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 AND EVALUATION OF A NATURAL LANGUAGE PROCESSING SYSTEM FOR RADICAL PROPAGANDA DETECTION USING THE MORAL FOUNDATIONS THEORY AND AFFECT ANALYSIS

PATRICIA ALONSO DEL REAL
JUNIO 2023

TRABAJO DE FIN DE GRADO

Título: Desarrollo y evaluación de un sistema de Procesamiento de Lenguaje Natural para la detección de propaganda radical utilizando la Teoría de los Fundamentos Morales y Análisis Afectivo

Título (inglés): Development and evaluation of a Natural Language Processing system for radical propaganda detection using the Moral Foundations Theory and Affect Analysis

Autor: Patricia Alonso del Real

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 AND EVALUATION OF A NATURAL LANGUAGE PROCESSING SYSTEM FOR RADICAL PROPAGANDA DETECTION USING THE MORAL FOUNDATIONS THEORY, AFFECT ANALYSIS AND SEMANTIC SIMILARITY

Patricia Alonso del Real

Junio 2023

Resumen

Durante los últimos años, el mundo ha experimentado una importante polarización de opiniones e ideologías, en gran parte debido a la gran cantidad de información que ofrece Internet. Esta herramienta ha allanado el camino para que los grupos organizados terroristas dispongan de más recursos y oportunidades para difundir su discurso extremista y perjudicial para la sociedad como conjunto.

Este trabajo continúa la investigación realizada sobre la detección de propaganda radical en textos mediante Procesamiento de Lenguaje Natural (PLN), ya que el análisis de claves sociales puede ayudar a examinar, identificar y prever usuarios extremistas. El objetivo se basa en ver qué planteamientos en cuanto a selección de características del texto ayudan a la hora de identificar estas claves y así poder clasificar la información obtenida en dos categorías: radical o no radical. La eficacia de cada modelo evaluado es medida cuantitativamente y, en algunos casos, explicada de manera gráfica gracias a técnicas para explicar las predicciones de modelos de aprendizaje automático como SHapley Additive exPlanations (SHAP).

Más concretamente, se lleva a cabo una evaluación del rendimiento de un enfoque basado en fundamentos morales en combinación con indicios afectivos y una técnica basada en la similitud semántica. Esta fusión en particular de métodos de extracción de características no se ha hecho antes, por lo que los resultados aportarán información de utilidad para el campo de la detección de propaganda radical en lenguaje escrito.

El resultado de este trabajo muestra cómo la moralidad es un concepto que puede ayudar eficazmente en la tarea propuesta mediante el desarrollo de léxicos basados en la Teoría de los Fundamentos Morales. Para ello, se hace uso de dos vocabularios distintos, ambos creados con el fin de aportar información extra y valor añadido al modelo de aprendizaje automático que se esté desarrollando. Los resultados obtenidos son comparados y analizados para ver cuál es más útil en qué situación.

Palabras clave: Detección de propaganda radical, Python, aprendizaje automático, Procesamiento de Lenguaje Natural (PLN), léxico, Teoría de los Fundamentos Morales, emociones, similitud semántica.

Abstract

Over the past few years, the world has experienced a significant polarization of opinions and ideologies, largely due to the vast amount of information available on the Internet. This tool has paved the way for organized terrorist groups to have more resources and opportunities to spread their extremist and harmful discourse to society as a whole.

This paper continues the research done on detecting radical propaganda in texts using Natural Language Processing (NLP), as the analysis of social cues can help to examine, identify and predict extremist users. The objective is to see which approaches to text feature selection help in identifying these cues and thus to classify the information obtained into two categories: radical or non-radical. The effectiveness of each evaluated model is quantitatively measured and, in some cases, graphically explained thanks to techniques to explain the predictions of machine learning models such as SHapley Additive exPlanations (SHAP).

More specifically, an evaluation of the performance of an approach based on moral foundations in combination with affective cues and a technique based on semantic similarity is carried out. This particular fusion of feature extraction methods has not been done before, so the results will provide useful information for the field of radical propaganda detection in written language.

The result of this work shows how morality is a concept that can effectively help in the proposed task by developing lexicons based on the Moral Foundations Theory (MFT). For this purpose, two different vocabularies are used, both created to provide extra information and added value to the machine learning model being developed. The results obtained are compared and analyzed to see which is more useful in which situation.

Keywords: Radical propaganda detection, Python, Machine Learning, Natural Language Processing (NLP), lexicon, Moral Foundations Theory (MFT), emotions, semantic similarity.

Agradecimientos

A mis padres Sara y Juan Carlos, a los que agradezco todo lo que me aportan cada día.

A mi hermano Jaime, mi mejor amigo.

A mi abuela Mª Carmen, quien siempre me pregunta cómo llevo los estudios para que no tenga que depender nunca de nadie.

A mis amigos, los de verdad, los que se alegran de mis éxitos como si fuesen suyos, al igual que yo de los suyos como si fuesen míos.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

The proliferation of social media and the internet in recent times has accelerated the pace and lowered the expenses associated with sharing information. As a consequence, this has facilitated numerous terrorist organizations to restructure themselves in order to amplify their capabilities to operate autonomously and cause significant harm to individuals, communities, and nations [2]. The spread of extremist beliefs and ideologies can result in violent actions, hate crimes, and social unrest [3]. Therefore, the detection of radical propaganda is crucial for preventing radicalization and promoting social harmony. Numerous global organizations and nations have formulated tactics and initiatives to counteract radicalization through social and computational text analysis [4]. The task of performing thorough manual inspections is impractical with the vast amount of text and relationships between information and individuals. Therefore, the development of computational techniques for detecting, analyzing, and preventing radicalization and extremism is essential.

There are many research studies that investigate the development of efficient automated approaches for detecting extremism [5]. An important illustration of this is the present emphasis of counter-terrorism organizations and governments on automatically detecting extremist profiles on social media in their attempt to combat extremist social network groups. By developing information technology systems that can recognize extremist content it would be possible to combat online radicalization [6].

It is undeniable that there is a growing tendency to utilize data mining approaches, such as machine learning, to investigate these matters related to extremist content [7]. The efficient implementation of technology-based methods for counter-terrorism remains an ongoing problem.

Many social-ideological differences within a country may stem from variations in the definitions of morality, as suggested by evidence. According to what the authors of [8] propose, liberals and conservatives in the US have unique perspectives on the social environment and depend on differing moral frameworks and ideologies. Also, the significance of emotions in comprehending terrorism has been emphasized in recent social science literature [9, 10].

This investigation is based on three main research questions (RQ):

RQ1: Does a moral foundations approach help in the task of detecting radicalization? If so, how? Research says that using moral values to radicalize groups of people is a tendency in terrorist cells. When a group collectively agrees on certain values and morals, it can wield significant influence, including the ability to legitimize and sometimes mandate violence against individuals who are a threat to the them [11]. Thus, this work studies the effect of incorporating moral values information through the comparison of two learning models for the task of radicalization detection.

RQ2: Can moral and emotion values combined with embedding-based similarity features be used effectively for propaganda detection? As said previously, moral foundations and emotions are useful tools when classifying text to detect radicalization. Word embeddings have also been utilized alone and in combination with other approaches for the same task [12]. With this information, this work analyzes whether the effect of merging these techniques may have a positive outcome in the overall classification performance.

Given these inquiries, this work suggests a machine learning system that can identify propaganda across news, magazines and social media. For this purpose, the suggested system produces textual representations that are utilized by the machine learning classifier and specifies the utilization of the subsequent types of features: moral values, emotions, similarity-based features that employ a word embedding model, unigrams, bigrams, features from the TF-IDF method and their combinations. A thorough assessment has been conducted on three pertinent datasets in order to evaluate the efficiency of this broad range of features. Furthermore, a computational technique that utilizes word frequency distributions to derive a domain-specific vocabulary has been used (*FreqSelect*). The objective of this method is to function as a fundamental resource for capturing the lexicon of a specific domain.

1.2 Project goals

The goal of this research is the evaluation of a novel combination of Natural Language Processing (NLP) techniques in the task of classifying text into radical or non-radical. A morality-based approach is introduced along with emotion and semantic similarity-based analysis. The specific objectives in order are:

1. Use of the different approaches and feature extraction methods, and their combinations.
2. Evaluation of the performance of each combination in comparison to the established baselines.
3. Use of SHAP explainers to illustrate the most important combinations of methods and how different features have an important role in text classification.

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 presents a comprehensive summary of the entire project, including its objectives. Its aim is to explain the rationale behind the development of this work, the issues it seeks to address and the approaches that have been useful in the task of confronting them.

Chapter 2 provides relevant literature regarding the different modules that integrate the full model.

Chapter 3 shows the tools and services that have made it possible to develop this research and how they have taken part in each stage of the investigation.

Chapter 4 describes the general architecture of the system and then provides more detailed information regarding each component.

Chapter 5 presents the materials that have been used in order to carry out the work, the methodology that has been followed, and the results obtained through the several experiments done with the different models proposed.

Chapter 6 culminates the project with conclusions derived from the research itself, an overview of the objectives that have been accomplished and a discussion about how in the future this work can be continued and enhanced.

CHAPTER 2

Related work

2.1 Moral Foundation Theory

Here it is provided an overview of the work done regarding the evaluation of moral values through analysis of textual data.

Following the line of reasoning that states that perspectives coming from multiple levels of analysis should be acquired prior to understanding the social pattern of radicalization [54], the approach that this work proposes uses the study line related to morality by analyzing the performance of two different lexicons as tools for classifying the input data based on the Moral Foundations Theory [53]. This psychological theory aims to model the differences in morality across different cultures, while also highlighting the presence of resemblances and recurring patterns. According to the theory, there are certain psychological systems that are innate and universally accessible. These foundational elements are then built upon by each culture through the development of virtues, narratives, and institutions. The foundations that have been taken into account in this work are 'care/harm', 'fairness/cheating', 'loyalty/betrayal', 'authority/subversion' and 'sanctity/degradation'.

Other studies have also been developed using the Moral Foundation Theory for investigation on diverse fields. One of the most representative tool used by computation approaches is the Moral Foundations Dictionary (MFD) [8], that allows researchers to assess moral foundations by means of a lexical resource, based on the LIWC framework [78]. In [8], four studies using four different methods were carried out in order to develop a theory which states the psychological foundations upon which political groups build their moral codes. In a more recent work, the MFD was utilized to conduct a manual analysis of 12 years' worth of coverage in the New York Times, with a focus on political discussion in the United States [56]. [57] analyzed the discourse on the potential exit of Greece from the European Union through the examination of approximately 8,000 tweets related to the topic. A comparison was made between the effectiveness of basic machine learning models, such as Maximum Entropy (ME) and Naive Bayes (NB), in utilizing unprocessed MFD features. Both studies came to the conclusion that machine learning in its pure form is more desirable than dictionary-based approaches as it achieves similar predictive accuracy with fewer assumptions.

Another study [58] employed a Latent Dirichlet Allocation (LDA) model to explore the existing variations between conservative and liberal ethical codes. This facilitated the unsupervised identification of topics related to morality. In [59], the authors applied the same structure to investigate moral argumentation in text, focusing on the US Federal shutdown of 2013. The study analyzed the influence of morals on intra- and inter-community disparities in political party retweets [60]. In addition, [61] proposes Latent Semantic Analysis (LSA) for representing the moral values of text through a model that uses a multiset of words in order to calculate a co-occurrence matrix, and subsequently, word vectors are extracted from it.

There are also some recent studies where the MFD has been utilized to identify moral values in lengthy political speeches over a period of time, such as in [34]. Likewise, [62] presented a technique called Distributed Dictionary Representations (DDR), which involves merging psychological dictionaries and semantic similarity to assess the prevalence of moral rhetoric on a particular subject. This technique has been utilized in other studies with the aim of identifying morals in the donation to charity [63] and also to include demographic embeddings within the language representations by expanding it [64]. Another study where the MFD and the DDR technique are combined and the *MoralStrength* lexicon is taken into account is [77], where the moral divergence between candidates from the Republican and Democrat parties is analyzed through the quantification of presidential debaters' moral judgments. This work is related to the mediatization of opinions, which can lead to their polarization.

In this work we explore the performance of two different lexicons coming from distinct works: the *MoralStrength* [18] and *eMFDscore* [32] lexicons. The MoralStrength resource expands the original MFD. After an initial preprocessing step was conducted on the word corpus obtained, wherein forms that matched the search but did not relate to a moral trait were eliminated. This procedure was carried out manually, taking into account both the gloss for the lemma provided by WordNet [79] and the moral trait associated with that word. Then, a division was done separating the obtained word corpus in “virtue” and “vice” lemmas so that an association strength between word and moral trait can be provided with value ranges from 1 to 9 (1 for words commonly linked to vices and 9 for words commonly linked to virtues).

The MoraStrength lexicon can be used in a variety of applications, such as in [55], where the authors monitorize the responses to the mask mandate due to COVID-19 and analyze the moral values behind each argument proposed, as well as the political leaning of people with one opinion or the other. This is based on the conceptual perspective of Moral Foundation Theory and Hofstede’s cultural aspects [80].

The other resource has also been used in this work is the *eMFDscore* lexicon [32]. With this approach, every word is assigned both foundation probabilities, that indicate the likelihood that each word is linked with each one of the five moral foundations, and sentiment scores, that reflect the average sentiment of the context related to the moral foundation where each word is used.

2.2 Word embeddings and emotion lexicons for online radicalization detection

Word embeddings and emotion lexicons are NLP methods that have been previously used for detecting radicalism in natural language processing. Word embeddings represent words as vectors in high-dimensional spaces [82]. This approach can be useful for capturing the meaning and context of words within a given text corpus. On the other hand, emotion lexicons are dictionaries that label a certain vocabulary with affective dimensions, such as joy or fear [81].

Data coming from social media and magazines has been analyzed with the aim of detecting radicalization by merging both approaches [12] [66]. The obtained results ratified the potential that the combination of these techniques has in this kind of text classification as the F1 score increased in comparison to its value when only using one of the methods.

In the end, the analysis carried out in works like this one is used to enhance law enforcement agencies (LEA) in their process of making decisions. In [65] it is explained that solutions resulting from analysis can be categorized into two primary types: network-based (or link) and content-based. The first type concentrates on virtual communities and graph characteristics, while the second deals with online behavior, linguistic style analysis, authorship identification, emotion analysis and usage mining.

Regarding, emotion analysis, it has been carried out in different fields, such as radical forums [42], extremist magazines [36] and social media platforms like Twitter [24].

In relation to the affective processing, there are distinct ways in which different authors have carried it out. Several studies utilize the polarity of sentiment analysis (e.g., valence) [41] [40]. Others, the strength of the vocabulary related to hatred and aggression [42]. Besides, Linguistic Inquiry and Word Count's categories are used to utilize the information obtained from emotions of both positive and negative nature, as well as sadness, anxiety and anger.

There is an ongoing investigation work where the response of the general public to a terrorist attack is tracked [44]. The authors conducted a sentiment analysis on both texts and images that were extracted from Facebook and noticed that while the sentiment was initially negative in the first few hours, it gradually shifted toward a positive valence as time passed. This happened with text, but for images, an opposite effect took place.

In the past, it was common for many Natural Language Processing (NLP) techniques to utilize bag-of-words characteristics when representing text for classification tasks. Then, word embeddings started to be commonly employed as features for a subsequent learning system [51]. Continuing this research direction, various papers address the incorporation of word embeddings as a component of feature extraction techniques for text categorization. In [68] they are utilized as characteristics used for a SVM classifier in order to identify polarity in texts. Following the assessment, the authors arrived at the conclusion that word embeddings may encompass semantic knowledge among words, thereby enabling the acquisition of valuable text representations. In [67], semantic meanings are captured by this approach using the *Word2Vec* model to create a system that can differentiate between tweets that endorse the Islamic State of Iraq and Syria (ISIS) and those that do not. The authors of [70] examine the efficacy of utilizing word embeddings for sentiment analysis and provide a summary of unsupervised embedding methods and their potential use in obtaining text representations. In addition, methods based on this approach are widely employed in open competitions, where participants strive to get the highest scores in a wide range of tasks [72] [71] [73]. Another interesting work [74] suggests the use of word embeddings within a novel framework designed to minimize computational complexity and demonstrates comparable evaluation metrics to those of more intricate neural models across multiple tasks. Besides, in [75] a merge of semantic similarity and word embeddings approaches is presented.

Recently, [69] explored the detection of hate speech in a forum dedicated to white supremacy. The model the authors suggest has been trained and evaluated on a balanced subset of a dataset containing roughly 2,000 sentences sourced from the Stormfront forum. Approaches like SVM, LSTM and Convolutional Neural Network (CNN) were utilized in order to identify hate speech. One significant constraint of this study is that it involves annotating sentences taken from paragraphs, without any extra information that could aid in the comprehension of the context of the sentences for precise labeling.

CHAPTER 3

Enabling technologies

3.1 Python

For this work, the programming language that has been used is Python. It is high-level, object-oriented interpreted language created by Guido van Rossum and first released in 1991. It is simple and versatile. It has a huge constellation of libraries particularly designed for machine learning and scientific computing, such as Pandas[1], Numpy[2], scikit-learn[3], spacy[4], gensim[5], nltk[6], gsitk[7] or matplotlib[8]. As seen in figure 3.1, last year Python was the most popular programming language among coders, even more popular than JavaScript.

[1] https://pandas.pydata.org/

[2] https://numpy.org/

[3] https://scikit-learn.org/stable/

[4] https://spacy.io/

[5] https://radimrehurek.com/gensim/

[6] https://www.nltk.org/

[7] https://gsi.upm.es/software/projects/gsitk/

[8] https://matplotlib.org/

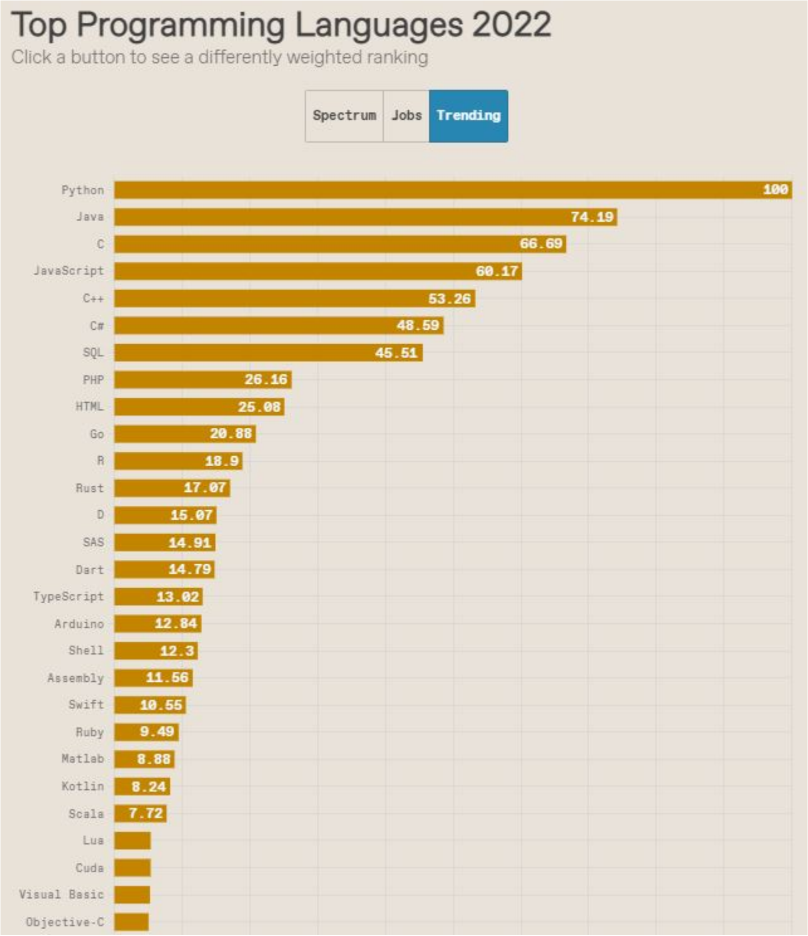


Figure 3.1: 2022's top programming languages [1].

Pandas is an open-source Python library extensively utilized for analyzing and manipulating data. It offers efficient and powerful tools for working with structured data. Some functionalities of this library are:

- Data structures: Pandas presents two main data structures (Series and DataFrame).
- Large dataset handling: By implementing mechanisms such as lazy evaluation and memory optimization, extensive datasets can be easily handled.
- Data cleaning and preprocessing
- Integration with other data analysis libraries

NumPy, which stands for Numerical Python, serves as a fundamental Python library for performing numerical computations. It equips developers with a potent array object and a comprehensive set of functions and tools to effectively manipulate arrays and matrices. Some functionalities of this library are:

- Multi-dimensional arrays: Arrays in NumPy can be one-dimensional, two-dimensional, or even higher-dimensional.
- Numerical operations: These include basic arithmetic operations, element-wise operations, linear algebra operations, Fourier transforms, random number generation...etc.
- Broadcasting: This allows performing operations between arrays of different shapes and sizes.
- Integration with other data analysis libraries

Scikit-learn is a popular Python library for machine learning that offers many tools and algorithms for many tasks, including classification, regression, clustering, and dimensionality reduction. It is built on top of other libraries in Python, such as NumPy, SciPy, and matplotlib. This library provides a diverse range of preprocessing techniques to address tasks such as data cleaning, scaling, handling missing values... etc. Regarding model selection, it has methods for cross-validation, hyperparameter tuning, and model evaluation metrics such as accuracy, precision, recall, F1-score... etc. It can be integrated with other Python libraries too.

SpaCy is an open-source library for tasks related to NLP in Python. Some features it includes are tokenization (segmenting text into words, subword units or punctuations marks called tokens), lemmatization (assigning the base forms of words, converting altered or derived words to their canonical form), Part-of-Speech (POS) tagging (assigning grammatical tags to each token, indicating their part of speech, such as noun, verb, adjective, etc), accurate sentence segmentation and identification of boundaries for a proper information extraction, word vector representations (word2vec, GloVe) for representing words as vectors of high dimensions or text classification.

Gensim is a topic modelling tool used in NLP. It is open-source and offers a comprehensive range of algorithms and tools designed for tasks like analyzing document similarity, clustering documents or preprocessing of text. Additionally, it incorporates functionality for popular word embedding models like Word2Vec and FastText (the latter used in the experiments) that allows training, loading, and using them.

Nltk, which stands for Natural Language Toolkit, is an open-source group of libraries used for NLP in Python. It provides access to over 50 corpora and lexical resources, like

WordNet. It is utilized for tasks like tokenization, stemming, tagging, parsing, and semantic reasoning.

Gsitk is a library that facilitates the workflow of projects based on NLP. It manages datasets, features, classifiers, and evaluation methods, simplifying the creation of an evaluation pipeline for quick implementation.

Matplotlib is a Python library that enables the creation of static, animated and interactive visualizations. Plots, charts and graphs can be created and customized using its functions and tools.

Jupyter Notebooks are an open source web application that enables the creation and sharing of documents incorporating live code, equations, visualizations, and narrative text. It supports Python language, among others. The notebooks consist of cells that can be executed independently, so the code enablesbe run in an incremental way and with an immediate result visualization.

JupyterHub is a multi-user version of the Jupyter Notebook environment. It runs in the cloud or on hardware.

3.2 Classifiers

The classifiers that have been utilized for making the predictions regarding the radicalness of the input texts are: Logistic Regression Classifier and Linear Support Vector Machine Classifier (Linear SVM). We will briefly describe them:

Logistic regression[9] is a supervised learning classification algorithm used to predict the probability of a target variable. The classifier based on this procedure uses the logistic or sigmoid function to transform a linear combination of input features. This function resembles an "S" shaped curve when plotted on a graph. It takes values between 0 and 1 and "squishes" them towards the margins at the top and bottom, labeling them as 0 or 1. Typically, a threshold is set to determine the value at which an example is assigned to one class or the other.

Support Vector Machine[10] is a supervised machine learning algorithm that can be used for both classification or regression challenges. It is especially efficient when the data can be linearly separated into two classes. In a linear SVM classifier, each data item is plotted in a n-dimensional space (being n the number of features) with the value of each

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[10] https://scikit-learn.org/stable/modules/svm.html

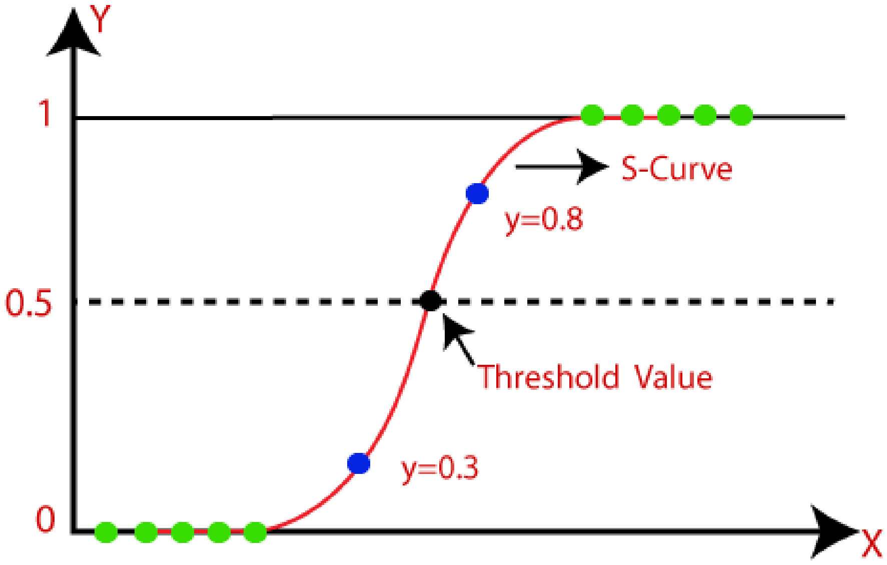


Figure 3.2: Logistic Regression classifier.

feature being the value of a particular coordinate. The algorithm aims to identify an optimal hyperplane that effectively separates the data points belonging to different classes. This classifier calculates a score that represents the instance's distance from the hyperplane, which can be positive or negative. By applying a threshold to the score, the classifier makes a binary classification decision.

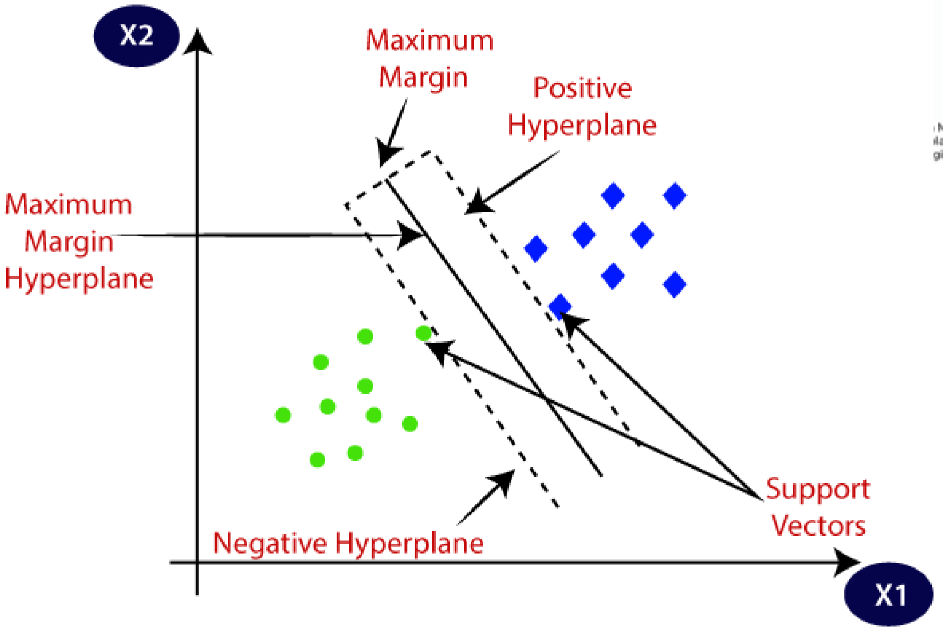


Figure 3.3: Linear SVM classifier.

CHAPTER 4

Models

4.1 Introduction

In this project, a machine learning model is proposed through the combination of different methods that have their origins in distinct study lines. These methods are three subunits: morality-based, emotion-based and embedding word similarity feature extraction. An illustration of the proposed model is shown in Figure 4.1 where it can be seen that the three submodules extract their own features from the natural language text input according to their internal algorithms. These features are presented as vectors in the model output and are concatenated, so a machine learning classifier generates a prediction using this concatenation. Regarding the classifier algorithms, Logistic Regression and a Linear Support Vector Machine model have been used.

4.1.1 Moral foundation features

Propaganda often appeals to people's moral values in order to persuade them to adopt a certain belief or point of view [13]. By analyzing the moral foundations present in text, we hypothesize that tactics used to manipulate people can be identified. For example, propaganda may use language that evokes feelings of loyalty and patriotism in potential

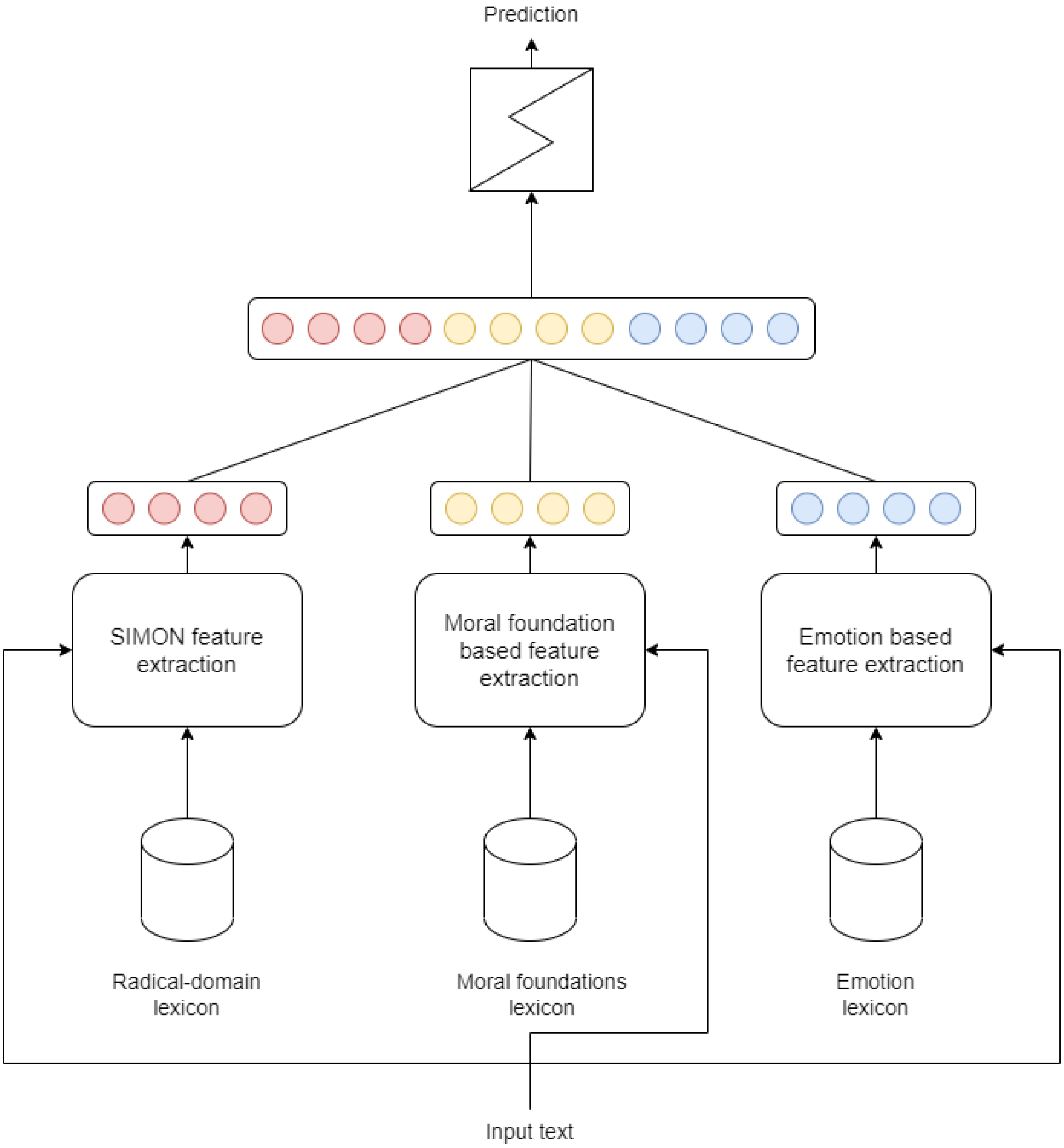

Prediction
SIMON feature extraction
Moral foundation based feature extraction
Emotion based feature extraction
Radical-domain lexicon
Moral foundations lexicon
Emotion lexicon
Input text

Figure 4.1: Architecture of the proposed model.

recruits to motivate them to support a specific political program. On this basis, we propose the use of two resources to extract different features regarding morality: *MoralStrength* and *eMFDscore*.

In *MoralStrength*, the procedure has several steps. First, the association strength between each word and a certain moral foundation is obtained. Then, for each text, the moral values acquired are summed and then divided by the number of words that contributed to get a value between 1 and 9. This process is repeated for each of the five moral foundations,

Algorithm 1 MoralStrength

```
Require: Moral lexicon composed by a vocabulary $T^{(s)}$ and annotations $\boldsymbol{A}$
Ensure: $\boldsymbol{v} \in \mathbb{R}^{n \cdot m}$, the final feature vector
for $i \leftarrow 1, n$ do
    for $j \leftarrow 1, m$ do
        for all $t_k \in T^{(s)} \cap T^{(i)}$ do
            $\boldsymbol{M}_{k,:} \leftarrow$ moralAnnotation($t_k, \boldsymbol{A}$)
            $\boldsymbol{S} \leftarrow$ sum($\boldsymbol{M}_{k,:}$)
            $\boldsymbol{D} \leftarrow \boldsymbol{S}/size(T^{(s)} \cap T^{(i)})$
            $\boldsymbol{v} \leftarrow$ append(D)
        end for
    end for
end for
```

so at the end a matrix of n rows and five columns is obtained, being n the number of texts in the dataset. In 1 it is presented the pseudocode behind the logic of the model. $T^{(i)}$ is the input text of a single instance of the dataset.

In *eMFDscore*, the foundation probabilities are obtained by counting the frequency of a word's annotation with a particular foundation and then dividing it by the total number of times the word was seen by annotators with this foundation assigned. Regarding the sentiment scores, a valence score that combines multiple factors is calculated for each annotation, which ranges from -1 (representing the most negative valence) to +1 (representing the most positive valence), indicating the general sentiment of the annotation. The average sentiment score based on the annotations where the word was used in a manner specific to a particular foundation is calculated for each word. This computes a vector of five sentiment values per word. Regarding the 'mapping' of the scores, the model has been configured so that each word is used as an indicator for all the foundations with the probabilities as weights.

4.1.2 Emotion based features

Radical speech and extremist discourses can entail the use of emotionally charged language and rhethoric with the aim of manipulating and radicalizing potential members of the organization [14][15]. This is why the study of the emotions that take place in a text is a very important step in the path of detecting radicalization. Also, sentiment analysis plays an important role in this task as it can identify the overall sentiment of a text, which can also offer an understanding of its likelihood to become radicalized.

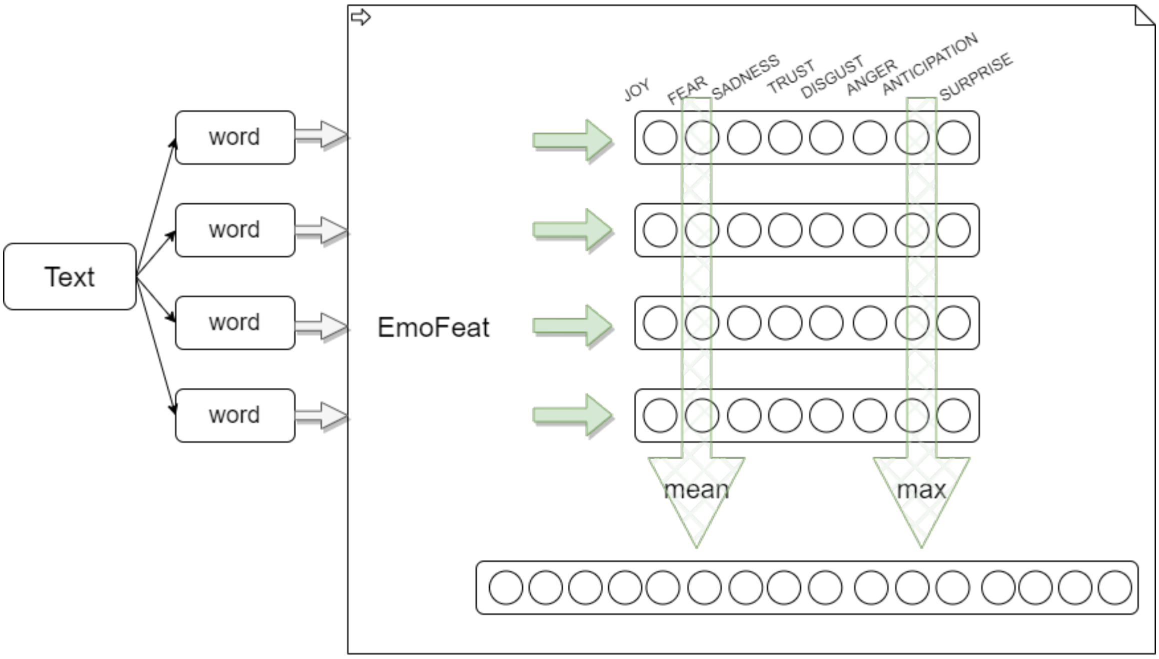


Figure 4.2: Diagram of the emotion feature extraction approach.

In this work, the utilization of the National Research Council emotion lexicon (NRC lexicon) is proposed to help with the extraction of features related to emotions and sentiments. As seen in previous research [12], the emotion-based approach is very useful when detecting extremism. What it is suggested is the use of a lexicon-based approach that utilizes statistical measures for encoding emotional attributes within textual content.

The algorithm that makes this possible, which we will refer to as *EmoFeat (Emotion Features)*, considers an emotion lexicon formed by a group of words $W^{(l)} = w_1^{(l)}, ..., w_i^{(l)}, ..., w_P^{(l)}$ and a vector of numeric annotations $L = [l_1, ...l_i, ...l_P]$. This lexicon has an annotation l_i for each term w_i, so there are P pairs of (w_i, l_i) values. Additionally, the vector l_i indicates the strength of each emotion for word $w_i^{(l)}$ in the lexicon. Besides that, as an emotion vector has dimensionality $l_i \in IR^m$, the computed emotion annotation matrix is $L \in IR^{Pm}$, where the m columns indicate the number of emotions taken into account in the lexicon. Finally, $W^{(i)} = w_1^{(i)}, ..., w_j^{(i)}, ..., w_I^{(i)}$ is defined as the set of I elements formed by the input words.

Taking into consideration the intersection $W^{(l)} \cap W^{(i)}$, the related emotion vector is extracted from L for each word w_k. The result of this process is a matrix with the emotion annotation for all the input words that exist in the lexicon. Then, different statistical metrics are employed to depict the matrix as a feature vector. The suggested metrics include the mean and the maximum. Consequently, a feature vector is derived with a dimension of

$n \cdot m$, where n represents the number of the chosen statistical measures (these measures can be used independently).

4.1.3 Embedding based semantic similarity

Distributed representations have become increasingly popular in natural language processing because they have various benefits over more traditional approaches [16]. One of the main advantages of these models is that they can capture the richness and complexity of the meanings of words in a way that cannot be done with simpler and more symbolic depictions. Word embeddings are the most outstanding technique for computing distributed representations. They usually involve training a neural network to anticipate specific aspects of the context in which a word appears using the vector representation of this word. The problem that arises here is the fact that pre-trained word embedding models' content does not entail any task-specific information as these models are trained from extensive datasets using unsupervised techniques.

Besides, there is currently an issue of data scarcity in the extremist language detection domain [12]. As a consequence, training specific word vectors in this domain does not represent an interesting direction. To avoid this limitation, we use the SIMilarity-based sentiment projectiON (SIMON) model [17].

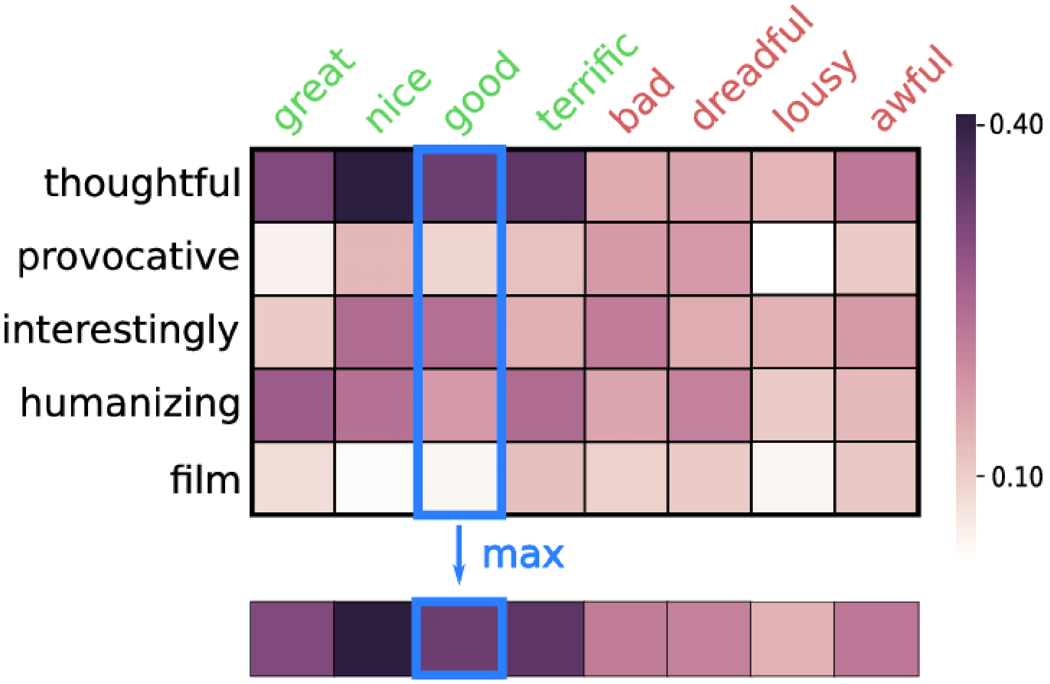


Figure 4.3: SIMON similarity computation.

SIMON is a feature extractor which uses a lexicon centered in a specific domain (in this work, in the radical-detection domain), which is extracted from the training dataset by calculating the frequency of appearance of the vocabulary within the dataset. In previous work [17], this method is called *FreqSelect* and its aim is to serve as a basic reference to capture the lexicon of a particular domain in a straightforward manner. The main proposal is the representation of a certain term that could be absent from the lexicon database. This approach involves projecting this word onto a group of sentiment words that have been extracted from a sentiment lexicon. As seen in Figure 4.3, this projection is done making use of the semantic similarity between words by means of a word embedding model as it contains semantic and syntactic information by converting the text into vectors of a predetermined length by applying a max function on each column.

The data is measured against the specific vocabulary and a vector is calculated with the representation of the similarity between them. With this method, material coming from both a word embedding model and a domain lexicon can be exploited. Furthermore, this model can be used when an extensive corpus is not available. The study of SIMON has already been developed in the detection of radicalization [12], the estimation of moral value [18], and the analysis of hate speech [19].

CHAPTER 5

Evaluation

The assessment of radicalization has been modeled as a binary classification task, being non-radical and radical the negative and positive classes, respectively. Thus, the techniques used for this task have been developed by leveraging the datasets, embeddings and lexicons described in Section 5.1, and adhering to the methodology outlined in Section 5.2. The results of this evaluation are in Section 5.3.

5.1 Materials

The datasets used in this study and their main features are shown in Table 5.1. These datasets are the following.

Pro-neu. This dataset is the amalgamation of two distinct English datasets gathered by [20]. On the one hand, one of the datasets, accesible on the net [21], is composed of 17,350 tweets from 112 different Twitter accounts that sympathize with ISIS. This collection process was carried out through the filtration of keywords (e.g., *Amaq, Dawla*), images, and their followers networks. On the other hand, these radical instances are balanced with a different set, composed of more than 122k tweets from more than 95k different Twitter accounts. Unlike the first dataset, this one has neutral or anti-ISIS messages. A filtration

Table 5.1: Table 1. Statistics of the datasets: number of instances, category balance (percentage), average number of words per instance and source.

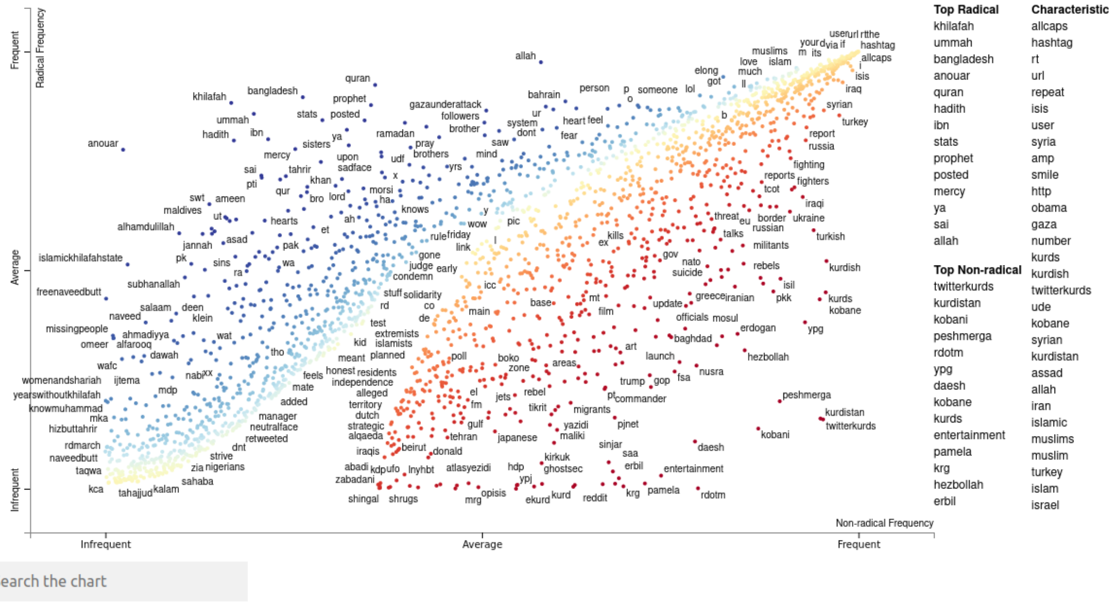
Dataset	**Instances**	**Balance (%)**	**Avgerage no. of words**	**Source**
Pro-Neu	224	50/50	18,646	Twitter
Pro-anti	1,132	50/50	36,352	Twitter
Magazines	468	68/32	950	Magazines
Jacobs	5,000	50/50	519	Dark Web

of keywords related with ISIS (e.g., *ISIS, IslamicState*) helped in the collection process. Please note not all of the 95k accounts were utilised: 112 of them were selected by filtering the ones that are not currently active.

Jacobs. This resorce was created by Scanlon and Gerber [22]. It is composed of a group of jihadist posts from private forums with origin in the dark web, which were already assembed in the Dark Web Portal Project [23]. These forums belong to the Ansar AlJihad Network, an extremist organization with ties to Al-Qaeda and well-liked among Western jihadists [22]. The classification carried out in this dataset is binary: propaganda and non-propaganda.

Pro-anti. This English-written dataset is the collection of tweets from 1,132 Twitter accounts done by [24] and can be divided into two groups; the first one contains 566 instances which were categorized as pro-ISIS, as they are users that share propaganda material from established pro-ISIS accounts that aims to incite or provoke. In an initial version of the dataset, there were 727 identified accounts but 161 were found either closed-down or unavailable for public access, so they were removed. The second group contains another 566 different instances extracted from anti-ISIS accounts. That is, a categorization done through the analysis of language use in accounts that oppose ISIS.

Figure 5.1 shows a visualization of the *Pro-anti* dataset. Half of the instances have been randomly selected because there were too much of them. The visualization depicted in this graph is a scatter plot presenting the frequency distribution of words in the dataset categorized as radical and non-radical. Each point on the plot represents a word, and its color indicates its frequency in relation to each category: blue for radical and red for non-



Radical document count: 281; word count: 5,882,649
Non-radical document count: 285; word count: 13,100,215

Figure 5.1: Word frequency for both radical and non-radical categories regarding half of the *Pro-anti* dataset. On the right, most frequent words for radical (Top Radical), non-radical (Top Non-radical) and both (Characteristic).

radical. To analyze the frequency within each category, the occurrence of different words is computed. Due to limited space, only a subset of word labels is displayed along the figure. The y-axis represents the frequency within the radical category, so words that frequently appear in radical texts are positioned towards the upper region of the plot.

Likewise, the x-axis represents the frequency within the non-radical category. Words that commonly appear in non-extremist texts are positioned towards the right side of the graph. Notably, the areas that help the most in this study are the top left (words frequent in radical texts), bottom right (words frequent in non-radical texts), and top right (words frequent in both neutral and radical texts) sections of the visualization as they reveal the most distinctive words associated with the neutral, radical, and overlapping categories. Examining these regions provides insights into how words are employed in these two categories. For instance, notable radical words found in the dataset include 'khilafah' and 'ummah'. The first one means 'caliphate', which denotes the position held by the leader responsible for the political affairs of the Muslim community or state, in particular during the period from 632 to 1258. 'Ummah' is the Muslim community itself. In contrast, non-radical texts frequently feature words such as 'twitterkurds' and 'kurdistan'. Kurdish people have suffered violence and injustice from the Islamic State; in fact, there exist militia groups against ISIS in this country [25], so it is logical that these terms are found in the non-radical category.

Magazines. This English-written dataset is presented in [12] and two parts can be differentiated. For the first one, the data came from two online magazines shared by the Islamic State of Iraq and the Levant radical organization [26]: Dabiq [27] and Rumiyah [28] magazines. 166 articles from 13 editions have been extracted from Dabiq (released in July 2014 and lasted two years) and 155 articles corresponding to 15 editions from Rumiyah (released in September 2016 and lasted one year). All content has been originally extracted from `jihadology.net`, a digital platform that addresses terrorism[1]. As balance of Dabiq and Rumiyah's texts, additional data is considered. Concretely, texts from two digital newspapers that deal with matters related to ISIS from a non-radical perspective: Cable News Network (CNN)[2] and The New York Times[3], which are both sources that give away their content at no cost through their APIs. To gather the data, a keyword-based filtration (e.g., *ISIS, Daesh*) was done for a 10-month period. For an increase in the value of the categorization, texts that did not address the ISIS issue where removed, as well as links, images and other media. As a result, 129 instances where added to the *Magazines* dataset form the CNN, and 23 from The New York Times.

As mentioned, word embeddings have been used in order to make it possible for machines to understand the data contained in the corpus when SIMON [29, 17] method was applied. Following previous research [12], we use the FastText embedding model[30]. This is a 300-dimension vector with a vocabulary size of 1,999,995 words in which the training domain is Wikipedia.

In addition, as described above, an objective of this work is to explore the impact of contextualization through domain-relevant lexicons. We study the effect of the following lexicons.

The ***MoralStrength*** lexicon is a resource containing nearly 3,000 manually annotated words and their association strengths with each moral trait (care, fairness, loyalty, authority, purity) where, for each word, it is provided a community-sourced numerical evaluation of Moral Valence. This lexicon is an expansion of the lemmas from the original Moral Foundations Dictionary (MFD), as presented in [18].

NRC Hashtag Emotion Lexicon [31] is an inventory of 16,862 English words and their relationship with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) selected concretely for

[1] http://www.jihadology.net/

[2] https://cnn.com/

[3] https://nytimes.com/

the Twitter platform. As observed in related works [12], emotion can have an impact of radicalization detection.

eMFD Lexicon [32] is made up of 3,270 English words. Every word is allocated 5 probability scores that indicate their association with each moral foundation, as well as 5 sentiment ratings that encode average sentiment of the foundation context in which each word appears.

Finally, it is worth mentioning that when using the SIMON feature extractor, a specific lexicon was generated and utilized for each dataset to extract dataset-oriented features. We use the *FreqSelect* method, presented in [12].

5.2 Methodology

We design a thorough evaluation to assess the different proposed models and their effectiveness in the task of radicalization detection. In every experiment, we use the macro-averaged F1 score for performance assessment. For each dataset, a k-fold cross-validation is performed, being $k = 10$. To study the effect of contextualization through diverse representations, we use different feature extractors, concatenating their resulting vectors. The ones used as the baseline are unigrams, bigrams, the TF-IDF method and SIMON. Unigrams provide speed, transparency, flexibility, and accuracy as a radical propaganda detection technique. Analyzing bigrams offered a more nuanced comprehension of propaganda techniques, a higher precision, and aid in the identification of complex propaganda methods that might evade detection by solely analyzing unigrams. The TF-IDF method is also used as it can provide a more targeted analysis and the detection of important themes.

At first, the only results taken into account were the ones obtained from extracting features with unigrams, bigrams and the TF-IDF method. These outputs were compared with those generated using the SIMON method. Following, we merged these two approaches to see if an improvement had taken place compared to the baseline. The process that followed this stage of evaluation was the combination of the feature extractors with the used lexicons. Additionally, we include an evaluation on the effectiveness of an unified representation that combines SIMON, emotion and morals.

Regarding the classifier algorithms, two have been used: Logistic Regression and Stochastic Gradient Descent Learning with a LinearSVM configuration.

5.3 Results

Tables 5.2 and 5.3 present the results of the evaluation considering the described models and their combinations with the contextual lexicons. Something very important that should be commented on is some of the results obtained in the *Pro-neu* dataset. This dataset has a lack of quality that makes it difficult to obtain a proper F1 score (100% is not a good result as a machine cannot be 100% right in its performance by nature). Taking this issue into account, we will focus on results that do not reach the total percentage.

It is noticeable that the F1 scores do not defer much from one classifier to the other and, in general, they are better when using Logistic Regression classifier. Regarding Table 5.2, it can be seen that the best scores are consistently obtained by combinations of representations. In the *Pro-neu* dataset, the highest score is obtained by merging the SIMON method with morals from MoralStrength and emotions from the NRC emotion lexicon. This shows the added value that this type of combination has on the classification of radical text. Table 5.3 demonstrates that the SIMON method is enough for obtaining a very high F1 score without any type of aggregate.

In *Pro-anti*, table 5.2 shows that the strongest result is obtained by joining SIMON and features obtained through the TF-IDF method. There is a very noticeable increase with respect to just TF-IDF, while a less noticeable increase with respect to just SIMON. Such a result indicates that the sparsity of the TF-IDF representations complements the SIMON model. Although the combinations with MoralStrength and eMFD decrease this baseline, this does not happen when using SIMON as the baseline, where both moral foundations and emotions approaches help to get better scores even when using them at the same time (SIMON + MoralStrength + NRC). In the case of unigrams, only SIMON has an added value and in the case of bigrams only MoralStrength decreases the baseline. On the other hand, 5.3 has the best score when combining the TF-IDF method and emotions. When using SIMilarity-based sentiment projectiON (SIMON) as the baseline, only morality helps to increase it. This does not entirely happen when the baselines are unigrams or bigrams, only SIMON makes it happen.

In *Magazines*, tables 5.2 and 5.3 demonstrate that the most outstanding result is the one resulting from the combination of bigrams and the SIMON method. Nevertheless, MoralStrength contributes to get a nearly as high mark when using the Logistic Regression classifier. Continuing with this classifier, an interesting point is that the first method mentioned is the only one that, in combination with the baseline, increases the score when unigram or TF-IDF approaches are used. With SIMON as the baseline, moral features seem

Method	Pro-neu	Pro-anti	Magazines	Jacobs
Unigrams	95.08	88.78	95.38	91.00
Unigrams + MoralStrength	95.08	88.60	94.64	90.84
Unigrams + NRC	96.42	88.60	95.11	91.04
Unigrams + SIMON	95.98	88.87	95.43	91.02
Unigrams + eMFD	96.87	88.60	95.14	91.16
Bigrams	91.50	84.97	96.86	86.93
Bigrams + MoralStrength	89.73	84.88	97.10	87.42
Bigrams + NRC	95.08	86.13	95.16	88.07
Bigrams + SIMON	94.19	86.13	**97.58**	90.50
Bigrams + eMFD	92.84	85.59	96.86	87.26
TF-IDF	87.50	85.76	86.93	91.73
TF-IDF + MoralStrength	86.15	84.61	85.67	91.32
TF-IDF + NRC	95.09	87.24	80.80	91.58
TF-IDF + SIMON	98.21	**89.09**	95.64	91.48
TF-IDF + eMFD	90.62	85.48	86.81	**91.84**
SIMON	98.21	88.03	94.35	90.04
SIMON + MoralStrength	98.21	88.48	93.11	90.14
SIMON + NRC	98.21	88.85	94.93	90.32
SIMON + eMFD	98.21	88.03	95.41	90.40
SIMON + MoralStrength + NRC	**99.11**	88.93	93.86	90.26

Table 5.2: Results with a Logistic Regression classifier.

Method	Pro-neu	Pro-anti	Magazines	Jacobs
Unigrams	94.63	84.16	92.99	90.88
Unigrams + MoralStrength	93.29	83.85	93.94	90.82
Unigrams + NRC	95.53	85.05	93.88	90.90
Unigrams + SIMON	95.08	87.42	94.86	90.26
Unigrams + eMFD	92.85	85.04	92.26	90.58
Bigrams	92.84	87.99	95.88	86.53
Bigrams + MoralStrength	91.50	85.32	93.20	86.87
Bigrams + NRC	93.29	85.77	94.37	87.27
Bigrams + SIMON	93.29	88.15	**97.33**	90.06
Bigrams + eMFD	94.64	85.31	96.13	86.86
TF-IDF	95.09	88.32	95.19	93.20
TF-IDF + MoralStrength	95.54	87.70	94.96	93.30
TF-IDF + NRC	96.43	**91.42**	91.29	**93.52**
TF-IDF + SIMON	**98.66**	89.20	93.39	92.82
TF-IDF + eMFD	94.20	88.49	94.48	93.50
SIMON	**98.66**	86.43	93.39	89.86
SIMON + MoralStrength	98.21	86.62	91.69	89.96
SIMON + NRC	100.00	85.92	94.39	90.28
SIMON + eMFD	**98.66**	86.35	93.39	90.22
SIMON + MoralStrength + NRC	100.00	85.74	93.18	90.14

Table 5.3: Results with a Linear SVM classifier.

to elevate the result obtained when eMFD. Emotions have an important role too in this results as they also enhance the score. Moving on to the other classifier, it is worth mentioning that emotions and morality when using MoralStrength improve the baseline results, which is positive. This does not happen when using TF-IDF. Although the combination of semantic similarity with MoralStrength and emotions does not provide the highest result, it is a very high one.

In *Jacobs* dataset the better approach that can be used in combination with others is the TF-IDF method. This can be seen in tables 5.2 and 5.3, where its merging with eMFD and NRC gives the best scores in each of them, respectively. When using unigrams as the baseline, only emotions can improve it when using Linear **svc!** (**svc!**) classifier but, if bigrams or SIMON are utilized, all of the rest of the approaches have an additional worth for both classifiers.

For a more visual understanding of the impact each combination of methods has on text classification, SHapley Additive exPlanations (SHAP) has been applied to the models with the best scores [33]. The SHAP method offers a thorough framework for interpreting the predictions generated by any machine learning algorithm. It entails explaining the output of machine learning models by assigning an importance score to each input feature that indicates how much each feature contributes to the prediction. SHAP is based on the Shapley value concept from cooperative game theory, where a fair assignation of the contribution of each player takes place in a cooperative game. In this context, the "players" are the input features, and the "game" is the prediction task.

In this work, we exploit the information contained in 'Beeswarm plots', which are a SHAP visualization tool for explaining the results obtained. Each point represents an instance of the dataset and its position on the x-axis indicates its SHAP value for a specific feature, which is an indicator of the contribution of that feature to the final model prediction for that instance. The density of the dots is shown through the y-axis, which gives information about the distribution of the feature values in the dataset. Also, the color of the dot indicates, in this case, the level of radicality of the instance, red for high radical value and blue for low radical value.

Figures 5.2, 5.3 and 5.4 show that moral values, emotions and features obtained through the SIMON model have an important role in the classification process (the words that appear are the most important ones).

As expected, as shown in figures 5.3 and 5.4, the moral value 'loyalty' is a feature with red positive SHAP values, which means that it is a feature that contributes to the final prediction of the model with an inclination towards extremism. This foundation is understood in

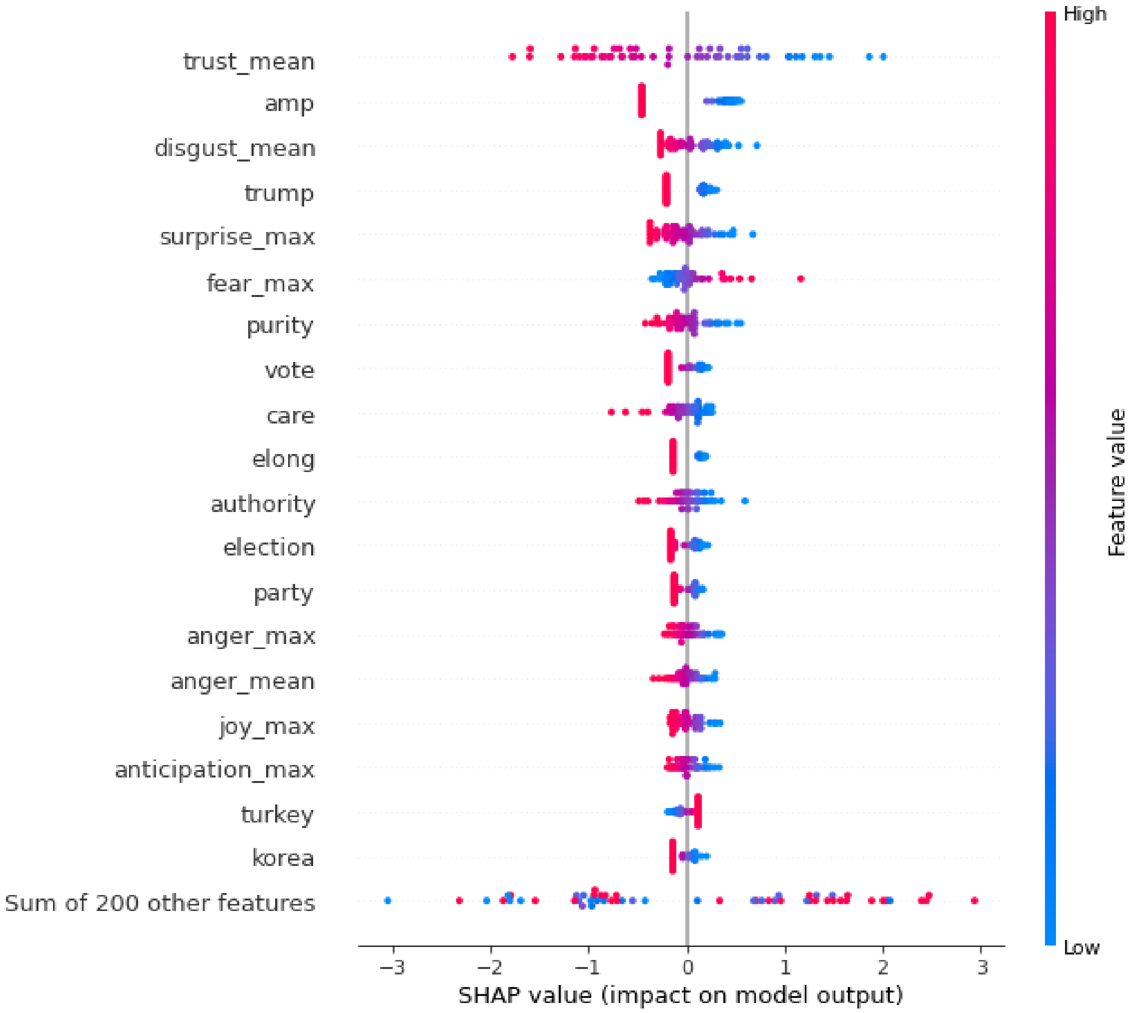


Figure 5.2: SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Pro-neu* dataset.

radical speeches as a virtue that requires continuous commitment to a particular ideology, cause or leader. The final aim of these discourses is to enforce conformity and suppress dissent within the group or movement. In contrast, the moral foundation 'authority' has blue positive SHAP values, as shown in figure 5.2. As mentioned, the *Pro-neu* dataset has anti-ISIS instances, so it is logical to think that these messages spread ideas where authoritarian forms of organization have negative connotations. It is worth mentioning the appearance of the word 'purity' in figures 5.2 and 5.3 as a feature with positive blue shap values. A possible explanation for this would be that purity-oriented individuals may associate extremist ideologies with immoral or impure actions that conflict with their own moral values.

Regarding emotions, figure 5.3 shows that 'anger' is an affect that helps to classify

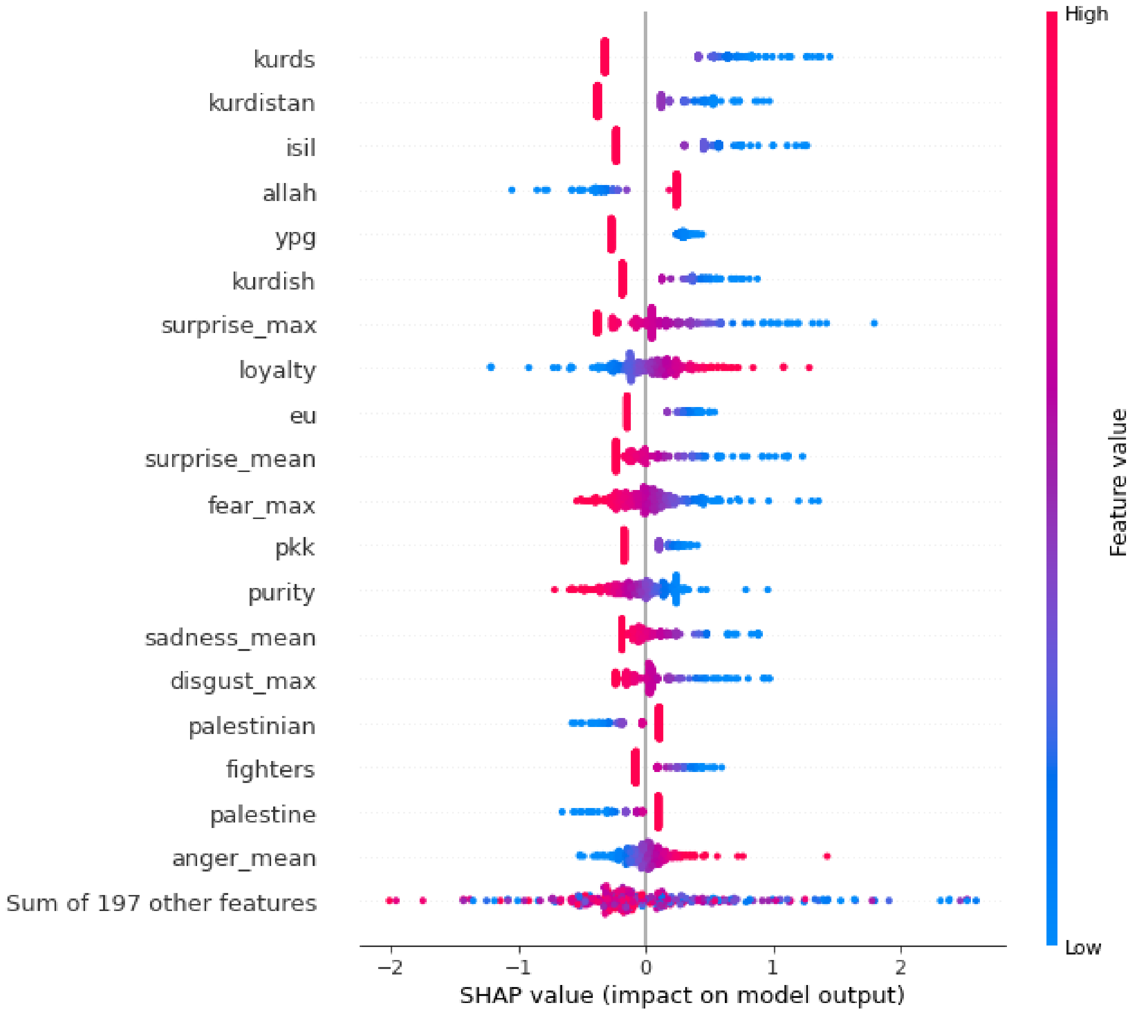


Figure 5.3: SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Pro-anti* dataset.

instances with a radical perspective as the extremist speech resides in hatred and fury, while 'sadness', 'fear' and 'disgust' contribute to the classification of texts with a non-radical discourse, which is a predictable result. Nevertheless, in figure 5.2, 'anger' is seen as an affect that aids in the categorization of non-radical texts, which can also be logical because, as it has been mentioned, *Pro-neu* dataset has also instances with texts that transmit ideas against radical beliefs, not just informative texts that could appear in the CNN. In 5.5 the most outstanding feature is the 'anticipation' emotion. This can be because radical discourses sometimes can show a vision or promise of a future outcome or transformation. They may stress the need for immediate action for apparent injustices that require extremist solutions.

Features obtained through SIMON method such as 'ypg' in figure 5.3 have expected

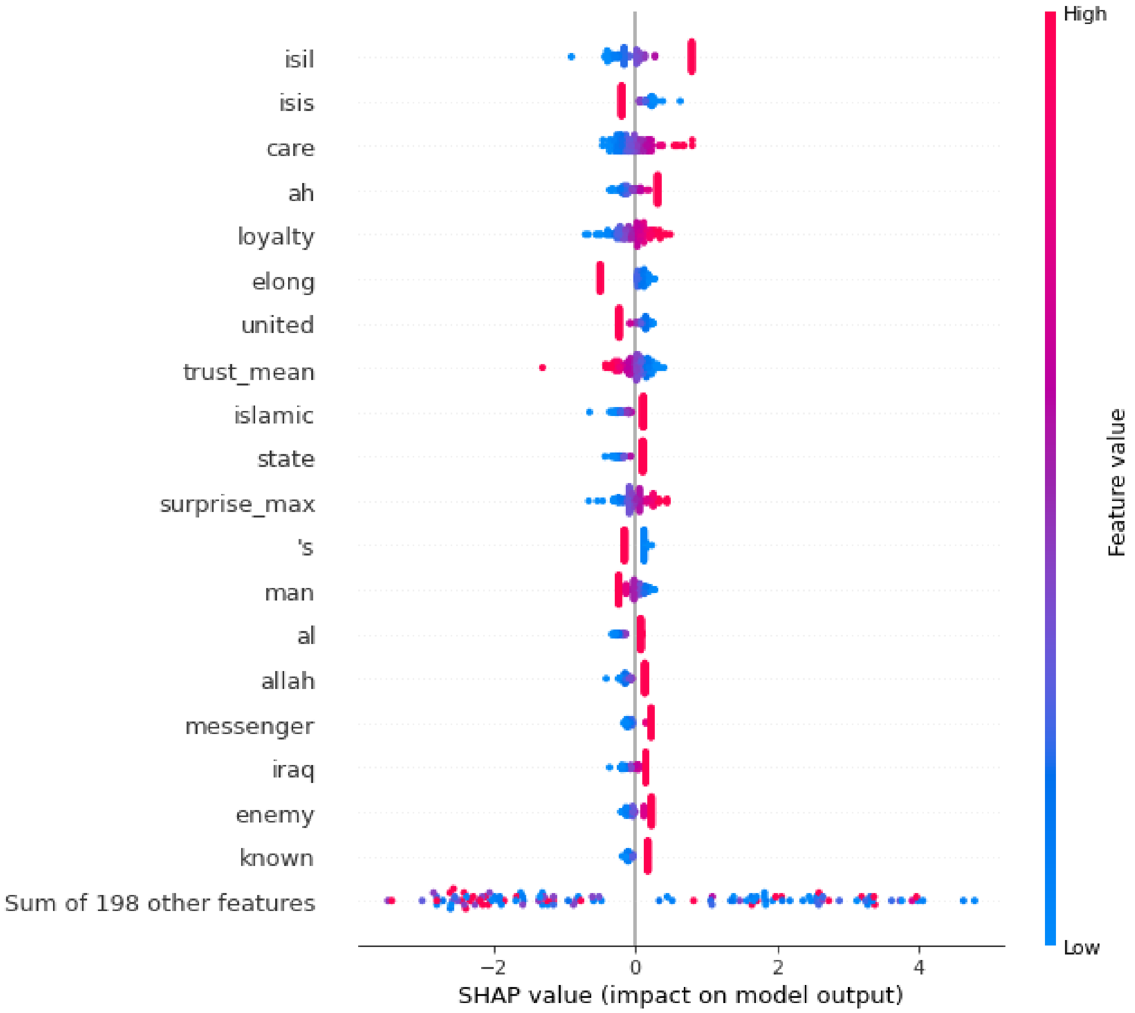


Figure 5.4: SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Magazines* dataset.

explanations with contributions to non-extremist discourses. This feature refers to "People's Protection Units", a militia group from Kurdistan active from 2011 that has been the partner of the United States coalition in Syria against the ISIS. This also explains that 'kurds', 'kurdistan' and 'kurdish' have high shap values regarding non-radicality. As a counterpoise, 'allah', a feature shown in figure 5.4, has a not too high positive shap value concerning radical texts. It is an Arabic word that means 'God' and holds important religious and cultural relevance for Muslims; in some cases, individuals or groups with radical or extremist ideologies may use religious language, including references to Allah, to legitimize their actions. 'Praise' and 'mercy' are similar examples shown in figure 5.5.

Figure 5.6 shows the beeswarm plot for the model in which bigrams and features from the SIMilarity-based sentiment projectiON (SIMON) model are combined, as it is a fusion

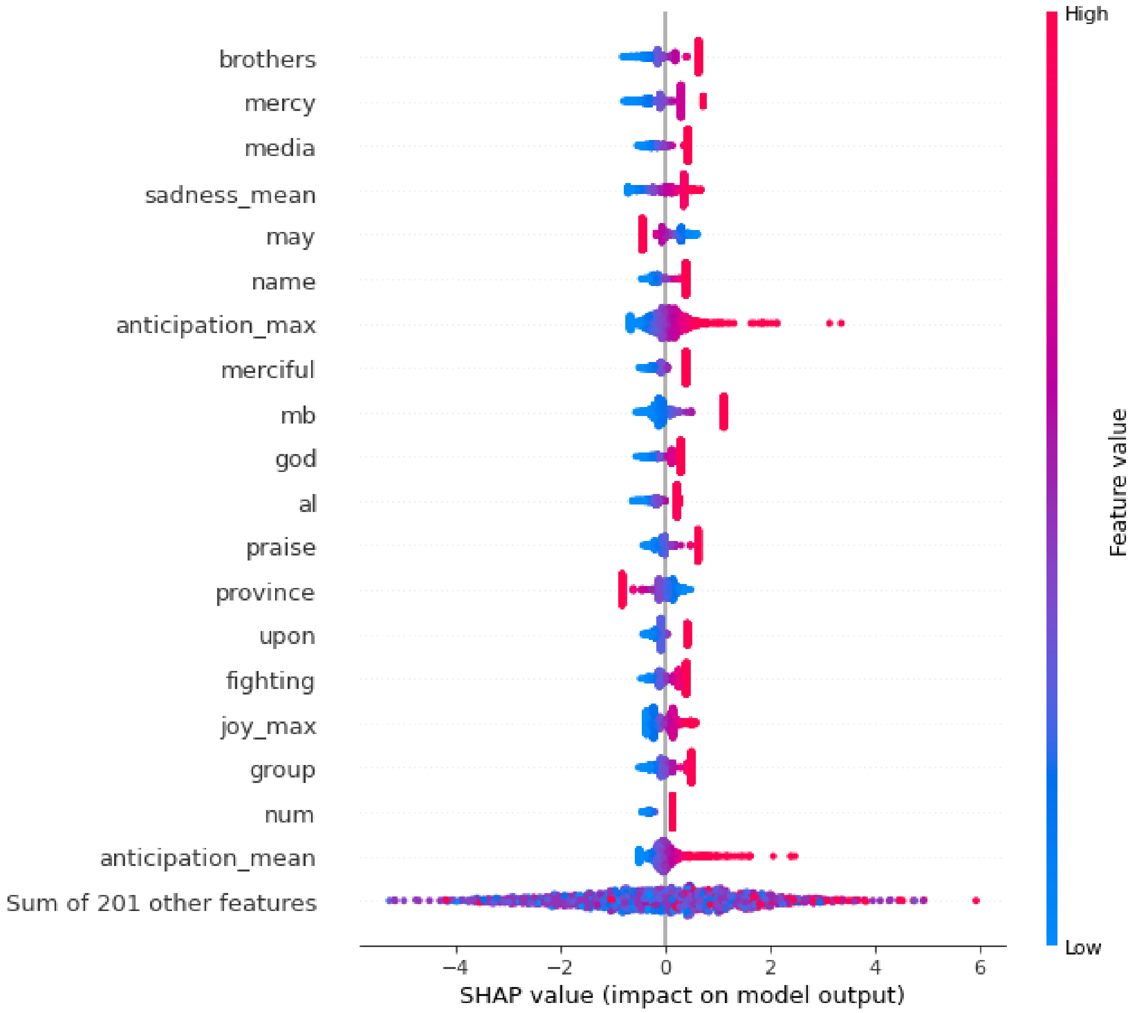


Figure 5.5: SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Jacobs* dataset.

that has provided a very good result in the experiments. For example, bigrams like ’al baghdadi’ (the former leader of the Islamic State of Iraq and Syria) and ’bashar al’ (the president of Syria) seem to have helped (with a very low shap value) to the categorization of radical text. SIMON features like ’war’, ’isis’ or ’fear’ have also carried out this function, which is something that was expected. The feature ’murtadd’ (an Arabic word that refers to an individual who rejects or abandons their previously embraced religious beliefs or faith) would be anticipated to operate as the previous ones described but, surprisingly, it has very low shap values for text classification towards the radical side.

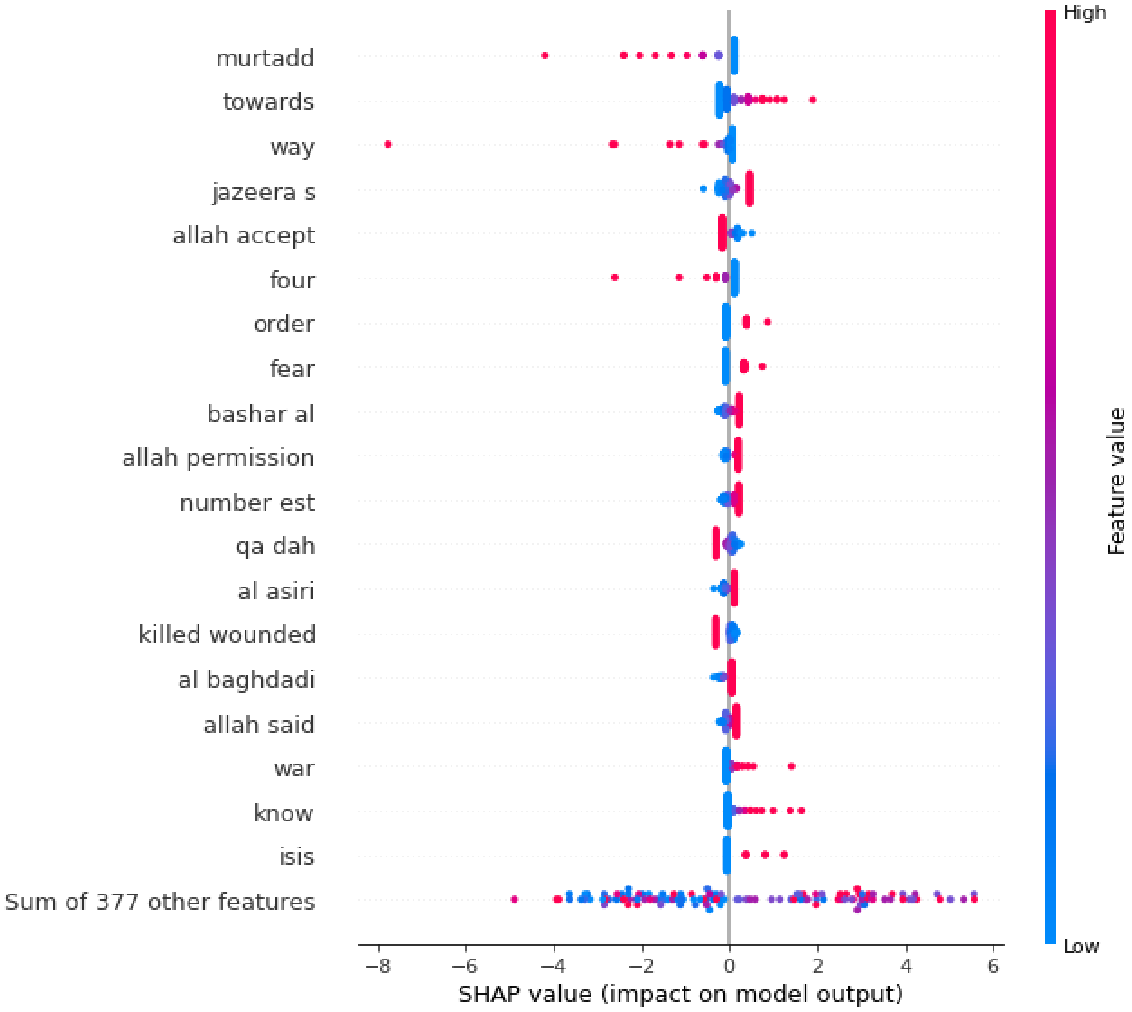


Figure 5.6: SHAP beeswarm plot for bigrams + SIMON with *Magazines* dataset.

CHAPTER 6

Conclusions and future work

In this chapter we will describe the conclusions extracted from this project, and the thoughts about future work.

6.1 Conclusions

This work follows the research done in [12], where detection of radicalization is carried out through the exploratory combination of emotions and similarity-based features with very good results. Here, we add moral values based on the Moral Foundations Theory and see how the affect the results as they have an important role in the task of identifying political inclinations [34]. It is worth mentioning that the scores obtained in this work are macro-averaged, which involves taking the arithmetic mean (unweighted mean) of the F1 scores for each class and treating all classes equally regardless of their support values.

Regarding the method used in this investigation, other studies do not use machine learning techniques to evaluate the radicalization present in texts [24, 35, 36, 37, 38, 39]. Our approach employs two commonly used machine learning algorithms: Logistic Regression and Linear Support Vector Machine (SVM). They were selected for this work because the objective is the classification of text according to their level of extremism employing a richer variety of features but avoiding the complexity of more complicated algorithms. These or comparable learning algorithms are also utilized in certain studies, including: [40, 41, 42, 43, 44, 45, 46]. Our approach differs mainly in the feature extraction stage. The proposed method utilizes moral values, emotions and similarity-based features that leverage the extensive lexicon contained in word embeddings (apart from unigrams, bigrams and features generated by the TF-IDF method, which also have important results in Tables 5.2 and 5.3. As far as we are aware, this type of approach for evaluating radicalization has not been previously suggested. What previous works have done is utilize lexicons by directly comparing words in the analyzed text with those included in the lexicon. Thus, such approaches fail to adequately model out-of-vocabulary terms.

The limitations that have been perceived in this work are the lack of resources available in the radicalization domain. On the one hand, there are not many radical datasets available, as some texts come from magazines intended for a very specific public. In addition, not all of them have good quality. Other datasets come from Twitter accounts (some of them have been banned or deleted). Also, it is not available to work with a corpus of significant size containing explicit annotations for detecting radicalization, so training a word embedding model with it is not accessible at the moment [47].

This work describes a machine learning system that utilizes various categories of characteristics to accomplish the objective of identifying radical propaganda in texts that come from both social media and magazines: moral features, emotions, similarity-based features obtained through word-embeddings, unigrams, bigrams and features acquired from the TF-IDF method. Besides, these approaches are combined with each other with the aim of improving the overall task performance. Regarding the first two, what they provide is supplementary information that gathers moral and emotional knowledge and adds context to the analyzed text creating features that can be very useful in the classification process. On the other hand, SIMON approach is used with a lexicon obtained by sorting words based on how frequently they appear in the training data. This is the mentioned method *FreqSelect*. Unigrams, bigrams and features from the TF-IDF method have also an added value as they help to identify noteworthy word associations. We assess all techniques using three datasets that can be found online.

In Section 4.1, three essential investigation inquiries are proposed. The first one (RQ1) asks if moral values are useful for identifying radical propaganda. In general terms, the F1 score improves with respect to the baseline when using bigrams or the SIMON model.

Regarding RQ2, as seen in 5.2, with a Logistic Regression classifier the combination of the three main approaches provides the best F1 score in *Pro-neu* dataset. For the rest of the datasets, except for *Magazines*, the outcome of this model is very close to the best result obtained in each of them. If the classifier used is a linear SVM, an enhancement can only be seen in *Jacobs* dataset. The values obtained through SIMON model are already very high, so it is a difficult task to elevate them.

6.2 Achieved goals

1. **Integration of feature extraction techniques that come from distinct approaches:** Several methods for extracting features from textual data have been used and combined in Jupyter Notebooks to see how they function in distinct scenarios. These modules had their origin in different study lines, such as affect signs, moral values, bigrams and semantic similarity computation, among others.

2. **Obtainment of the different F1 scores in order to see which models perform a more efficient classification:** The results regarding the effectiveness of each model obtained in the experiments using the proposed classifiers have been analyzed and gathered in the tables shown in this document.

3. **Visual explanations of how features have an added value in diverse combinations of models:** For a graphic representation of the razionalization of the different predictions, SHapley Additive exPlanations (SHAP) has been utilized since it provides a structure for analyzing and breaking down the individual contribution of each feature in the model to the final prediction. The plots obtained have been also commented and explained.

6.3 Future work

Regarding the technology that has been used in general terms and how it can be further improved in future work, this research has selected the automated detection path. This path has two ways: Graph and Machine Learning (ML). According to [7], the Graph approach is not as popular as the ML approach among the studies that have been carried out as it is employed to uncover supporters of known radicals and to identify the level of extremist influence, but fails to detect the novel radical users. On the contrary, the most popular technique, ML, uses algorithms that efficiently discover newly radicalized users and continuously learn from the transforming extremist material.

Despite the fact that the results achieved are quite promising, we consider that it is very important to keep a thorough study on the use of language as it progressively varies with the pass of time. An essential task that should be improved in future work so that more reliable scores can be obtained is the gathering of information that makes up the dataset. In addition, how this information is classified can also be enhanced: the binary categorization of texts (radical and non-radical) can be substituted by a classification based on a wider spectrum. This would provide a more detailed performance and a greater understanding of the nature and impact of the propaganda.

Technologies that the research community should keep investigating are advanced deep learning architectures such as neural networks and transformer models in order to effectively capture the connections and meaningful associations within propaganda texts. Also, the implementation of active learning techniques would ease the path of researchers as the models would be updated with new information. Consequently, this would decrease the dependence on manually annotated data.

APPENDIX A

Economic budget

This appendix details an adequate budget to bring about the project, including physical resources, project structure, human resources and taxes.

A.1 Physical resources

A computer is needed to execute the code of this project. The one that has been used has these characteristics:

- **RAM:** 128 GB
- **CPU:** 12 cores Intel(R) Xeon(R) CPU E5-2430 v2 @ 2.50GHz
- **Storage:** 1 TB

A.2 Project structure

The activities that have been carried out for the development of this work are shown alongside the number of days necessary to complete each of them in the table below:

Activity	Days
Learning about Python and NLP with tutorials	14
Trying different combinations of methods for *Magazines* dataset	40
Trying different combinations of methods for all the datasets at the same time	30
Gathering of the results in an Excel form	5
Using SHAP to obtain explanations for the most important results	14
Writing of the article for IEEE Access	35
Writing of the document	21
Total	159

Table A.1: Project structured by tasks

A.3 Human resorces

For this section, both the time needed to develop the project and the salary of an engineer are taken into account. Let's consider that a month has approximately twenty-two working days and four hours per working day (part-time schedule). The time cost of this project is estimated to be around 636 hours (seven months). The expected total monthly compensation for engineers involved in the development of this type of software is expected to reach 450 euros on average. With this data, the total cost dedicated to the development of the software results in around 3,150€ (before taxes).

A.4 Licenses

As described in Chapter 3, the tools that have been utilized are open-source, so there are no costs for licenses.

A.5 Taxes

According to software taxation [48], a tax of 15% of the product value must be considered. The support and regulation for this measure are governed by Spanish law, specifically Statue 4/2008. This situation would only be examined if there was interest from a foreign company in the sale.

APPENDIX B

Impact of this project

This appendix reflects, quantitatively or qualitatively, on the possible impact of this project on different fields.

B.1 Social impact

A better classification of the discourses that are spread on the Internet would lead to a better control of the information that reaches people all over the world. Thus, less people would be exposed to harmful and violent messages, which would reduce the polarization of opinions that takes place nowadays on the net and the societal stability would be enhanced. Additionally, it would be more difficult for terrorist organizations to recruit new members, at least digitally. There would be more people focused on improving the world we live in as distractions related to extremist discourses would be removed.

B.2 Economic impact

The economic consequences that this project would have are the following. First, as said in Section B.1, there would be a more stable society with fewer social tensions, which would lead to the development of an environment more suitable for economic growth. This would also motivate business owners to invest and operate more assuredly. Second, as said in Section B.1, less people would be radicalized and distracted from their work, so productivity would increase and the economic output too.

B.3 Environmental impact

The impact that this work has on the environment resides in two ideas. First, a reduction in terrorist recruitment would lead to a reduction in terrorist activity, which involves conflicts that can have destructive consequences in the environment. The other aspect that must be taken into account is the costly nature of developing and implementing Artificial Intelligence algorithms and models.

B.4 Ethical implications

The most important ethical implication that can be found in this project is what information is understood and classified as radical. There are no political implications in this work, and the processed data is treated in the most objective possible way.

Bibliography

[1] Stephen Cass. Top programming languages 2022, Nov 2022.

[2] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, 2018.

[3] Cori E Dauber and Carol K Winkler. Radical visual propaganda in the online environment: An introduction. *Visual propaganda and extremism in the online environment*, pages 1–30, 2014.

[4] Lorenzo Vidino. *Countering Radicalization in America:*. JSTOR, 2010.

[5] Saja Aldera, Ahmad Emam, Muhammad Al-Qurishi, Majed Alrubaian, and Abdulrahman Alothaim. Online extremism detection in textual content: A systematic literature review. *IEEE Access*, 9:42384–42396, 2021.

[6] Elizabeth Bodine-Baron, Todd C. Helmus, Madeline Magnuson, and Zev Winkelman. *Examining ISIS Support and Opposition Networks on Twitter*. RAND Corporation, Santa Monica, CA, 2016.

[7] Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and Ketan Kotecha. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404, 2021.

[8] Jesse Graham, Jonathan Haidt, and Brian Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96:1029–46, 06 2009.

[9] Jacquelien van Stekelenburg. Radicalization and violent emotions. *PS: Political Science &; Politics*, 50(4):936–939, 2017.

[10] Stephen K. Rice. Emotions and terrorism research: A case for a social-psychological agenda. *Journal of Criminal Justice*, 37(3):248–255, 2009.

[11] Clark McCauley and Sophia Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, 20(3):415–433, 2008.

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

[13] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[14] Clark McCauley and Sophia Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, 20(3):415–433, 2008.

[15] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Survey track.

[16] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[17] Oscar Araque, Ganggao Zhu, and Carlos A. Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346–359, 2019.

[18] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184, 2020.

[19] Diego Benito, Oscar Araque, and Carlos A. Iglesias. GSI-UPM at SemEval-2019 task 5: Semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 396–403, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[20] Miriam Fernandez, Moizzah Asif, and Harith Alani. Understanding the roots of radicalisation on twitter. WebSci '18, page 1–10, New York, NY, USA, 2018. Association for Computing Machinery.

[21] Fifth Tribe. How isis uses twitter, Nov 2019.

[22] Jacob R Scanlon and Matthew S Gerber. Automatic detection of cyber-recruitment by violent extremists. *Security Informatics*, 3(1):1–10, 2014.

[23] Hsinchun Chen, Edna Reid, Joshua Sinai, Andrew Silke, and Boaz Ganor. *Terrorism informatics: Knowledge management and data mining for homeland security*, volume 18. Springer Science & Business Media, 2008.

[24] Matthew Rowe and Hassan Saif. Mining pro-isis radicalisation signals from social media users. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):329–338, Aug. 2021.

[25] Michael Stephens. Facing isis: The kurds of syria and iraq. *IeMed Mediterranean Yearbook. European Institute of the Mediterranean, Barcelona*, 2015.

[26] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[27] Dabiq: The strategic messaging of the islamic state.

[28] Remy Mahzam. Rumiyah – jihadist propaganda & information warfare in cyberspace. *Counter Terrorist Trends and Analyses*, 9(3):8–14, 2017.

[29] Oscar Araque, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. Gsitk: A sentiment analysis framework for agile replication and development. *SoftwareX*, 17:100921, 2022.

[30] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.

[31] S M Mohammad and S Kiritchenko. Using hashtags to capture fine emotion categories from tweets, 2015.

[32] Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text - behavior research methods, Jul 2020.

[33] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[34] Justin Garten, Reihane Boghrati, Joe Hoover, Kate M. Johnson, and Morteza Dehghani. Morality between the lines : Detecting moral sentiment in text. 2016.

[35] Anna Jurek, Maurice D. Mulvenna, and Yaxin Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4:1–13, 2015.

[36] Matteo Vergani and Ana-Maria Bliuc. The evolution of the isis' language: a quantitative analysis of the language of the first year of dabiq magazine. 2015.

[37] Tawunrat Chalothorn and Jeremy Ellman. Using sentiwordnet and sentiment analysis for detecting radical content on web forums. 2012.

[38] Shadi Ghajar-Khosravi, Peter J. Kwantes, Natalia Derbentseva, and Laura Huey. Quantifying salient concepts discussed in social media content: A case study using twitter content written by radicalized youth. *Journal of terrorism research*, 7:79, 2016.

[39] Daniel López-Sáncez, Jorge Revuelta, Fernando de la Prieta, and Juan M. Corchado. Towards the automatic identification and monitoring of radicalization activities in twitter. In Lorna Uden, Branislav Hadzima, and I-Hsien Ting, editors, *Knowledge Management in Organizations*, pages 589–599, Cham, 2018. Springer International Publishing.

[40] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F. Smeaton. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. ASONAM '09, page 231–236, USA, 2009. IEEE Computer Society.

[41] Hassan Saif, Thomas Dickinson, Leon Kastler, Miriam Fernández, and Harith Alani. A semantic graph-based approach for radicalisation detection on social media. In *Extended Semantic Web Conference*, 2017.

[42] A. Abbasi and Hsinchun Chen. Affect intensity analysis of dark web forums. *2007 IEEE Intelligence and Security Informatics*, pages 282–288, 2007.

[43] Hsinchun Chen. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *2008 IEEE International Conference on Intelligence and Security Informatics*, pages 104–109, 2008.

[44] Prateek Dewan, Anshuman Suri, Varun Bharadhwaj, Aditi Mithal, and Ponnurangam Kumaraguru. Towards understanding crisis events on online social networks through pictures. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 439–446, New York, NY, USA, 2017. Association for Computing Machinery.

[45] Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, and Nico Prucha. Detecting jihadist messages on twitter. In *2015 European Intelligence and Security Informatics Conference*, pages 161–164, 2015.

[46] Swati Agarwal and Ashish Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, 2015.

[47] Miriam Fernandez and Harith Alani. Contextual semantics for radicalisation detection on twitter. 2018.

[48] FRANCISCO DE and TORRE DÍAZ. La tributación del software en el irnr. algunos aspectos conflictivos. *Cuadernos de Formación. Colaboración*, 22(10), 2010.

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

[50] J. Fernando Sánchez-Rada. Design and Implementation of an Agent Architecture Based on Web Hooks. Master's thesis, ETSIT-UPM, 2012.

[51] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[52] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 154–160, New York, NY, USA, 2022. Association for Computing Machinery.

[53] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press, 2013.

[54] Jean Decety, Robert A Pape, and Clifford Ian Workman. A multilevel social neuroscience perspective on radicalization and terrorism. *Social Neuroscience*, 13:511 – 529, 2018.

[55] Yelena Mejova, Kyrieki Kalimeri, and Gianmarco De Francisci Morales. Authority without care: Moral values behind the mask mandate response, 2023.

[56] Scott Clifford and Jennifer Jerit. How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, 75, 07 2013.

[57] Livia Teernstra, Peter van der Putten, Liesbeth Noordegraaf-Eelens, and Fons Verbeek. The morality machine: Tracking moral values in tweets. In Henrik Boström, Arno Knobbe, Carlos Soares, and Panagiotis Papapetrou, editors, *Advances in Intelligent Data Analysis XV*, pages 26–37, Cham, 2016. Springer International Publishing.

[58] Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, 11(1):1–14, 2014.

[59] Eyal Sagi and Morteza Dehghani. Measuring moral rhetoric in text. *Social Science Computer Review*, 32(2):132–144, 2014.

[60] Eyal Sagi and Morteza Dehghani. Moral rhetoric in twitter: A case study of the u.s. federal shutdown of 2013, May 2017.

[61] Rishemjit Kaur and Kazutoshi Sasahara. Quantifying moral foundations from various topics on twitter conversations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2505–2512, 2016.

[62] Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods, Instruments, and Computers*, 50(1):344–361, 2 2018.

[63] Joe Hoover, Kate Johnson, Reihane Boghrati, Jesse Graham, and Morteza Dehghani. Moral framing and charitable donation: integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, 4(1), 2018. Funding Information: This work has been funded in part NSF IBSS 1520031. Publisher Copyright: © 2018 The Author(s).

[64] Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. Incorporating demographic embeddings into language understanding. *Cognitive Science*, 43(1):e12701, 2019.

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

[66] Óscar Araque and Carlos Angel Iglesias. An ensemble method for radicalization and hate speech detection online empowered by sentic computing. *Cognitive Computation*, 14:48–61, 2021.

[67] Mariam Nouh, Jason R. C. Nurse, and Michael Goldsmith. Understanding the radical mind: Identifying signals to detect extremist content on twitter. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 98–103, 2019.

[68] Dongwen Zhang, Hua Xu, Zengcai Su, and Yunfeng Xu. Chinese comments sentiment classification based on word2vec and svmperf. *Expert Syst. Appl.*, 42:1857–1863, 2015.

[69] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[70] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[71] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *International Workshop on Semantic Evaluation*, 2018.

[72] Cynthia Van Hee, Els Lefever, and Veronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *International Workshop on Semantic Evaluation*, 2018.

[73] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[74] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification, 2016.

[75] Oscar Araque, Ganggao Zhu, Manuel García-Amado, and Carlos Angel Iglesias. Mining the opinionated web: Classification and detection of aspect contexts for aspect based sentiment analysis. In Carlotta Domeniconi, Francesco Gullo, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu, editors, *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain*, pages 900–907. IEEE Computer Society, 2016.

[76] Robert Pelzer. Policing of terrorism using data from social media. *European Journal for Security Research*, 3(2):163–179, 2018.

[77] Mengyao Xu, Lingshu Hu, and Glen T Cameron. Tracking moral divergence with ddr in presidential debates over 60 years. *Journal of Computational Social Science*, 6(1):339–357, 2023.

[78] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[79] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[80] Geert Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* sage, 2001.

[81] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[82] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.