaccines—a machine learning approach', *Mathematics* **10**(3), 467.

[7] Alahmary, R. M., Al-Dossari, H. Z. and Emam, A. Z. [2019], Sentiment analysis of saudi dialect using deep learning techniques, *in* '2019 International Conference on Electronics, Information, and Communication (ICEIC)', IEEE, Auckland, New Zealand, pp. 1–6.

[8] Aldayel, H. K. and Azmi, A. M. [2016], 'Arabic tweets sentiment analysis – a hybrid scheme', *Journal of Information Science* **42**(6), 782–797.

[9] AlFutamani, A. A. and Al-Baity, H. H. [2021], 'Emotional analysis of arabic saudi dialect tweets using a supervised learning approach', *Intelligent Automation & Soft Computing* **29**(1), 89–109.

[10] Alhumoud, S., Albuhairi, T. and Altuwaijri, M. [2015], Arabic sentiment analysis using weka a hybrid learning approach, *in* 'Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management', SCITEPRESS - Science and Technology Publications, pp. 402–408.

[11] Alhuri, L. A., Aljohani, H. R., Almutairi, R. M. and Haron, F. [2020], Sentiment analysis of covid-19 on saudi trending hashtags using recurrent neural network, *in* '2020 13th International Conference on Developments in eSystems Engineering (DeSE)', IEEE, p. 299–304.

[12] Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., Aljaafary, S. K. and Alshamrani, F. M. [2020], 'A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in saudi arabia', *International Journal of Environmental Research and Public Health* **18**(1), 218.

[13] Alomari, E., Mehmood, R. and Katib, I. [2020], Sentiment analysis of arabic tweets for road traffic congestion and event detection, *in* 'Smart Infrastructure and Applications', EAI/Springer Innovations in Communication and Computing, Springer International Publishing, p. 37–54.

[14] AlYami, R. and AlZaidy, R. [2020], Arabic dialect identification in social media, *in* '2020 3rd International Conference on Computer Applications Information Security (ICCAIS)', IEEE, p. 1–2.

[15] Amurta [n.d.], 'The art of data mining for turning data into insights', Retrieved May 8, 2023, from `https://www.amurta.com/blogs/the-art-of-data-mining-for-turning-data-into-insights/`.

[16] Assiri, A., Emam, A. and Al-Dossari, H. [2018], 'Towards enhancement of a lexicon-based approach for saudi dialect sentiment analysis', *Journal of Information Science* **44**(2), 184–202.

[17] Ben-Hur, A. and Weston, J. [2010], A user's guide to support vector machines, *in* 'Springer', Vol. 609, pp. 223–239.

[18] Bramer, M. [2007], Measuring the performance of a classifier, *in* 'Principles of Data Mining', Springer, pp. 173–185.

[19] Centers for Disease Control and Prevention [2023], 'Sars-cov-2 variant classifications and definitions'.
URL: *https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html*

[20] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. [2020], 'A comprehensive survey on support vector machine classification: Applications, challenges and trends', *Neurocomputing* **408**, 189–215.

[21] Charbuty, B. and Abdulazeez, A. [2021], 'Classification based on decision tree algorithm for machine learning', *Journal of Applied Science and Technology Trends* **2**(01), 20–28.

[22] Chowdhury, S. M. and Hoque, M. [2019], A review paper on comparison of different algorithm used in text summarization, *in* 'Intelligent Data Communication Technologies and Internet of Things: ICICI 2019', Vol. 38, p. 114.

[23] Contreras-Valdes, A., Amezquita-Sanchez, J. P., Granados-Lieberman, D. and Valtierra-Rodriguez, M. [2020], 'Predictive data mining techniques for fault diagnosis of electric equipment: A review', *Applied Sciences* **10**(3), 950.

[24] Datareportal [2023], 'Digital 2023 saudi arabia', Retrieved April 25, 2023, from `https://datareportal.com/reports/digital-2023-saudi-arabia`.

[25] Domingos, P. [2012], 'A few useful things to know about machine learning', *Communications of the ACM* **55**(10), 78–87.

[26] Dreiseitl, S. and Ohno-Machado, L. [2002], 'Logistic regression and artificial neural network classification models: A methodology review', *Journal of Biomedical Informatics* **35**(5–6), 352–359.

[27] Fan, S. [2018], 'Understanding the mathematics behind support vector machines'.
URL: *https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/*

[28] Google [2023], 'Google sheets', `https://workspace.google.com/intl/en_ie/products/sheets/`.

[29] Gosavi, A. [2008], 'Reinforcement learning: A tutorial survey and recent advances', *INFORMS Journal on Computing* **21**(2), 178–192.

[30] Gupta, V. and Lehal, G. S. [2009], 'A survey of text mining techniques and applications', *Journal of Emerging Technologies in Web Intelligence* **1**(1), 60–76.

[31] Han, J., Kamber, M. and Pei, J. [2012], *Data Mining: Concepts and Techniques (3rd ed.)*, Morgan Kaufmann Publishers.

[32] Hemakala, K. and Santhoshkumar, M. [2018], 'A survey of sentiment analysis evaluation metrics', *Sensors* **18**(7), 4550.

[33] Hotho, A., Nürnberger, A. and Paass, G. [2005], 'A brief survey of text mining', *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* **20**, 19–62.

[34] IBM [n.d.], 'What is natural language processing?', Retrieved May 12, 2023, from `https://www.ibm.com/topics/natural-language-processing`.

[35] Jo, T. [2018], *Text Mining: Concepts, Implementation, and Big Data Challenge*, Springer Science+Business Media.

[36] Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R. and Mora, H. [2020], 'A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making', *Industrial Marketing Management* **90**, 523–537.
URL: *https://www.sciencedirect.com/science/article/pii/S0019850118307612*

[37] Kingdom of Saudi Arabia Ministry of Health [2021], 'Saudi ministry of health launches national campaign to raise awareness about influenza'.
URL: *https://www.moh.gov.sa/Ministry/MediaCenter/News/Pages/NEWS-2012-11-04-002.aspx*

[38] KNIME [2023], 'Knime analytics platform', `https://www.knime.com/knime-analytics-platform`.

[39] Kuhlen, R. [1992], *A morphological approach to language processing*.
URL: *https://arxiv.org/abs/2203.08909*

[40] Lavanya, P. and Sasikala, E. [2021], Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey, *in* '2021 3rd International Conference on Signal Processing and Communication (ICPSC)'.

[41] Liu, B. [2012], *Sentiment Analysis and Opinion Mining*, Vol. 5 of *Synthesis Lectures on Human Language Technologies*.

[42] Liu, T. Y. [2009], Learning to rank: From pairwise approach to listwise approach, *in* 'Proceedings of the 24th International Conference on Machine Learning', pp. 129–136.

[43] Mahesh, B. [2020], 'Machine learning algorithms - a review', *International Journal of Science and Research (IJSR)* **9**, 381–386.

[44] Maimon, O. and Rokach, L. [2005], Data mining and knowledge discovery handbook || support vector machines, *in* '10.1007/b107408(Chapter 12)', p. 257–276.

[45] Microsoft [2023], 'Python in visual studio code', `https://code.visualstudio.com/docs/languages/python`.

[46] Moghaddam, S. [2015], Beyond sentiment analysis: Mining defects and improvements from customer feedback, *in* 'Advances in Information Retrieval. ECIR 2015.', Vol. 9022 of *Lecture Notes in Computer Science*, Springer, p. 44.

[47] Mohataher, M. [2023], 'Arabic stop words', `https://github.com/mohataher/arabic-stop-words`.

[48] Mohbey, K. K. [2019], 'Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset', *Journal of Data, Information and Management* .

[49] Mohotti, W. and Premaratne, S. [2017], 'Analysing sri lankan lifestyles with data mining: Two case studies of education and health', *Kelaniya Journal of Management* **6**(1), 10.4038/kjm.v6i1.7523.

[50] Osman, A. S. [2019], 'Data mining techniques: Review', *International Journal of Data Science Research* **2**(1).
URL: *http://ojs.mediu.edu.my/index.php/IJDSR/article/view/1841*

[51] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W. and Yu, S. [2022], 'A survey on deep learning for textual emotion analysis in social networks', *Digital Communications and Networks* **8**(5), 745–762.

[52] Phillips, T. and Abdulla, W. [2021], 'Developing a new ensemble approach with multi-class svms for manuka honey quality classification', *Applied Soft Computing* **111**, 107710.

[53] Pranckevičius, T. and Marcinkevičius, V. [2017], 'Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression

classifiers for text reviews classification', *Baltic Journal of Modern Computing* **5**(2), 221.

[54] Python Software Foundation [2023], 'About python', `https://www.python.org/about/`.

[55] Rahman, S. A. E., AlOtaibi, F. A. and AlShehri, W. A. [2019], Sentiment analysis of twitter data, *in* '2019 International Conference on Computer and Information Sciences (ICCIS)', IEEE, p. 1–4.

[56] Rajput, A. [2020], Natural language processing, sentiment analysis, and clinical analytics, *in* 'Innovation in Health Informatics', Elsevier, p. 79–97.

[57] Rustam, F., Ashraf, I., Mehmood, A., Ullah, S. and Choi, G. [2019], 'Tweets classification on the base of sentiments for us airline companies', *Entropy* **21**(11), 1078.

[58] Saad, S. and Saberi, B. [2017], 'Sentiment analysis or opinion mining: A review', *International Journal on Advanced Science, Engineering and Information Technology* **7**, 1660.

[59] Sailunaz, K. and Alhajj, R. [2019], 'Emotion and sentiment analysis from twitter text', *Journal of Computational Science* **36**, 101003.

[60] sari, B. A., Alkhaldi, R., Alsaffar, D., Alkhaldi, T., Almaymuni, H., Alnaim, N., Alghamdi, N. and Olatunji, S. O. [2022], 'Sentiment analysis for cruises in saudi arabia on social media platforms using machine learning algorithms', *Journal of Big Data* **9**(1), 21.

[61] Saudi Ministry of Health (@SaudiMOH) [2020], '[tweet in arabic]'.
URL: *https://twitter.com/SaudiMOH/status/1234523092581523457?lang=ar*

[62] Saudi Press Agency [2020], '(at: 30. september 2023)'.
URL: *https://stgcdn.spa.gov.sa/viewfullstory.php?lang=arnewsid=2168181*

[63] Saudi Press Agency [2021*a*], '(at: 30. september 2023)'.
URL: *https://stgcdn.spa.gov.sa/viewfullstory.php?lang=arnewsid=2301159*

[64] Saudi Press Agency [2021*b*], '(at: 30. september 2023)'.
URL: *https://stgcdn.spa.gov.sa/viewfullstory.php?lang=arnewsid=2244988*

[65] Saudi Press Agency [2022], '(at: 30. september 2023)'.
URL: *https://www.spa.gov.sa/2334814*

[66] *saudiEAR* [n.d.], https://github.com/d7o-ae/saudiEAR/.

[67] Schmidt, M., Roux, N. L. and Bach, F. [2017], 'Minimizing finite sums with the stochastic average gradient', *Mathematical Programming* **162**(1-2), 83–112.

[68] Shah, K., Patel, H., Sanghvi, D. and Shah, M. [2020], 'A comparative analysis of logistic regression, random forest and knn models for the text classification', *Augmented Human Research* **5**(1), 12–.

[69] Singh, R., Singh, R. and Bhatia, A. [2018], 'Sentiment analysis using machine learning techniques to predict outbreaks and epidemics'.

[70] Siolas, G. and d'Alche Buc, F. [2000], Support vector machines based on a semantic kernel for text categorization, *in* 'Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000', p. 205–209 vol.5.

[71] Statista [2023], 'Leading social networks worldwide as of january 2023, ranked by number of active users (in millions) [graph]', In Statista. Retrieved April 25, 2023, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[72] Tweepy [2023], 'Tweepy: A python library for accessing the twitter api', https://www.tweepy.org/.

[73] Twitter [2023], 'Twitter api for academic research', https://developer.twitter.com/en/products/twitter-api/academic-research.

[74] Verma, R. [2012], 'Sentiment analysis: A survey on design framework, applications and future scopes', *Computational Intelligence* **28**(2), 188–221.

[75] Wang, X. and Grimson, E. [2007], Spatial latent dirichlet allocation, *in* 'Proceedings of Neural Information Processing Systems Conference (NIPS)'. **URL:** *https://papers.nips.cc/paper/3131-spatial-latent-dirichlet-allocation.pdf*

[76] World Health Organization [2020], 'Archived: Who timeline - covid-19'. **URL:** *https://www.who.int/news/item/27-04-2020-who-timeline—covid-19*

[77] World Health Organization [2023], 'Coronavirus disease (covid-19) pandemic - world health organization (who)'. **URL:** *https://www.who.int/europe/emergencies/situations/covid-19*

[78] Zimbra, D., Abbasi, A., Zeng, D. and Chen, H. [2018], 'The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation', *ACM Transactions on Management Information Systems* **9**(2), 1–29.

نموذج محسّن لتحليل المشاعر باستخدام تعلم الآلة في اللهجة السعودية: تطبيق على تطعيمات كوفيد-١٩

عبدالرحمن أسامة حلمي مصطفى

بحث مقدم لنيل درجة الماجستير في تقنية المعلومات

إشراف

أ . د طارق محمد أحمد

كلية الحاسبات وتقنية المعلومات
جامعة الملك عبدالعزيز
جدة، المملكة العربية السعودية
ذو القعدة ١٤٤٥ هـ - مايو ٢٠٢٤ م


بسم الله الرحمن الرحيم

إهداء

أُهدي هذه الأطْروحة إلى والدي العزيز، الذي طالما كان الُمعلَم الأول والمربي المبجّل. وإلى أمي الغالية، التي لم يفارق اسمي لسانها في دعائِها، والتي كانت تُشجعني على إنجازِ هذا العمل. و إلى رفيقتي في الحياة، زوجتي العزيزة، التي كانت ولا زالت ترافِقني في رحلتي. و إلى (نور) حياتي، ابنتي التي شاركتنا هذه الرحلة جنيناً حتى بدأت تَنطِق اسمي.

شُكراً لكم.

المستخلص

يُعَد تحليل الآراء (SA) وتحليل المشاعر (EA) من مجالات البحث العملية التي تهدف إلى الكشف التلقائي والتعرف على المشاعر التي تعبر عنها النصوص، وتحديد الآراء الضمنية تجاه موضوع مُعيّن. وبالرغم من إمكانية استخدام أي من هذين المصطلحين بدلا من الآخر في معظم الأحيان، إلا أن هناك اختلافات طفيفة بينهما. الهدف الرئيسي من تحليل الآراء هو تحديد القطبية التي يعبر عنها نص ما من خلال التمييز بين الآراء الإيجابية والسلبية والمحايدة. ويتعلق تحليل المشاعر باكتشاف المزيد من تصنيفات المشاعر، مثل السعادة والغضب والحزن والخوف. يساعد تحليل المشاعر على استخلاص نتائج أكثر دقة وتفصيلا وأكثر ملاءمة للمجال الذي يُطبّق فيه.

يتناول هذا العمل بشكل معمق تحليل المشاعر في إطار اللهجة العربية السعودية، مع التركيز على المشاعر المتعلقة بحملات التطعيم ضد وباء كوفيد-١٩. يعالج هذا البحث الحاجة المتفاقمة إلى المزيد من البحوث التي تهدف إلى تطوير نموذج تعلم آلي فعال لتحليل المشاعر في النصوص المكتوبة باللهجة السعودية، وبخاصة في مجال الرعاية الصحية والتطعيمات، وذلك في ظل غياب قاعدة بيانات لتحليل المشاعر المتضمَنة في التغريدات الموسومة. باستخدام منهج بحث منظم، قمنا بإعداد قاعدة بيانات مكونة من ٣٣٣٧٣ تغريدة، وتصنيفها ومعالجتها بشكل أولي. ساعد استخدام ست وثلاثين تجربة تعلم آلي تشمل آلة المتجهات الداعمة (SVM) والانحدار اللوجستي ونماذج شجرة اتخاذ القرارات وثلاث تقنيات لتجذير الكلمات وأربع طرق لاستخلاص

الملامح، على تعزيز فهم الآراء العامة المتعلقة بحملات التطعيم ضد وباء كوفيد-١٩. حقق نموذج الانحدار اللوجستي الذي استخدمناه نسبة دقة قدرها ٧٤.٩٥٪. أظهرت نتائج البحث مشاعر إيجابية في الغالب، وبخاصة السعادة، لدى المواطنين السعوديين. يساهم هذا البحث برؤى معمقة قيمة فيما يتعلق بمجال الرعاية الصحية ورصد الآراء العامة واتخاذ القرارات، كما يقترح اتجاهات مستقبلية لتحسين أداء النماذج واستكشاف تطبيقات أوسع نطاقا تتعلق باللغات واللهجات.

الكلمات المفتاحية : التنقيب عن البيانات، معالجة اللغة الطبيعية، تحليل الآراء، تحليل المشاعر، تعلم الآلة، آلة دعم المتجهات، الانحدار اللوجستي، شجرة اتخاذ القرارات، كوفيد-١٩.