UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DEVELOPMENT OF A SYSTEM FOR ANALYSIS OF THE EFFECT OF VOCABULARIES ON THE EFFECTIVENESS OF RADICAL TEXT DETECTION.

DHIRAJ TEJWANI JETHANI
ENERO 2022

TRABAJO DE FIN DE GRADO

Título:	Desarrollo de un sistema para el análisis del efecto de los vocabularios en la eficacia de la detección de texto radical.
Título (inglés):	Development of a System for Analysis of the Effect of Vocabularies on the Effectiveness of Radical Text Detection.
Autor:	Dhiraj Tejwani Jethani
Tutor:	Óscar Araque Iborra
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DEVELOPMENT OF A SYSTEM FOR ANALYSIS OF THE EFFECT OF VOCABULARIES ON THE EFFECTIVENESS OF RADICAL TEXT DETECTION.

Dhiraj Tejwani Jethani

Enero 2022

Resumen

El terrorismo moderno ha convertido a las redes sociales y las herramientas proporcionadas por internet en su principal medio para distribuir textos radicales propagandísticos, generando empatía para radicalizar a las personas y las sociedades. La radicalización basada en texto entre la población nos lleva a pensar en la importancia de las palabras elegidas conscientemente para este propósito. Las palabras tienen el poder de incitar sentimientos y emociones en el lector y también son una gran técnica de ingeniería social para aislar a individuos o grupos de personas y convencerlos de realizar acciones que puedan beneficiar al autor del texto.
El objetivo principal de este proyecto es examinar la importancia de las palabras utilizadas en artículos o revistas de índole radical de **ideología islámica**, analizando los sentimientos y emociones asociados con ellas y el papel de estos sentimientos en la detección de textos radicales. Para llevar a cabo este cometido, el trabajo se ha divido en tres partes.

La primera parte que se centra en la extracción de caracterízticas basadas en análisis de sentimientos y emociones. Creando así un vocabulario que contiene palabras claves con las que se llevará a cabo la parte de detección de radicalismo.
La segunda parte busca crear un vocabulario de palabras, en este caso enfocándose en los términos más frecuentes utilizadas en los textos, para llevar a cabo el cometido de detección de radicalismo.
Y una tarcera parte, la cual podríamos definirla como una combinación de las dos primeras, en este caso se tiene en cuenta, para la creación del vocabulario, tanto las características sentimentales y emocionales como la frecuencia con los términos aparecen en los escritos.

Finalmente estos métodos son evaluados en su eficacia en la tarea de clasificación de escritos extremistas, obteniendo unos resultados positivos, en especial el enfoque basado en la extracción de características a través del análisis de sentimientos, emociones y frecuencias tuvo un buen desempeño, pero con algunas limitaciones debido a la naturaleza de los textos usados para la evolución.

Palabras clave: Características de sentimiento, Aprendizaje automático, Emo-

ciones, Radicalización, Texto radical propagandístico, Vocabulario

Abstract

Modern terrorism has made social networks and the tools provided by the internet its spearhead for distributing propagandistic radical texts, creating empathy to radicalise people and societies. Text-based recruitment among the population leads us to think about the importance of words consciously chosen for this purpose. Words have the power to instigate feelings and emotions in the reader and are also a great social engineering technique to isolate individuals or groups of people and convince them to accomplish actions that may benefit the writer of the text.
The main objective of this project is to examine the importance of words used in radical **Islamic ideology** articles and magazines, analyzing the sentiments and emotions associated with them, and the role of these in detecting radical texts. To accomplish this task, the work has been divided into three parts.

The first part focuses on extracting features based on sentiment and emotion analysis, creating a vocabulary containing keywords that will be used in the radicalism detection part.
The second part aims to create a vocabulary of the most frequent terms used in texts, with the purpose of detecting radicalism.
And a third part, which could be defined as a combination of the first two, in this case, both sentiment and emotional characteristics as well as the frequency of the terms used in the articles are taken into account in creating the vocabulary.

Finally, these methods are evaluated for their effectiveness in classifying extremist texts, and obtaining positive results, in particular, the approach based on feature extraction through sentiment, emotion and frequency analysis had a good performance, but with some limitations due to the nature of the texts used for evaluation.

Keywords: Sentiment feature extraction, Machine Learning, Emotions, Radicalisation, Propagandistic radical text, Vocabulary

Agradecimientos

Quiero expresar mi más sincero agradecimiento al tutor de este Trabajo Fin de Grado, Oscar Aráque, por guiarme y apoyarme durante todo este proceso. Agradezco también a mis padres por su gran esfuerzo y apoyo para que pudiera llegar hasta aquí. Sé que nunca leerán este trabajo, pero quiero dedicárselo a ellos.Y por último, quiero agradecer a Marta de Luis, por su paciencia y comprensión durante los momentos de estrés y ansiedad generados por esta carrera. Su apoyo ha sido fundamental para poder completar este trabajo

Acknowledgment

I would like to express my sincere gratitude to my thesis advisor, Oscar Aráque, for guiding and supporting me throughout this process. I also want to thank my parents for their great effort and support that helped me to reach this point. I know they will never read this work, but I want to dedicate it to them. Finally, I want to thank Marta de Luis for her patience and understanding during the moments of stress and anxiety caused by this degree. Her support has been essential in completing this work.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter we will introduce the context on which this project is based, the structure followed in this work with a brief description of each of the parts and we will mention the objectives we want to achieve in this study.

1.1 Context

The rapid advancement of technology in recent years has led to unprecedented connectivity and access to information for individuals around the world. However, this proliferation of information also presents some challenges. In particular, certain organizations exploit the anonymity and confidentiality of the internet to spread radical propaganda, including that related to Islamic radicalization.

In light of this issue, our project aims to address the detection of radicalism in text. We intend to accomplish this by analyzing the words used in radical propaganda, extracting features based on emotions and sentiments, and creating a vocabulary that can assist our model in identifying radical propaganda in text.
There are three major parts in this case study.

Data analysis is the main focus of the first stage, where different dataset characteristics are analysed and extracted to enhance the model's prediction capabilities.
The model is trained using the extracted features in the second phase, and several algorithmic techniques, such as Logistic Regression, Decision Tree Classifier and Random Forest Classifier are tested.
Finally, the third phase of this study involves the evaluation of the results obtained and the drawing of conclusions based on the results.

To support this endeavour, a database has been provided to us by Araque et al. [28], it contains articles from reputable newspapers such as the New York Times (Sect.3.1.1) and CNN (Sect.3.1.1), which provide a source of Western perspective on the topic of Islamic radicalization. Additionally, we have also collected a database of articles from Islamic magazines such as Dabiq (Sect. 3.1.1), Rumiyah (Sect.3.1.1), and Al Jazeera (Sect.3.1.1).

1.2 Project goals

In this section, we will mention the objectives we want to achieve with this project:

- G1 Acquire knowledge of the fundamental concepts and principles of machine learning, and familiarize oneself with the tools and libraries that are commonly used in this field. Develop the capability to effectively handle large quantities of data and the aptitude for interpreting and analyzing the information.
- G2 Develop a model that achieves at least 80% accuracy in detecting radicalism in text.
- G3 Develop an understanding of the use of words in radical propaganda.

1.3 Structure of this document

In this section, we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 - Introduction It introduces the context, the problem we want to address, and how we plan to address it.
Chapter 2 - Enabling Technologies Describes the technologies and tools we have used throughout this project, and how we have used them and what they have been used for

within the project.

Chapter 3 - Developed method Details the methods used to accomplish our goal, including the techniques and steps we took to analyze and process the data, as well as the specific algorithms and models that we used to detect radicalism in text.

Chapter 4 - Evaluation Presents the evaluation of our model, including the results of our testing, the accuracy and performance metrics, and the comparison of our results. It also contains an analysis of the errors made by the model and recommendations for future improvements.

Chapter 5 - Conclusions Highlights the research conclusions of our study, including the results of our analyses and the efficiency of our approach in detecting radicalism in text. It also includes a discussion of the implications of our study and the opportunity for additional research in this area.

CHAPTER 2

Enabling Technologies and Related work

The project utilizes several enabling technologies to achieve its objective of examining the role of words in radical texts and detecting radicalism.
These technologies include machine learning, natural language processing, and Python data manipulation and visualization libraries. The project utilizes an open-source machine learning library for predictive data analysis to extract features from text using algorithms provided by the library. In this chapter, all these technologies will be detailed, as well as publications related to this work.

2.1 Machine Learning

Machine learning is a type of artificial intelligence that enables computer systems to learn and improve their performance, the way humans learn [20], on a specific task without being explicitly programmed.
It involves the use of algorithms and statistical models to analyze and make predictions or decisions based on data input. Machine learning has the ability to analyze large and complex data sets, identify patterns and relationships with the data, and make predictions or decisions based on those patterns.

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the model is trained on labelled data, meaning that the data is already labelled with the correct output [17]. In unsupervised learning, the model is not given any labelled data and must find patterns and relationships in the information on its own. In reinforcement learning, the model learns through trial and error, receiving rewards or penalties for specific actions.
Machine learning has a wide range of applications in many different industries, including healthcare, finance, transportation, and more. It is a rapidly growing field that has the potential to transform the way we interact with and analyze data.

2.1.1 Types of Machine Learning

1. **Supervised learning**: Involves training the model on data that has already been labelled. After learning patterns from the data, the model may then predict outcomes from new data [21]. The Model is then adjusted based on the differences between the predictions and the actual output in order to improve its accuracy. This process is repeated until a satisfactory result is achieved [21].

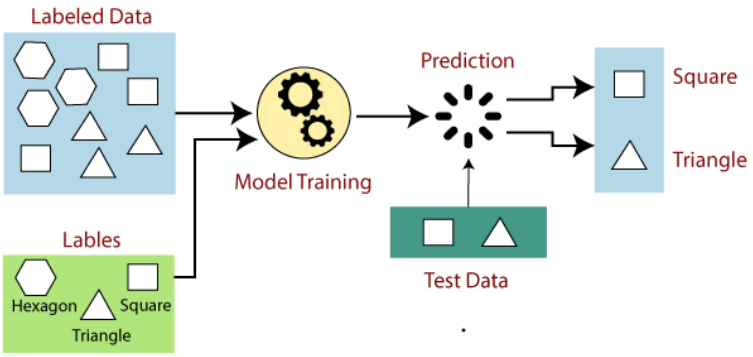

Figure 2.1: Supervised learning [18]

There are two main categories of supervised learning [14]:

- **Classification**: The goal is to predict a categorical label, the correct output is already provided, and is able to make predictions on new unseen data based on the patterns learned from the training data [36]. Different algorithms like logistic regression, decision trees and support vector machines (SVMs) can be employed for categorization. These algorithms operate by identifying patterns in training data and applying those patterns to forecast outcomes for new data.

One of the main benefits of classification is that it allows the model to make predictions about the class or category that new data belongs to.

Classification has several limitations:

(a) Only data that is similar to the training data sets may be used by the model to generate predictions. If the fresh input is significantly different from the training input, the model may not be able to make predictions accurately.

(b) Large amount of labelled data is required in order to be effective, this may be time-consuming and costly.

(c) Overfitting issues may occur, it performs well on the training data but poorly in new inputs.

The classification method is widely used despite these limitations. Some examples of how classification is used are: Spam filtering, Credit fraud detection, Medical diagnosis, Sentiment analysis, and Image classification.

- **Regression**: The objective of this method is to predict a continuous output, such as salary or probability, which is the only difference from the previously mentioned method [11].

  Linear regression, polynomial regression and decision trees are algorithms that can be used for regression. These algorithms work by finding patterns in the training input and using this to make predictions on new data.

  There are some weaknesses to regression:

  (a) Many regression algorithms, like linear regression, assume a linear relationship between the input features and the output. The model may not able to predict accurately if the relationship is nonlinear.

  (b) Significant impact can be produced by extreme values on the model's predictions if these outliers are not handled properly.

  (c) Collinearity, the correlation between input features can cause the model to be inaccurate in its predictions.

  Despite having certain limitations, regression is a popular technique that has achieved successful results in various scenarios, such as, Predicting stock prices, Evaluating performance, Predicting house prices, and Weather predictions.

2. **Unsupervised learning**: The model is provided with unlabelled training data. It is instructed to find patterns and correlations without the need for human intervention [23]. Unsupervised learning algorithms can identify unusual or anomalous data points that may not be immediately apparent to humans.

It does not have a specific goal or target, so this kind of learning may not be effective as supervised learning for a task that requires a specific output label, also it may not be able to identify all of the patterns and relationships in the input if the data is noisy or has a complex structure.

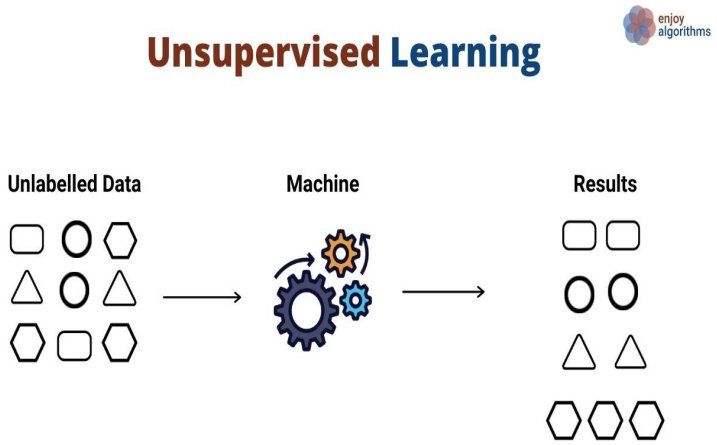

Figure 2.2: Unsupervised learning [19]

There are three main types of unsupervised learning:

- **Clustering**: The data is divided into groups or clusters [37], based on the patterns and relationships the model has found. Natural groupings in the data will appear. K-means, hierarchical clustering, and density-based clustering are examples of the various kinds of clustering algorithms.

  There are limitations to clustering [33]:

  (a) Determining the number of clusters is one of the main challenges of clustering. The number of clusters may be known a priori, but it must be determined through experimentation in many cases.

  (b) Clustering algorithms are sensitive to noise and extreme values, which can result in poor cluster assignments.

  (c) Output may not be as interpretable as supervised learning algorithms.

  Clustering has a wide range of practical applications, such as Market segmentation, Customer profiling, Text classification and Anomaly detection, are few of them.

- **Association**: It is used to identify the relationships between variables in a dataset. It is widely used in market basket analyses, where the goal is to identify

goods that are frequently purchased together [2][3]. The basic idea is to find the relationship between variables in the data that occur with a high frequency.

Few limitations to association rule learning:

(a) Large numbers of rules are created in association algorithms, which can be laborious to understand, and also can be challenging to manage.

(b) The context in which the rules were developed has not been taken into account.

(c) Association rules are typically limited to binary variables, which can limit the applicability to complex datasets.

There are various practical applications of Association algorithms like Grocery basket analysis, Fraud detection, Medical care

- **Dimensionality reduction**: In order to simplify the data for better visualisation and analysis, dimensionality reduction reduces the number of entries while retaining the most important information.

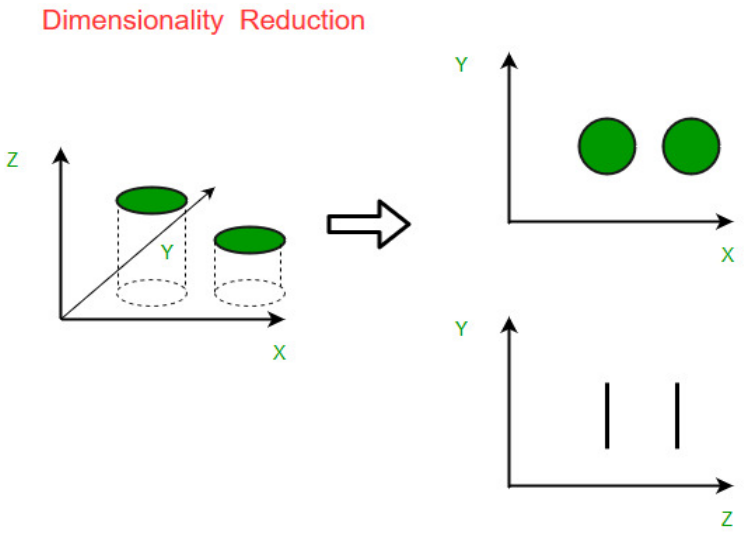

Figure 2.3: Dimensionality reduction [10]

There are several approaches that can be utilized to achieve this reduction:

(a) Principal component analysis (PCA): The data are projected into a reduced dimensional space, which maximises the variance of the data. This is helpful for identifying patterns in the data and reducing the total of dimensions while maintaining the most key information.

(b) Singular value decomposition (SVD): The approach relies on the mathematical division of a matrix into smaller matrices, which includes cutting

the information into its sub-components such as patterns and trends while retaining as much important information as possible.

(c) Independent component analysis (ICA): The ICA is based on the concept that the data is a mixture of independent sources, and involves the separation of data into components that are statistically independent of each other. The goal is to identify the underlying structure of the input data and to extract the most important information.

3. **Reinforcement learning**: With the purpose of rerouting and developing behaviour to achieve a specific objective, reinforcement learning is a technique of trial-and-error learning in which the model is rewarded for outcomes that are helpful to us and nothing if the results are not what we wanted.

   Reinforcement learning has many practical applications, such as autonomous vehicles, process optimisation and others.

   Some common examples of reinforcement learning algorithms are [44]:

   - **Q-Learn**: It is a reinforcement algorithm that provides a reward to the model for making optimal decisions, the goal of the model is to learn the relationships between decisions in order to maximise rewards. To train the model, a matrix is created in which the states, actions and rewards are stipulated. Each time the model performs an action in a certain state, the matrix is updated using the following formula 2.1.

$$Q(s,a) = (1-\alpha) * Q(s,a) + \alpha * (r + \gamma * max(Q(s',a'))) \tag{2.1}$$

   s is the current state
   a is the action taken
   r us the reward obtained
   s' is the next state
   a' is the action taken in the next stage
   α *is learning factor*
   γ *is the discount factor*

   - **Proximal Policy Optimization (PPO)**: The training is made stable and effective by using an iterative policy optimization strategy, which involves making minor adjustments at each iteration to prevent generating a significant abrupt change. In order to establish a link and maximise rewards, the model considers both past and present reward policies.

- **Actor-critic**: This algorithm consists of two neural networks as main components: the actor and the critic. The actor is a model that learns to make optimal decisions by commanding the state as input and producing an action as output. The output is evaluated and gets a reward. The critic is the other model that learns to evaluate the actor's policies. It takes as input the actor's action and the state and produces a reward value for the output. These values are used to evaluate the actor and provide feedback for decision improvement.This two neural network algorithm is optimal in environments where a direct reward is difficult to evaluate.

2.1.2 Types of algorithm

The decision-making and prediction processes are carried out by algorithms, which are the essential component of the learning task and based on which the models are developed. Depending on the type of machine learning in use and the issue to address, there are an endless number of different varieties of algorithms.

The following is a description of some of the supervised machine learning algorithms [35], which can be used for both classification and regression.

- **Logistic Regression**: A logistic or sigmoid 2.2 function, which is a mathematical function visually represented as an S-shaped curve, is the basis of logistic regression. Any value may be mapped with this function to a probability value between 0 and 1. After the logistic function is obtained, the optimal coefficients for the independent variables are looked for, so that the projected probability for each training example is as close to the actual probability as possible. Through an optimization procedure aimed at maximizing the probability of the data, these coefficients are generated.

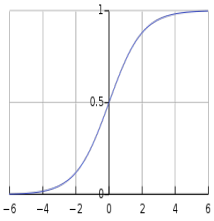

Figure 2.4: Sigmoide function [16]

$$S(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

Where x is a linear combination of the independent variables, that is, $x = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n$, where $w_0, w_1, w_2, ..., w_n$ are coefficients or weights.

The logistic function is used to model the probability of a binary dependent variable, it is calculated as the probability of the positive class divided by the sum of the probabilities of both classes, that is:

$$P(Y = 1|X) = S(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

$$P(Y = 0|X) = 1 - S(x) \tag{2.4}$$

- **K-Nearest Neighbors (K-NN)**: The K-Nearest Neighbors algorithm is a method for categorising an object based on its relationship to other objects that are similar to it. The unidentified object is assigned the most frequent class among the k surrounding objects after a brief search of them. It is only based on the measurement of the space between data points and makes no assumptions about the distribution of the data. As a result, it is adaptable to various sorts of data and versatile. It does have certain limitations, though. For example, in order to perform successfully, a significant volume of data is necessary, and the appropriate distance metric must be used for the situation.

- **Support Vector Machines (SVMs)**: The optimal discrimination between the various classes in the feature space is aimed to be achieved by the main goal of SVMs through the location of a hyperplane. The hyperplane, in a manner that causes it to maximize the margin or the distance between the hyperplane and the nearest data points from each class referred to as the support vectors, is selected. SVMs can handle high-dimensional and non-linearly separable data effectively by utilising a kernel method, and they are particularly helpful in issues when the data is not linearly separable in the original feature space. SVMs are also resistant to overfitting, especially when the data is noisy or contains a large number of outliers.

- **Decision Trees**: The core of this algorithm is the construction of a tree from the input data. Each node classifies the data based on a condition applied to a feature. The features that split the data most effectively are used to choose the conditions. Finding the decision tree that can most accurately predict the target variable from the input information is the objective. Starting at the root node, the decision tree algorithm iteratively divides the data into subsets, choosing the feature and the threshold that produces the most homogenous collection of samples for each branch. Until a stopping

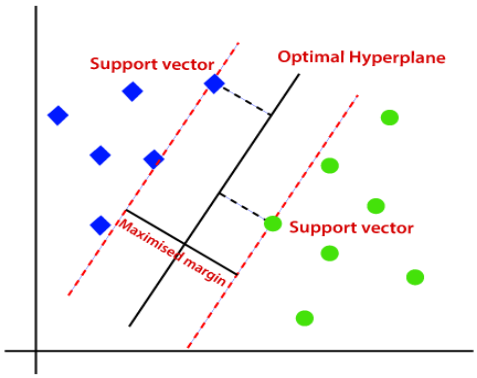

Figure 2.5: Random Forest diagram [13]

requirement, such as a maximum depth or a minimum number of samples in a leaf node, is satisfied, the procedure is repeated for each subgroup.

- **Random Forest**: The goal of Random Forest is to build plenty of decision trees during the training stage and then combine their predictions to enhance performance. The approach selects a subset of the characteristics and training data at random and then fits a decision tree to this subset. This procedure is done several times, producing the forest, which is a collection of decision trees. The predictions of each tree are then combined by taking a majority vote for classification.

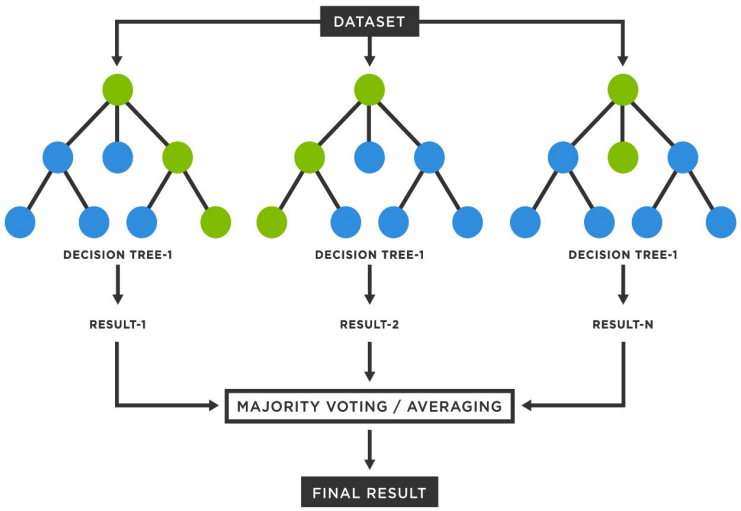

Figure 2.6: Random Forest diagram [13]

- **AdaBoost**: Several low-effective classifiers are combined to create a strong classifier. The idea behind AdaBoost is to train a set of classifiers individually, in which errors

made by the previous classifier are corrected. The aim is to combine all classifiers, with a weight equivalent to the effectiveness of their performance. AdaBoost works by adjusting the weights of each instance in each iteration. The Main advantage of AdaBoost is that it can be used with any base classifier, making it very flexible.

2.1.3 Feature extraction

The process of turning raw text data into a set of features, or attributes, that may be used to describe the text data, is known as feature extraction in natural language processing (NLP). These attributes may then be supplied to machine learning algorithms used for NLP tasks like sentiment analysis, text categorization, and others.

There are several methods that can be used for feature extraction in natural language processing (NLP):

1. **Bag of words (BoW)**: It is a straightforward and popular method that is used as a basis for language analysis and processing. The basic idea behind BoW is to represent text as a set, where each word means a value depending on the frequency at which the word appears [4]. Once the text representation is created from this technique, it is used as an input vector to machine learning models for classification.
   Here is an example of how the BoW Model can be used:

Word	Count
Hello	1
my	1
name	1
is	1
Dhiraj	1

Table 2.1: Word frequency in the sentence ”Hello, my name is Dhiraj”

2. **Term Frequency-Inverse Document Frequency (TF-IDF)**: It is a widely used method to measure the importance of an item within a collection. The basic idea

behind this method is to balance the importance of a term based on its frequency within a single document (TF) and how common it is within a collection of documents (IDF).
The equation that models this method is:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \tag{2.5}$$

- TF(t,d): is the frequency of the element 't' in the document 'd'.
- IDF(t,D): is the inverse frequency of element 't' in a collection of documents 'D'.

We can define TF(t,d) as:

$$TF(t, d) = \frac{n_t}{|d|} \tag{2.6}$$

- t: element for which the frequency is calculated.
- d: is the document
- n_t: number of times that the element 't' has appeared in the document 'd'
- |d|: total number of elements in the document 't'.

IDF can be defined as:

$$log(\frac{|D|}{df_t}) \tag{2.7}$$

- $|D|$: is the total number of documents in a collection.
- df_t: is the number of documents that contain the element 't'.

3. **Word Embeddings**: Words can be represented as continuous-valued vectors in a high-dimensional space using word embeddings. The meaning of the words and their relations to other words is captured by these vectors. These vectors can be used as input to machine learning models. There are several words embeddings methods, GloVe (Global Vectors for Word Representation)[43] is a technique that factors a global word-word co-occurrence matrix built by Standford, it is based on the belief that a word's meaning can be deduced from the context in which it appears.

   Another popular method is called FastText[30], the idea behind this method is to use sub words information to generate a meaningful vector, for example, a vector representation for "cat" can be formed by combining the character n-grams present in the sub-words("ca", "at").

2.2 Natural Language Processing

Natural language processing (NLP)[45] is one of the branches of artificial intelligence, which gives computer systems the ability to understand the text as well as a human would. NLP combines computational linguistics, statistics and machine learning to give meaning to words. Real applications of this technology are used, such as language translators, information classification, natural language generation, sentiment and emotion detection, spam detection, and more. Natural language processing goes beyond simple word translation to include meaning interpretation, taking into consideration exceptions and the context of individual words within a phrase.

The tools and methods that enable this will be examined in this section.

2.2.1 Pandas

Pandas [40] is a very popular free software Python library. This library allows us to manage, visualise and structure data. Pandas allow the import of data from any kind of formats such as CSV, JSON, SQL, Excel and others. It allows the structuring of data in one-dimensional arrays called series or in two-dimensional structures called DataFrame.

Our ability to analyse, sort, extract, filter, and visualise data is made possible by the large number of predefined functions that are included in this library.

2.2.2 NLTK

Natural Language Toolkit (NLTK)[29] is a library that provides us with a large number of tools and resources of great help and variety for natural language processing. One of the key features of the library is the number of corpora it provides, which can be used for model maintenance. For this work, it has been of great help for the preprocessing stage, since it allows the lemmatization, stemming and elimination of words with no information content.

It contains a large lexical database that provides us with synonyms and antonyms of words, very useful for language processing. It also allows the tokenization of the text, a process that consists of separating the text into phrases or words.

2.2.3 Scikit-learn

Scikit-learn[42] (also known as sklearn) is a well-known open-source Python library for machine learning. It is built on top of the NumPy and SciPy libraries for simple use. Sklearn provides a large number of classification, regression, clustering and pre-processing tools. It provides a simple API [31] that allows us to rapidly prototype models and evaluate their performance.
Scikit-learn has a large number of pre-made models, such as those mentioned in the section 2.1.2. In the Google Summer of Code programme, David Cournapeau worked on this project. Since 2017, the library has been in a stable state and continues to be widely used by practitioners and researchers in the field.

The utilization of this technology greatly facilitated the implementation of this project.

2.2.4 Matplotlib

It is an open-source library used for graphics generation and visualisation[34]. Using Matplotlib, users may have extremely fine-grained control over the visual aspects of plots, including the ability to change anything from the colour and line width of lines to the biblical text font size. It also provides an API, which gives users the option to build plots using a high-level interface.

Matplotlib is widely used in the academic and industrial fields, because of its versatility and the quality of the plots. This library is commonly used in combination with other libraries such as pandas, seaborn and others.

2.3 Performance Metrics

Performance metrics are used for the evaluation of the machine learning model. Metrics are chosen depending on the type of problem the model solves and the characteristics of the data.

Here are some of the most common performance metrics used in the field.

1. **Accuracy**: It is defined as correctly classified items over the total number of items. Precision is a simple and very intuitive metric, it is widely used as a first-instance

evaluation of a model.
On the other hand, it is not a metric that works well on unbalanced data, it can lead to false positive or false negative predictions.
In summary, precision is a widely used metric in machine learning despite its limitations.

2. **Precision**: Measure the number of correct positive predictions made by a model, divided by the total number of positive predictions made. A model with high precision is an indicator of low false positive predictions that have been made by the model, and a low precision indicates a high number of false positive predictions.
   Precision metrics are usually used in conjunction with other metrics such as recall and F1 score to get a complete evaluation of a model.

3. **Recall**: Metric to evaluate the effectiveness of a model in correctly identifying positive instances, measuring the number of correct positive predictions divided by the total number of actual positive instances.
   A high recall is a capacity to correctly identify a large percentage of positive instances, which reveals that it has a low number of false negative predictions. In contrast, a low recall indicates a high number of false negative predictions made by the model.

4. **F1-score**: It combines precision and recall to give a single measure. It is defined as:

   $$\text{F1} = \frac{2*(precision*recall)}{(precision+recall)}$$

   F1 score is especially valuable when there is an imbalance of positive and negative instances in the data. A score of 1 on this measure indicates that the model has perfect precision and recall, in contrast, a score of 0 indicates that a model has no true positive.

2.4 Related Work

In this section, we will review related work that is relevant to our proposal.

The paper by professors Oscar Araque and Carlos A. Iglesia, titled *"An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity"* [28]. has been the inspiration for this project.

This paper suggests a feature extraction method based on semantic similarity-based

vectorization and feature extraction based on the emotional lexicon.

It achieves great outcomes and argues that emotions can be used as an indicator for racialism detection.

We take a similar approach to this suggestion, in addition to extracting features based on emotions and sentiment, we also propose a frequency-based vectorization.

CHAPTER 3

Developed method

This chapter presents the methodology employed in the development of the machine learning model. It details the techniques utilized and the specific procedures that were implemented in the course of this study.
The chapter includes a comprehensive outline of the step-by-step process that was followed, the elements that were incorporated into the study, information regarding the dataset employed, and a complete analysis of the decisions that were made throughout the course of this work.

3.1 General view

The methodology we will follow for the development of our model is as follows:

First, a superficial analysis of the dataset, with the objective of looking for features that may be important for the extraction.

Secondly, a preprocessing procedure is essential in any natural language processing model. In this stage, we aim to clean the text of elements that are not useful for analysis such as

numeric characters, and punctuation marks, among others.

And finally, we reach the fundamental part of this work, the feature extraction stage, in which we exploit the features of the dataset with the different types of extraction that we propose in this study.

3.1.1 Data used

The dataset used, provided by Grupo de Sistemas Inteligentes called "Radical Magazine Corpus", consists of 468 articles from 5 different authors, including, The New York Times, CNN, Dabiq, Rumiyah y Al Jazeera.

- **The New York Times**: Widely considered one of the most reputable and influential newspapers in the world based in the United States [12]. It is known for its national and international coverage of news. The New York Times is also known for its rigorous and unbiased reporting.
- **CNN**: One of the most widely broadcasted news channel bases in the United States [5]. It is known for providing 24-hour worldwide coverage of news. CNN is considered a credible and reliable source of news and information.
- **Dabiq**: Online magazine used by terrorist groups active in the Middle East [6]. It has been used as a propaganda tool by ISIS to promote extremist, violent, and anti-Western ideologies. It was published in several languages, including English, to reach global adherents.
- **Rumiyah**: One of the Islamic State's most prominent online magazine for propaganda purposes[15]. It was used to spread extremist ideology, recruit new members and claim responsibility for terrorist attacks.
- **Al Jazeera**: A news organization based in Qatar [1], provides news and information through television broadcasts and websites. It has been criticized for airing controversial and divisive content.

The articles featured in the online magazines Dabiq, Rumiyah and Al Jazeera are considered to be a tool of radical propaganda. These publications are known for their use of misleading information to promulgate their extremist ideologies and justify their actions. On the other hand, articles found in CNN and The New York Times are considered neutral in their coverage of radicalism and radical groups.

1. **Id**: Identification number, a unique identifier assigned to a particular article.

2. **Headline**: This is the title of the article

3. **ArticleBody**: Is the main content of the article, is where the primary information is presented.

4. **Author**: The author is the newspaper, website or magazine where the article was published, there are 5 authors in this data set, Al Jazeera, CNN, Dabiq, Rumiyah, and The New York Times.

5. **Label**: This column provides information on the classification of the article, specifically whether it is considered radical (indicated by a value of 1) or neutral (indicated by a value of 0).

	id	headline	articleBody	author	label
0	3d3b60a0	In the Words of the Enemy	Attacks continue\nof its operations to other r...	Dabiq	1
1	966f6826	THE ALLIES OF AL-QĀ'IDAH IN SHĀM: PART 4	that from the nullifiers of Islam was "backing...	Dabiq	1
2	ed36209b	Wisdom	A pagan church in Finland\nleft not knowing w...	Dabiq	1
3	c5d65817	TAWHĪD AND OUR DUTY TO OUR PARENTS	{And [mention], when Luqmān said to his ...	Dabiq	1
4	dd0802a1	THE VIRTUES OF RIBĀṬ FOR THE CAUSE OF ALLAH	al-Basrī \nsaid \n(rahimahullāh) \nAl-Hasan in...	Dabiq	1
...	...	...	...	...	...
2552	fd7fad4e	Trump administration unveils new counterterror...	President Donald Trump has signed off on a new...	CNN	0
2553	4231da90	Dutch police foil terrorist plot to target 'la...	Dutch police say they have foiled a major terr...	CNN	0
2554	c1f9432f	Justice for Victims of ISIS	To the Editor:\n\nRe "14 Death Sentences in 2 ...	The New York Times	0
2555	0ab1a3a6	Video Purports to Show Tajikistan Attackers Pl...	A day after claiming its first attack in Tajik...	The New York Times	0
2556	c059c17f	A Second Chance for an Ivy League ISIS Recruit	"We've watched him closely," said Seth DuCharm...	The New York Times	0

468 rows × 5 columns

Figure 3.1: Radical Magazine Corpus Dataframe

The dataset is characterized by an imbalanced distribution, as evident in Figure 3.2, with 67.52% of the data labelled as radical, and 32.48% as neutral.

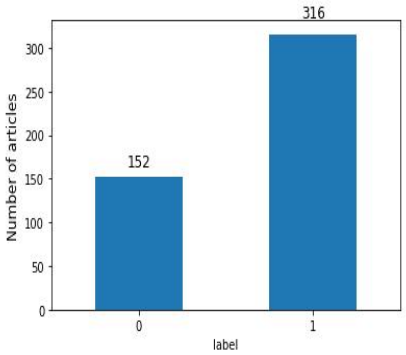

Figure 3.2: Unbalanced labelled data

An examination of the authors in the data set reveals an imbalance regard to the distribution of authors among the articles. The table 3.1 shows the number of articles written by different authors in the dataset.

Author	Number of Articles
Al Jazeera	69
CNN	129
Dabiq	126
Rumiyah	121
The New York Times	23

Table 3.1: Number of articles by Author

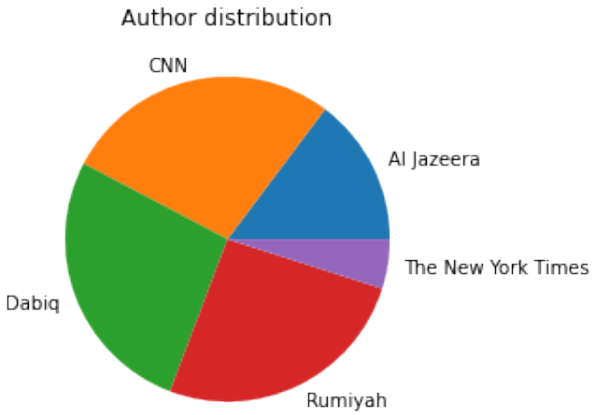

Figure 3.3: Distrubution of articles by Author

Having obtained the information from the dataset, the next step of the project will involve the preprocessing process.

3.1.2 Preprocessing

Preprocessing also referred to as data cleaning, is a crucial step in the data analysis process. It plays a key role in enhancing the performance of models and significantly improves the results.
It is important to note that preprocessing also has certain limitations, shush as data loss and a lack of interpretability in the process among others.

In this stage, we will examine and implement the techniques used for preprocessing. The goal is to ensure that the data is cleaned and prepared for further analysis or modelling.

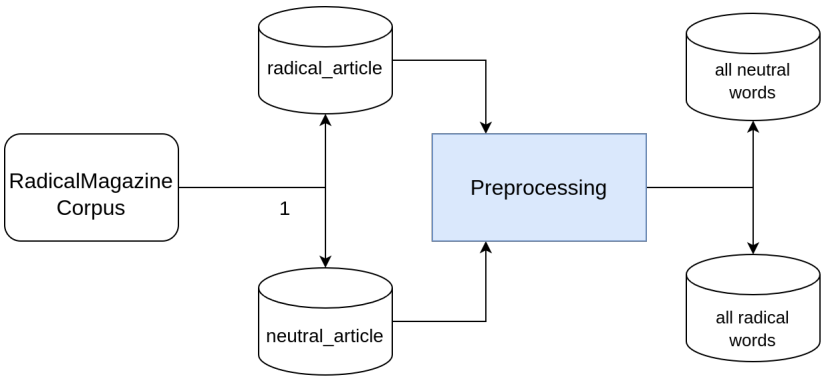

Figure 3.4: Preprocessing diagram

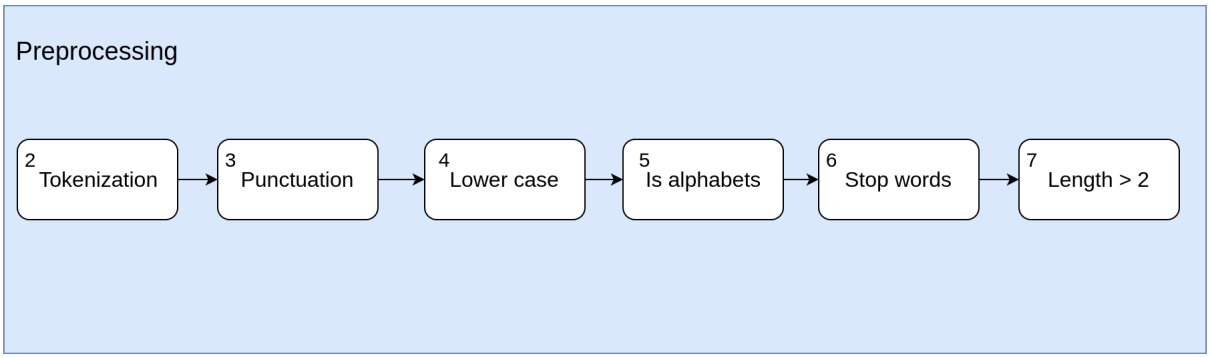

Figure 3.5: Preprocessing diagram 2

Figure 3.4 illustrates the various steps involved in preprocessing.

1. **Segmentation**: The dataset has been divided into two parts, with items classified as

radical (labelled as 1), and neutral (labelled as 0). Furthermore, the ”Headline” and ”ArticleBody” columns have been concatenated in order to cover all the words present in an article. The purpose of this is to provide a more comprehensive representation of the articles.

2. **Tokenization**: The subsequent step is the tokenization of the article, which involves breaking down the text into smaller units referred to as tokens[22]. The Natural Language Toolkit (NLTK) (Sect.2.2.2) is used to perform this task, which provides various functions to assist in breaking the text into individual words. This process allows for a more fine-grained analysis of the text.

3. **Punctuation**: Using the Python String library, it is possible to filter out all characters in the text that are not alphabetic or numerical. This is a frequently employed preprocessing step in natural language processing tasks, it allows the removal of all non-essential characters.

4. **Lowercase**: The use of lowercase words in NLP is motivated by the fact that certain models and algorithms are case-sensitive. This implies that the handling of uppercase words may not be equivalent to that of lowercase words. It is a common practice to convert all words to lowercase before further processing.

5. **Is alphabets**: This function is useful to filter non-alphabetic characters from text, such as digits. Typically digits do not provide much meaningful information and are often removed. NLP algorithms and models are not designed to handle numerical data, which can lead to errors in the analysis of the text.

6. **Stop words**: Stop words are considered to be of little value in text analysis. Examples of stop words include ”a”, ”an”, ”the”, ”and”, ”or”, ”in”, ”on”, ”of”, etc. These words are frequently present in text but do not contain much meaningful information.
   NLTK library in Python provides a list of stop words that can be used to remove stop words from the text. It reduces the dimensionality of the data and makes analysis more efficient.

7. **Length**: This step was taken after observing that the articles contained two-character words that did not provide any information, so it was decided to remove them from the texts.

8. **Unique words**: In the last stage of pre-processing, in order to reduce the amount of data and make the analysis more efficient, two collections of vocabulary are made containing words that have been used only in radical or neutral texts. In other terms,

all words that are present in both, radical and neutral texts, have been removed. The purpose behind this decision is to focus on words that are key to these texts being classified under one label or the other

As illustrated in figure 3.6, after completing the preprocessing stage, up to point 7, it can be observed that the total number of words present in the radical text is much higher than in neutral texts. This phenomenon can be attributed to the difference in the nature of the texts. The articles classified as radical tend to be more persuasive, and tend to be longer. They usually include more elaborated arguments to persuade the reader. On the other hand, articles classified as neutral tend to present facts and information in a more straightforward manner, resulting in shorter texts.

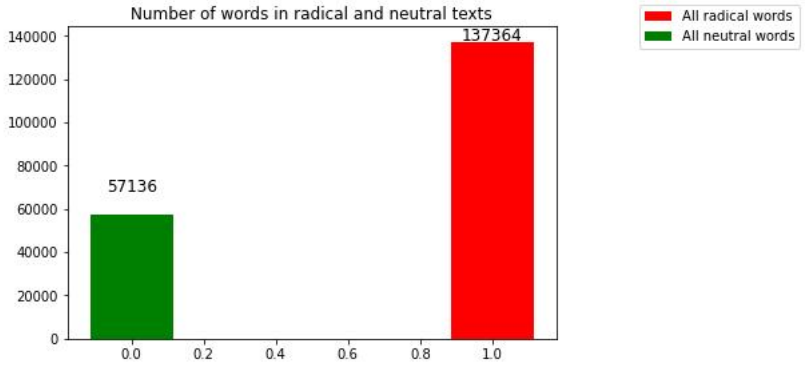

Figure 3.6: Total number of words in radical and neutral texts

After the implementation of the Unique word technique 8, it becomes apparent that the vocabulary size has been greatly reduced.

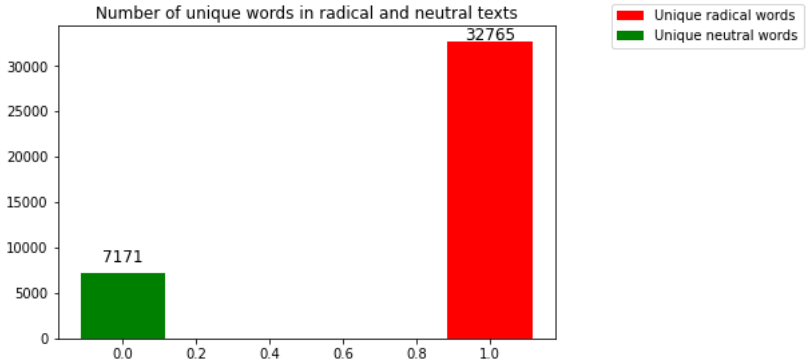

Figure 3.7: Total number of unique words in radical and neutral texts

3.2 Feature Extraction

After the preprocessing stage, the next step in the data machine learning pipeline is feature extraction. Feature extraction is the process of selecting and extracting the most relevant features from the data, as detailed in section (2.1.3), in order to improve the performance of the model.

In our case, we will focus on three types of features. The first one, which we will refer to as **EmoSelect**, will focus on the selection of features based on their emotional content The second one, named **EmoFreqSelect**, is a feature extraction method that combines both frequency-based and emotion-based selection. In contrast to EmoSelect, which only focuses on selecting features based on their emotional content, EmoFreqSelect also takes into account the frequency of occurrence of the features in the data. The third one, the **TF-IDF** method will also be employed.

The objective is to evaluate and compare the performance of the vocabulary created by our feature extraction method against the vocabulary generated by the TF-IDF method. This comparison aims to determine whether the vocabulary created by our method, which is based on the frequency of occurrence and emotional content of the words, improves the performance of the model compared to the vocabulary generated by the TF-IDF method.

3.2.1 EmoSelect

In this study is essential to establish a clear definition of what is understood as the emotions of words and what are the sentiments of words.

Pang, BO and Lee, are researchers who have made significant contributions to the field of sentiment analysis. They are best known for their work on movie review sentiment classification. The Pang, Lee and Bo Sentiment Lexicon, also known as the Opinion Lexicon, is a publicly available resource developed by Bo Pang, Lillian Lee, and Owen Rambow. The lexicon includes words and their associations with **positive** and **negative** sentiments [41].

Crowdsourcing word-emotion association lexicon is a method proposed by Saif Mohammad [39], a researcher at the National Research Council (NRC) of Canada. In the NRC Word-Emotion lexicon developed by Saif Mohammad said *"Plutchik (1962, 1980, 1994) proposes a theory with eight basic emotions. These include **joy, sadness, anger, fear, disgust, surprise trust and anticipation**"* (Saif Mohammad said, 2013).

For our study, we take into account that there are two broad categories of sentiments: **positive** and **negative**. Furthermore, there are eight basic emotions that are commonly associated with words: **anger, fear, anticipation, trust, surprise, sadness, joy, and disgust**.

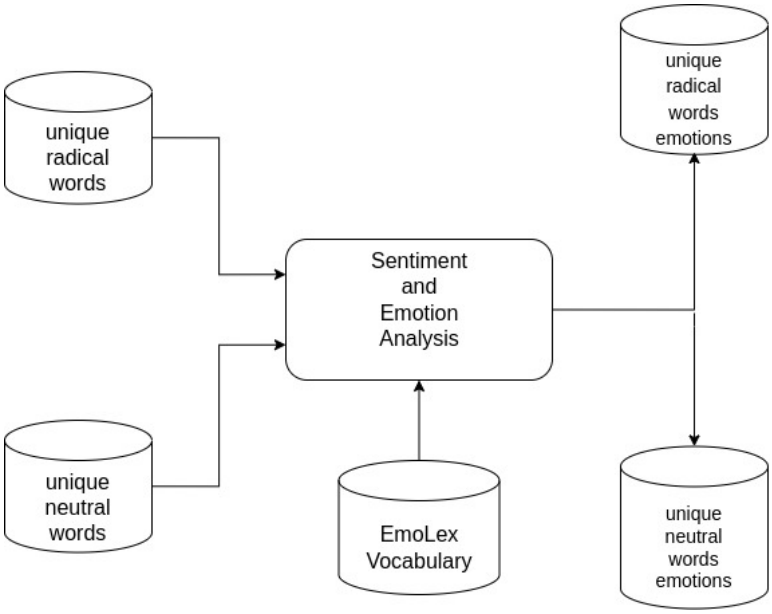

Figure 3.8: EmoSelect Diagram

This stage involves the collection of the set of words obtained from the previous process 3.1.2 and the utilization of EmoLex [39], which is a manually created list of words, to check for their association with each emotion and sentiment.

EmoLex, is a lexicon of words and their associated emotions, which was created manually by Saif Mohammad, it contains 14,182 words.

emotion	word	anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	trust
0	premier	0	0	0	0	0	0	1	0	0	0
1	tallies	0	0	0	0	0	0	0	0	0	0
2	snarling	1	0	0	0	0	1	0	0	0	0
3	superintendent	0	0	0	0	0	0	0	0	0	0
4	dogged	0	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1108	confessions	0	0	0	0	0	0	0	0	0	0
1109	seminar	0	0	0	0	0	0	0	0	0	0
1110	unborn	0	0	0	0	0	1	0	0	0	0
1111	dissimilar	0	0	0	0	0	0	0	0	0	0
1112	darkened	0	0	0	1	0	1	0	1	0	0

1113 rows × 11 columns

Figure 3.9: Dataframe of unique neutral emotion words

emotion	word	anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	trust
0	demonstrated	0	0	0	0	0	0	0	0	0	0
1	expansion	0	0	0	0	0	0	0	0	0	0
2	gravity	0	0	0	0	0	0	0	0	0	0
3	stream	0	0	0	0	0	0	0	0	0	0
4	replenish	0	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
2009	cosmetic	0	0	0	0	0	0	0	0	0	0
2010	peaked	0	0	0	0	0	0	0	0	0	0
2011	conglomerate	0	0	0	0	0	0	0	0	0	1
2012	hurting	1	0	0	1	0	1	0	1	0	0
2013	afford	0	0	0	0	0	0	1	0	0	0

2014 rows × 11 columns

Figure 3.10: Dataframe of unique radical emotion words

As can be observed in Figures 3.9 and 3.10, the size of the set of words has been significantly reduced in comparison to the previous set. This is attributed to the fact that Emolex only contains 14,182 words. As a result, any words that are not included in Emolex have been eliminated from our set.

It is possible to conduct a brief comparison of the emotions and sentiments that are prevalent in each type of text.

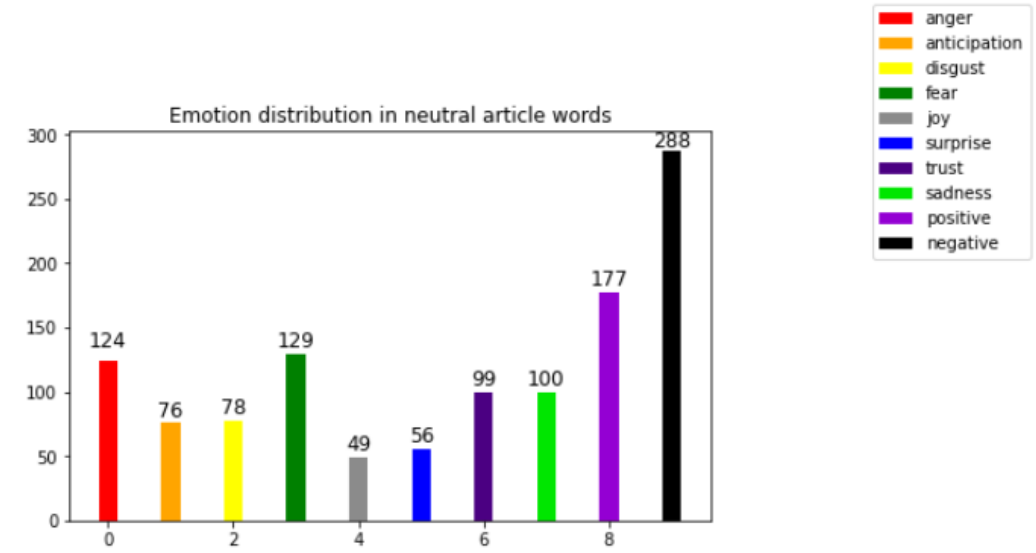

Figure 3.11: Neutral text emotion distribution

As illustrated in Figure 3.11, the most prominent emotions in neutral texts are fear, anger, and sadness. It is important to note that a word can be associated with multiple emotions.

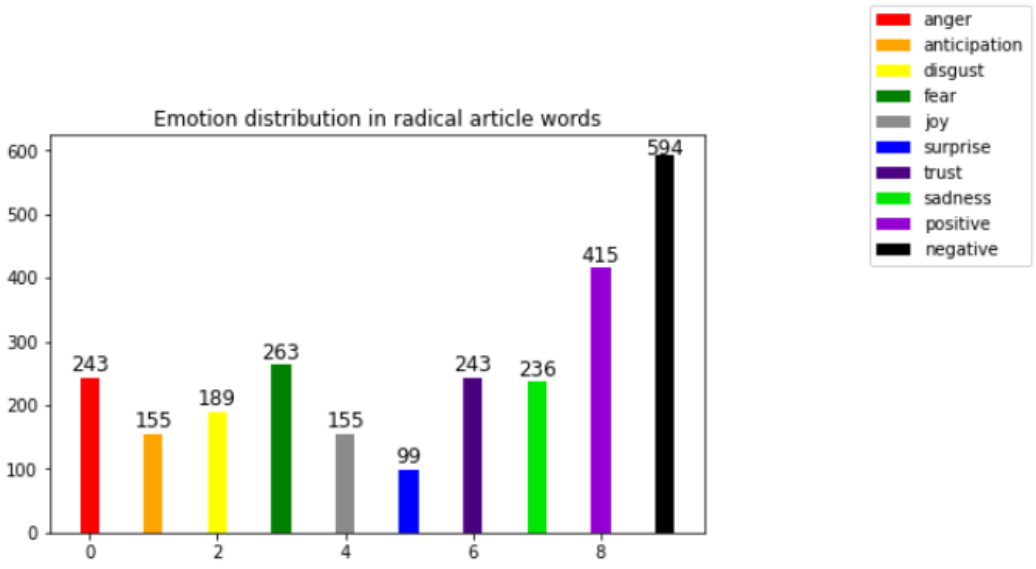

Figure 3.12: Radical text emotion distribution

On the other hand, we observe that in radical texts 3.12, the prominent emotions are fear, anger and trust.

Subsequently, we aggregated the emotions associated with each word and identified the top 50 words with the highest number of emotions present.

As can be seen in figure 3.13, we have identified the unique words associated with emotions in neutral texts. And these are the words that will make up our first set of words in the vocabulary collection.

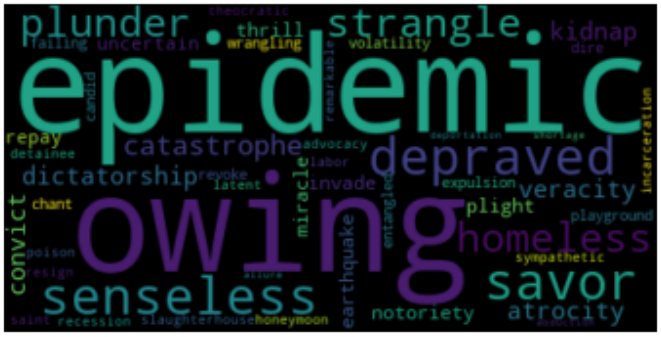

Figure 3.13: Set 1,top 50 with the highest number of emotions in neutral texts **(EmoSelect 1)**

In another hand, in figure 3.14, we have identified the unique words associated with emotions in radical texts.

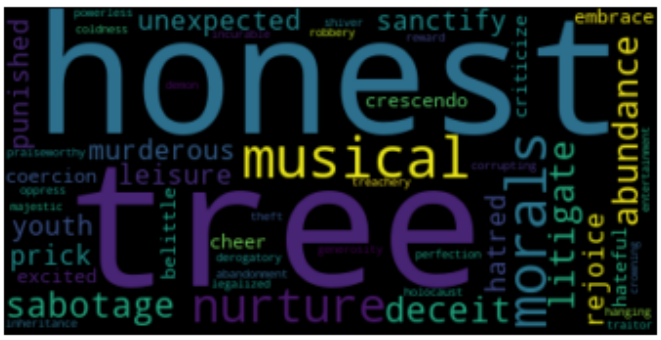

Figure 3.14: Set 2, top 50 with the highest number of emotions in radical texts **(EmoSelect 2)**

3.2.2 EmoFreqSelect

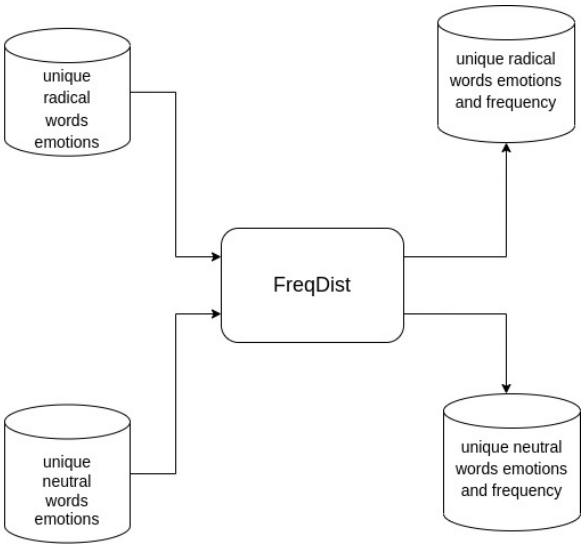

Figure 3.15: EmoFreqSelect Diagram

In this stage, we perform the same exercise as in the previous section, but in this case, we not only take into account the emotions and sentiments, but also the frequency with which the word appears in the text. To do this, we use the NLTK library's FreqDist function to identify the top 50 words with the highest frequency of occurrence and emotional content in neutral and radical texts.

We define **EmoFreqSelect 1** as the set of unique words present only in neutral texts with the highest number of emotions and high frequency in the text 3.16. We define

Figure 3.16: Set 3, top 50 with the highest number of emotions and high frequency in neutral texts **(EmoFreqSelect 1)**

EmoFreqSelect 2 as the set of unique words present only in radical texts with the highest number of emotions and high frequency in the text 3.17.

Figure 3.17: Set 4, top 50 with the highest number of emotions and high frequency in radical texts **(EmoFreqSelect 2)**

3.2.3 TF-IDF

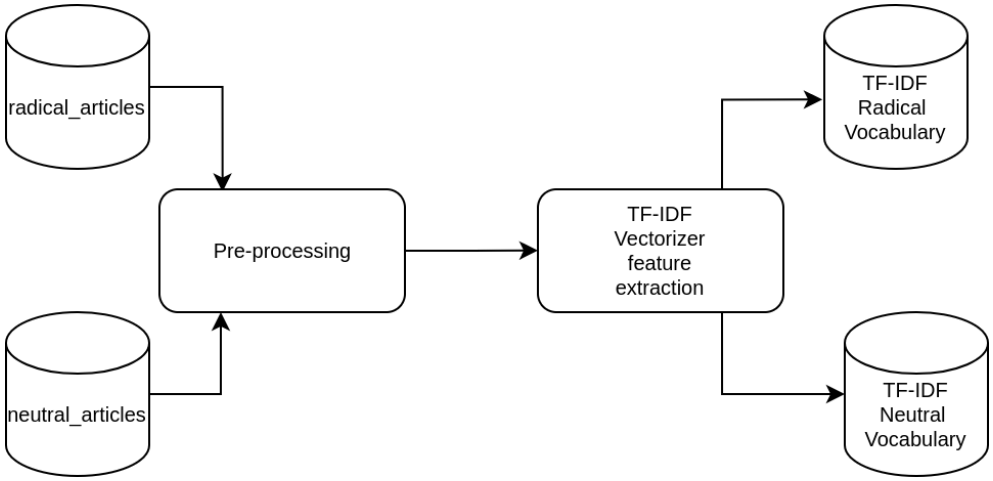

Figure 3.18: TF-IDF Diagram

In this stage, the process of feature extraction is implemented as it would typically be done in a machine learning model. Specifically, the TF-IDF [2] method is used. In the first instance, the TF-IDF method is applied only to neutral texts and subsequently, it is applied only to radical texts. We have decided to set mindf =1 and maxdf=1, this means that all terms present in at least one document will be considered and those that appear in more than one document will not be considered.

Taking this into consideration, we have obtained the following sets of vocabulary. TfIdf 1 is defined as the vocabulary set obtained by training the TF-IDF algorithm with exclusively neutral texts.

Figure 3.19: Set 5, top 50 words applying TF-IDF on neutral articles **(TfIdf 1)**

TfIdf 2 is defined as the vocabulary set obtained by training the TF-IDF algorithm with exclusively radical texts.

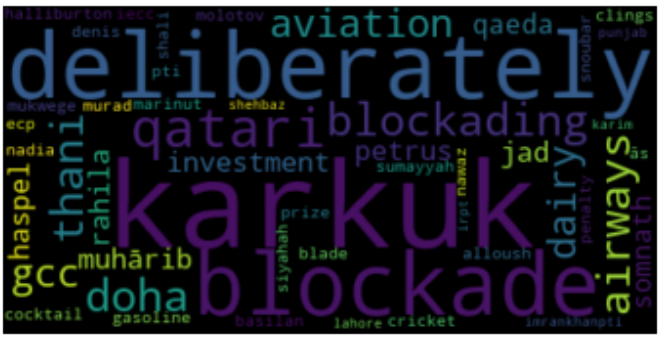

Figure 3.20: Set 6, top 50 words applying TF-IDF on neutral articles **(TfIdf 2)**

In summary, we have obtained a collection of 6 sets of words, **EmoSelect 1**, **EmoSelect 2**, **EmoFreqSelect 1**, **EmoFreqSelect 2**, **TfIdf 1** and **TfIdf 2**, which we will refer to as vocabulary. The purpose of this is to determine which option for feature extraction is the most optimal for detecting radicalism.

CHAPTER 4

Evaluation

In this chapter, we will discuss the evaluation of the vocabulary obtained in the previous chapter. We will describe the classification and regression algorithms used, and finally, we will evaluate the results obtained.

4.1 Classification and evaluation

This step is crucial for our model, as it determines whether the previous steps have been successful. To accomplish this, we train our model's classification algorithms and evaluate their precision in predicting the output labels.

In order to evaluate the effectiveness of the vocabulary obtained in the previous chapter, we have chosen to use classification algorithms 2.1.2 as there are well suited for predicting discrete outputs 4.1.

Multinomial Naive Bayes
Bernoulli Naive Bayes
Logistic Regression
Support Vector Classification
Decision Tree Classifier
K-Nearest Neighbors Classifier
Random Forest Classifier
Ada Boost Classifier
Extra Trees Classifier
Gradient Boosting Classifier

Table 4.1: Classification algorithms used

As observed in figure 4.1, these are the 10 algorithms that we will use for the evaluation process. As we can see they are all classifiers, this choice is made because the predictions to be done are of a discrete character. In order to ensure a fair evaluation, we have elected to use the default settings for the classifiers employed in this study. Adjusting the parameters of the classifier would then be assessing the performance of the algorithms rather than the effectiveness of the vocabulary.

The steps that will be followed in the evaluation can be seen in diagram 4.1:

1. Vectorize the Radical Magazine Corpus using the TF-IDF method
2. Provide to the vectorization algorithm a set of vocabulary.
3. Evaluate the effectiveness of our feature extraction model in order to determine the best set of words for detecting radicalism.

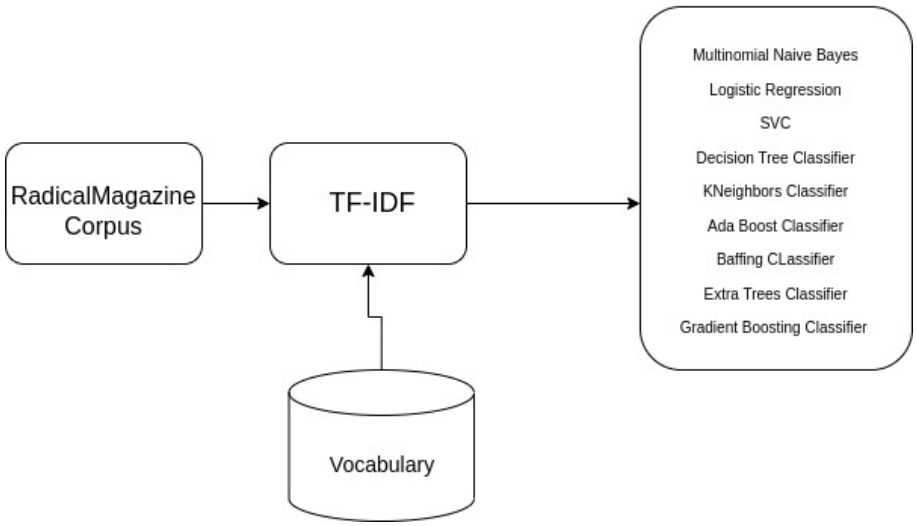

Figure 4.1: Classification diagram

4.2 Results

To evaluate the results, we will compare the EmoSelect 1, EmoFreqSelect 1, and TfIdf 1 vocabularies, as they have been defined in section 3.2 as the words that have been extracted from neutral articles. Secondly, the results of the vocabularies extracted from radical texts, EmoSelect 2, EmoFreqSelect 2, and TfIdf 2 will be evaluated.
Lastly, a global comparison will be made to determine which set is the best, regardless of the type of text from which the vocabulary has been extracted.

For the evaluation of the results, we will take into account the F1 Score, as it is ideal for imbalanced data, which is the case for our dataset.

4.2.1 EmoSelect 1, EmoFreqSelect 1 and TfIdf 1 results

As we can see in the image, the EmoSelect 1 and EmoFreqSelect 1 vocabularies, using the SVC (Support Vector Classification) algorithm, have achieved the best results in categorizing the articles. However, the same cannot be said for the TfIdf vocabulary.

When comparing the other algorithms, in general, both EmoSelect 1 and EmoFreqSelect 1 are on par. We observe that the Logistic Regression algorithm has a significant drop in performance when we remove frequency as a feature. Additionally, the MNB (Multinomial Naive Bayes) algorithm does not perform well on vocabularies with emotions.

It should be highlighted that the use of emotional features in the vocabulary improves classification accuracy by 8% when compared to vocabulary without emotional features if we compare the highest F1 Score measurements.

	Algorithm	Accuracy	Precision	F1Score
0	MNB	0.712766	0.712766	0.832298
1	BNB	0.712766	0.712766	0.832298
2	LR	0.712766	0.712766	0.832298
3	SVC	0.712766	0.712766	0.832298
6	RFC	0.702128	0.709677	0.825000
8	ETC	0.702128	0.709677	0.825000
4	DTC	0.691489	0.706522	0.817610
7	ABC	0.691489	0.706522	0.817610
9	GBC	0.691489	0.706522	0.817610
5	KNC	0.351064	0.800000	0.207792

(a) TfIdf 1 results

	Algorithm	Accuracy	Precision	F1Score
3	SVC	0.840426	0.817073	0.899329
2	LR	0.808511	0.788235	0.881579
8	ETC	0.787234	0.770115	0.870130
1	BNB	0.765957	0.752809	0.858974
4	DTC	0.765957	0.752809	0.858974
6	RFC	0.765957	0.752809	0.858974
7	ABC	0.765957	0.752809	0.858974
9	GBC	0.765957	0.752809	0.858974
5	KNC	0.755319	0.744444	0.853503
0	MNB	0.723404	0.720430	0.837500

(b) EmoFreqSelect 1 results

	Algorithm	Accuracy	Precision	F1Score
3	SVC	0.840426	0.817073	0.899329
8	ETC	0.787234	0.770115	0.870130
4	DTC	0.776596	0.761364	0.864516
6	RFC	0.776596	0.761364	0.864516
7	ABC	0.776596	0.761364	0.864516
1	BNB	0.765957	0.752809	0.858974
9	GBC	0.765957	0.752809	0.858974
5	KNC	0.755319	0.744444	0.853503
0	MNB	0.723404	0.720430	0.837500
2	LR	0.723404	0.720430	0.837500

(c) EmoSelect 1 results

Figure 4.2: Neutral text vocabularies results

We see that the Logistic Regression algorithm works well on vocabularies with frequency features, this may be due to the fact that there is a linear correlation between word frequency and the nature of the text. Looking at these results, we can affirm that the SVC algorithm works very well on emotional inputs.

4.2.2 EmoSelect 2, EmoFreqSelect 2 and TfIdf 2 results

In this case, the situation changes, the TfIdf 2 vocabulary improves the classification compared to the other vocabularies. In this situation, the Decision Tree Classifier and the Random Forest Classifier have performed the best.

This is particularly interesting, as it suggests that the TfIdf approach, which is based on the importance frequency of words in a corpus, is more effective for classifying text.

	Algorithm	Accuracy	Precision	F1Score
4	DTC	0.797872	0.792683	0.872483
6	RFC	0.797872	0.792683	0.872483
8	ETC	0.797872	0.792683	0.872483
2	LR	0.776596	0.761364	0.864516
7	ABC	0.776596	0.773810	0.860927
0	MNB	0.765957	0.752809	0.858974
5	KNC	0.765957	0.758621	0.857143
9	GBC	0.765957	0.764706	0.855263
1	BNB	0.755319	0.755814	0.849673
3	SVC	0.712766	0.712766	0.832298

(a) TfIdf 2 results

	Algorithm	Accuracy	Precision	F1Score
2	LR	0.712766	0.712766	0.832298
9	GBC	0.712766	0.712766	0.832298
0	MNB	0.606383	0.678571	0.754967
3	SVC	0.617021	1.000000	0.632653
6	RFC	0.617021	1.000000	0.632653
8	ETC	0.617021	1.000000	0.632653
7	ABC	0.585106	1.000000	0.589474
1	BNB	0.574468	1.000000	0.574468
4	DTC	0.563830	1.000000	0.559140
5	KNC	0.553191	1.000000	0.543478

(b) EmoFreqSelect 2 results

	Algorithm	Accuracy	Precision	F1Score
2	LR	0.712766	0.712766	0.832298
3	SVC	0.712766	0.712766	0.832298
4	DTC	0.712766	0.712766	0.832298
6	RFC	0.712766	0.712766	0.832298
7	ABC	0.712766	0.712766	0.832298
8	ETC	0.712766	0.712766	0.832298
9	GBC	0.712766	0.712766	0.832298
1	BNB	0.680851	0.703297	0.810127
0	MNB	0.670213	0.700000	0.802548
5	KNC	0.489362	1.000000	0.441860

(c) EmoSelect 2 results

Figure 4.3: Radical text vocabularies results

On this occasion, we see that there is no linear relationship between the frequency and the nature of the text.

4.2.3 EmoFreqSelect 1 and TfIdf 2 results comparisons

Upon comparison of the results of EmoFreqSelect 1 and TfIdf 2, it is evident that EmoFreqSelect 1 which was extracted from only neutral text, holds a better F1 Score when it comes to classifying the articles in our dataset.

This suggests that the vocabulary extracted from neutral texts is superior in terms of classification accuracy, and incorporating sentiment analysis in the extraction process improves its performance.

	Algorithm	Accuracy	Precision	F1Score
3	SVC	0.840426	0.817073	0.899329
2	LR	0.808511	0.788235	0.881579
8	ETC	0.787234	0.770115	0.870130
1	BNB	0.765957	0.752809	0.858974
4	DTC	0.765957	0.752809	0.858974
6	RFC	0.765957	0.752809	0.858974
7	ABC	0.765957	0.752809	0.858974
9	GBC	0.765957	0.752809	0.858974
5	KNC	0.755319	0.744444	0.853503
0	MNB	0.723404	0.720430	0.837500

(a) EmoFreqSelect 1 results

	Algorithm	Accuracy	Precision	F1Score
4	DTC	0.797872	0.792683	0.872483
6	RFC	0.797872	0.792683	0.872483
8	ETC	0.797872	0.792683	0.872483
2	LR	0.776596	0.761364	0.864516
7	ABC	0.776596	0.773810	0.860927
0	MNB	0.765957	0.752809	0.858974
5	KNC	0.765957	0.758621	0.857143
9	GBC	0.765957	0.764706	0.855263
1	BNB	0.755319	0.755814	0.849673
3	SVC	0.712766	0.712766	0.832298

(b) TFIdf 2 results

Figure 4.4: Comparison chart

CHAPTER 5

Conclusions and Future Work

5.1 Introduction

In this section, we will present the conclusions that have been drawn from our work, including the objectives that were set out at the beginning and the manner in which we would have proceeded in future research.

5.2 Conclusions

We can conclude that sentiment analysis has many limitations when analyzing a large number of words, as most sets of words do not cover all the words in the language. The set of words, Emolex used for emotional analysis contains around 14,000 words. The Oxford English Dictionary contains 600,000 words, which means that Emolex only covers approximately 2.33% of the total number of words in the English language.

In terms of sentiments, we have only been able to classify them as positive or negative, but in reality, there are grey areas, that are neither positive nor negative. In this case, it would be difficult to qualify those words.

In our study, we did not take into account the context in which the word is found, the meaning changes dramatically when the context of the word is taken into account.

Based on our research results, it cannot be concluded that sentiment analysis is the best method for detecting radicalism in text. However, it can be stated that the choice of text and vocabulary used for text vectorization can have a positive impact on the results. Specifically, the presence of words commonly found in neutral articles can improve the classification of radical texts, as words not found in emotional lexicon libraries can generally be considered neutral.

We have learned that texts of a radical nature tend to have a higher number of words compared to non-radical texts 3.11. The predominant emotions in radical texts are typically anger, fear and trust 3.12. It is interesting to note that in neutral texts, the predominant emotions are also anger and fear. Neutral texts tend to have a higher prevalence of sadness, as they often report sad news. Another thing that is very curious is that both types of texts present an enormous amount of negative sentiment.
It seems that there is a linearity between the frequency of words and the nature of the texts, if neutral vocabulary is used as input to the vectorizer.

In conclusion, while sentiment analysis has its limitations, it can still be a useful tool in detecting radicalism in text, especially when it is combined with other methods and when careful consideration is given to the choice of text and vocabulary used.

5.3 Objectives achieved

1. G1 Acquire knowledge of the fundamental concepts and principles of machine learning, and familiarize oneself with the tools and libraries that are commonly used in this field. Develop the capability to effectively handle large quantities of data and the aptitude for interpreting and analyzing the information: This goal has been achieved and exceeded, as we have developed a machine learning model based on emotion analysis, using a variety of tools and libraries, which are reflected in this document.

2. G2 Develop a model that achieves at least 80% accuracy in detecting radicalism in-text: We have achieved a model that exceeds 80% in F1 Score, thus we can consider this objective as fulfilled.

3. G3 Develop an understanding of the use of words in radical propaganda: We can assert

that this goal has not been fully completed. It is true that we have achieved a great understanding in terms of vocabulary, emotions, and sentiments used in radicalism recruitment. But we are still far from having a comprehensive understanding of the problem.

5.4 Future Work

The model might be further developed in the future by including new data and optimising the feature extraction procedure. Feature extraction based on the semantic similarity of words [28] and their context could be added. Additionally, the accuracy of the model could be improved by fine-tuning the parameters of the classification algorithms.

In addition, reading the dataset's articles and researching the most recent techniques for detecting and preventing online radicalization might help to investigate and comprehend how radicalism is utilised for recruiting online. Techniques such as network-based analysis could be used for this purpose [32].

Testing the model and comparing it to other cutting-edge models might be valuable to determine whether it performs better or worse. Such as this study [26], which develops a model for detecting radicalism in tweets.

Other future work could consist of expanding the concept to fit other linguistic and cultural systems. In the world, there are other types of religious radicalism on which we can adjust our model [24].

Integrating other data sources, like social media or online forums, into the model, will help it become more effective at detecting extreme propaganda in practical situations. The way of communicating has changed with the arrival of new social networks such as Instagram or TikTok. The integration of radical data from these sources could give a major impulse to our model.

APPENDIX A

Impact of this project

In this appendix, the impacts related to this work are presented from the economic, social, environmental and ethical perspectives.

A.1 Social impact

The social impact of this project could be significant in terms of its ability to detect and prevent the spread of radical ideologies through the internet and social media. The study aims to identify and isolate people or groups who may be in danger of radicalization by examining the language and emotions present in radical Islamic publications. This can reduce the harmful effects of radicalization on society, such as increasing extremism and violence. This research might also help create a safer and more welcoming online space for all users by offering tools for locating and eliminating extremist information online.

Furthermore, this project could also have an impact on the way society understands the relationship between emotions and violent radicalization. With the support of this information, anti-radicalization initiatives and tactics may be created, helping to prevent the radicalization of individuals from the beginning.

A.2 Economic impact

This project might have a significant economic impact if the solution created as a result of it was intended to be sold to government agencies or corporations for the identification and prevention of extremist ideas. Sales or licencing may be the sources of revenue. The approach may also be applied in a number of other fields, including social media and marketing, which would increase its economic effect.

A.3 Enviromental impact

Environmental effects may result from the classification of text using computational resources. Energy is heavily utilised by data centres, which contain the computers and other equipment required to process huge amounts of data. Since the project in our scenario was conducted by using personal computers instead of data centres, the environmental impact may probably be less serious.

A.4 Ethical impact

The project focuses on detecting the spread of radical Islamic ideologies, this could raise ethical concerns related to freedom of speech and the targeting of specific groups or communities. The use of data and words from a particular religious group could lead to stigmatization and discrimination of that religious group. It could promote prejudice and misconceptions about devotees of that religion.

In addition, there is concern that the project would result in populations who are currently marginalised becoming even more excluded.

The project analysis methodologies of fairness and correctness are other topics of dispute. Since sentiment and emotion analysis is a young area, there are still many unresolved problems regarding the precision and reliability of these techniques. The possibility of bias in the analysis exists.

APPENDIX B

Economic budget

This appendix gives an estimate of the project development costs, including the necessary resources, materials, and expenses. The expenses for the project can be divided into three main categories, Salaries of physical resources, Equipment and materials, and Software licenses.

B.1 Salaries of Physical resources

The total amount of hours that should be put into this project would be 300 hours provided that the final degree project has 12 credits ECTS and that each ECTS credit is equivalent to 25[8]hours of work (12 credits x 25 hours per credit) An engineer specialized in machine learning can cost around 30,000[9] euros gross per year. There are around 1,800 working hours in a year [25], assuming a 40-hour workweek. As a result, if a machine learning engineer cost 30,000 euros gross a year, their hourly salary would be around 16.67 euros (30,000/1800).

The amount is around 5,001 euros when we multiply 300 hours by the hourly rate of 16.67 euros. The project is anticipated to take at least 300 hours of labour, which, assuming a 40-hour workweek, translates to around 7.5 weeks of work.

B.2 Equipment and materials

In our specific case, the cost of equipment and materials is relatively low as only a computer and a storage device will be needed.

- Computer: A basic computer with 12 GB of RAM, 512 GB SSD, and a new generation processor may cost 1000 euros.
- Data storage device: Making backups of the model just requires an external hard disc, it may cost 30 euros

B.3 Sofware licenses

Since the libraries and materials used are free and easily available online, we have no expenses at this point.

B.4 Total budget

Worker	Hours worked	Hourly cost	sub total
1	300	16.67 €	5,001.00€

Table B.1: Salaries

PC	Storage Device	sub total
1,000€	30€	1,030.00€

Table B.2: Equipment

The total budget is 6,031 € at least, we have not taken into account the taxes for this budget.

Bibliography

[1] Al jazeera - wikipedia.

[2] Association rule learning - javatpoint.

[3] Association rule learning - wikipedia.

[4] Bag-of-words model. Page Version ID: 1119717306.

[5] Cnn - wikipedia, la enciclopedia libre.

[6] Dabiq (revista) - wikipedia, la enciclopedia libre.

[7] El ascenso del radicalismo hindú - ojo europeo en la radicalización.

[8] El sistema universitario español. il sistema universitario spagnolo.

[9] Experta en inteligencia artificial - empleo en it.

[10] Introduction to dimensionality reduction.

[11] Machine learning regression explained - seldon.

[12] The new york times - wikipedia, la enciclopedia libre.

[13] Random forest. Page Version ID: 1129334743.

[14] Regression vs classification in machine learning - javatpoint.

[15] Rumiyah (magazine) - wikipedia.

[16] Sigmoid function. Page Version ID: 1122727350.

[17] Supervised learning - wikipedia.

[18] Supervised machine learning - javatpoint.

[19] Supervised, unsupervised and semi-supervised learning.

[20] What is machine learning? | IBM.

[21] What is supervised learning?| IBM.

[22] What is tokenization | tokenization in nlp.

[23] What is unsupervised learning? | IBM.

[24] *Youth and violent extremism on social media: mapping the research - Alava, Séraphin, Frau-Meigs, Divina, Hassan, Ghayda - Google Libros.*

[25] horas de trabajo anuales. ¿cómo calcularlas? - asesorías.

[26] Swati Agarwal and Ashish Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, *Distributed Computing and Internet Technology*, pages 431–442, Cham, 2015. Springer International Publishing.

[27] Oscar Araque. Design and Implementation of an Event Rules Web Editor. Trabajo fin de grado, Universidad Politécnica de Madrid, ETSI Telecomunicación, July 2014.

[28] Oscar Araque and Carlos A. Iglesias. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access*, 8:17877–17891, 2020.

[29] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[30] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[31] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[32] Denzil Correa and Ashish Sureka. Solutions to detect and analyze online radicalization : A survey, 2013.

[33] Rafael Garcia-Dias, Sandra Vieira, Walter Hugo Lopez Pinaya, and Andrea Mechelli. Clustering analysis. *Machine Learning: Methods and Applications to Brain Disorders*, pages 227–247, 1 2019.

[34] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[35] Akinsola Jet and Hinmikaiye J O. Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology*, 48, 2017.

[36] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190, 11 2006.

[37] T Soni Madhulatha. An overview on clustering methods. 2:719–725, 2012.

[38] Batta Mahesh. Machine learning algorithms -a review. 01 2019.

[39] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 03 2013.

[40] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[41] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 01 2008.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[44] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

[45] Wikipedia. Procesamiento de lenguajes naturales — Wikipedia, the free encyclopedia. http://es.wikipedia.org/w/index.php?title=Procesamiento%20de%20lenguajes%20naturales&oldid=147980501. [Online; accessed 17-January-2023].