UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF A MACHINE LEARNING SYSTEM FOR THE DETECTION OF PROPAGANDA IN RADICAL TEXTS USING TRANSFORMERS

PABLO REAL BAEZA

JUNIO 2022

TRABAJO DE FIN DE GRADO

Título:	Diseño y Desarrollo de un Sistema basado en Aprendizaje Automático para la Detección de Propaganda en Textos Radicales mediante Transformers
Título (inglés):	Design and Development of a Machine Learning System for the Detection of Propaganda in Radical Texts using Transformers
Autor:	Pablo Real Baeza
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

DESIGN AND DEVELOPMENT OF A MACHINE LEARNING SYSTEM FOR THE DETECTION OF PROPAGANDA IN RADICAL TEXTS USING TRANSFORMERS

Pablo Real Baeza

Junio 2022

Resumen

La difusión de Propaganda ha sido una actividad generalizada a lo largo de la historia. Sin embargo, en los últimos años el auge the ideologías basadas en contenidos radicales que hacen uso de esta actividad ha empezado a preocupar a las instituciones europeas, las cuales han comenzado a financiar proyectos como H2020 Participation para controlar la expansión de estas ideas.

Asimismo, la limitación a la libertad de prensa que Rusia ha puesto en práctica como resultado a la invasión de Ucrania ha demostrado que este problema ha de ser seriamente considerado.

Por ello, el objetivo de este proyecto es encontrar una solución para detectar Propaganda en un entorno radical y/o violento, gracias al Aprendizaje Automático, específicamente a las técnicas de Procesamiento de Lenguaje Natural (PLN) que hacen uso de Transformers.

Para así conseguir el fin anteriormente mencionado, diferentes módulos de Python han sido usados para procesar datos y visualizar los resultados. Destaca la presencia de SHAP, implementado para obtener un enfoque más intuitivo, extrayendo la información adquirida durante el entrenamiento de todas las palabras incluidas en las fuentes de datos.

El aspecto más reseñable de este proyecto es el uso de Tranformers en los diferentes clasificadores. Por ello, su arquitectura ha sido estudiada minuciosamente, seleccionando aquellos aspectos más relevantes para así comprender su potential.

Además, cuatro fuentes de datos han sido utilizadas para la detección de Propaganda en varios temas, incluyendo el COVID-19, las elecciones presidenciales de EEUU o el yihadismo. Por lo demás, los resultados son mostrados de forma gradual, empezando por aquellos ejemplos más particulares antes de poder ver el aprendizaje con los datasets completos.

Finalmente, se han extraído diversas conclusiones, analizando los objetivos originales y describiendo los problemas más relevantes, considerando posibles mejoras para el futuro.

Palabras clave: Propaganda, Aprendizaje Automático, PLN, Transformers, SHAP, entrenamiento, clasificador, Dataset

Abstract

The diffusion of Propaganda has been a generalized activity over the course of history. Nevertheless, in recent years the growth of ideologies based on radical contents that make a repeated use of this activity has started to concern the European institutions, who have started to finance projects as H2020 Participation to control the expansion of these ideas.

Moreover, the limitation to the freedom of press which Russia has executed as a result of the invasion of Ukraine has proven that this issue has to be seriously considered.

Therefore, the objective of this project is finding a solution to detect Propaganda inside a radical an/or violent context, by means of Machine Learning, specifically Natural Language Processing (NLP) techniques which make use of the Transformers technology.

As a means to achieve the aforementioned aim, different Python modules have been used in order to process data and visualize the results. Noteworthy the presence of SHAP, implemented in order to give a more intuitive approach, extracting the information which has been acquired when training from every word included in the data sources.

The most significant aspect of this project is the use of Transformers in the different classifiers. Therefore its architecture is thoroughly developed, selecting its most relevant parameters so as to comprehend its potential.

Additionally, four data sources have been used to approach the detection of Propaganda from various topics, including COVID-19, 2020 US Presidential Election or jihadism. Otherwise, the results are displayed gradually, starting from particular examples before showing the outcome of training with the whole datasets.

In the end, several conclusions have been extracted, analyzing the original objectives and describing the most relevant issues that have been faced, considering possible improvements for the future.

Keywords: Propaganda, Machine Learning, NLP, Transformers, SHAP, train, classifier, Dataset

Agradecimientos

Me gustaría agradecer la confianza que han mostrado en mí las siguientes personas:

A mi tutor, Óscar Araque Iborra, por ayudarme a sumergirme en el desconocido mundo de la inteligencia artificial y tener la paciencia suficiente para demostarme que esto era posible.

Al GSI y en especial a su Director, Carlos Ángel Iglesias, por ofrecerme todas las facilidades, incluyendo el acceso al laboratorio donde disponía de todos los recursos necesarios para realizar este proyecto.

A mis compañeros de laboratorio: Diego, Inés y Sara, por hacerme disfrutar de esos momentos libres tomando el café matutino cuando no podíamos más.

A mis padres, por apoyarme y darme todas y cada una de las facilidades que he necesitado para finalizar esta etapa.

A mis amigos, porque gracias a esta universidad he conocido a gente maravillosa, en especial a Santi, Gabri y Javi porque sé que son para toda la vida.

Y a Andrea, por darme la ayuda final que necesitaba para terminar este trabajo y, sencillamente, por hacerme feliz.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

The veracity of the news which we read is one of the main topics of discussion in nowadays society. As events happen rapidly, the media does not have enough time to verify every article which is published. Furthermore, an increasing number of people, especially youngsters, tend to inform themselves by means of the social media, instead of traditional newspapers, even including their online versions. The main issue regarding this content is that it is simply not verified at all, circumstance which can lead to the diffusion of alternative versions of an event, the "fake news".

This environment, perfect breeding ground for radical content, is where the European Competitive Project H2020 Participation was born [3]. Inside it, both narratives and dynamics of expansion of propaganda in social media and Internet are studied, in order to prevent that a wider group of people is attracted by this radical content.

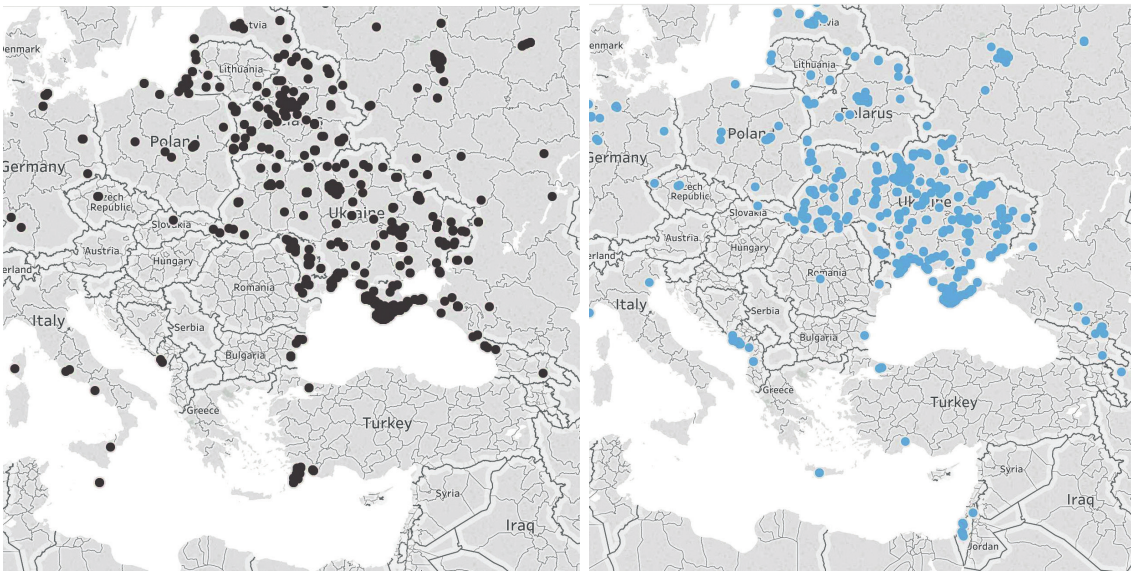
Given this circumstance, the importance of this project for the European Union is decisive, considering that a considerable number of these ideas are clearly contrary to the main principal of the EU since its foundation, democracy. Therefore their expansion, especially inside European territory, is a significant threat for the future of the Union.

In addition, recent studies [4] reveal that the potential heterogeneity of the social media networks does not necessarily start an enriching discussion where distinct opinions are described, since the citizens may be limited to search for supportive news which emphasise their perspective and radically counter their opposite.

Consequently, as the single target of this content is to be contrary to an idea, this disagreement can produce a devaluation of the politics, which enhances the scepticism towards democracy. Nevertheless, it has been noticed that this factor has a shorter impact when referring to institutional websites or traditional media, since their publications are generally aligned to more democratic perspectives.

Furthermore, the importance of the diffusion of propaganda is currently even more significant in the European context, considering that it is one of the main weapons which is used by Russia against the Western countries, especially nowadays when waging a war against Ukraine. However, this praxis is nothing but new, as it has been used against the former Soviet states, including Estonia, Latvia, Lithuania or the proper Ukraine almost since the fall of the Soviet Union [1].

Moreover, as seen in Figure 1.1, the use of propaganda by Russia is especially considerable in this country since the annexation of Crimea in 2014, particularly on this region and recently at Donetsk and Lugansk, where the roots of the pro-Russian supporters are deep.



(a) Pro-Russia (b) Pro-Ukraine

Figure 1.1: Comparison of Activist Community Geotagged Tweets [1]

Thus, given that the expansion of propaganda is even more accused in the previous months, due to to this military conflict, the development of classifiers that detect this content should be taken into account in order to solve this issue. Moreover, considering that they can be implemented by means of the most recent technology of the Natural Language Processing (forward referenced as NLP) field.

1.2 Project goals

The objectives of the project include the following:

1. To apply the acquired knowledge of the Transformer [2] by training a series of deep learning algorithms in order to predict both propaganda and radical content.
2. To obtain a adequate result of this training that would demonstrate the potential of this tool.
3. To understand the behaviour of the Transformer by examining the relationships that are created among the different words of a sentence.
4. To extract a number of dictionaries where the knowledge acquired when training is applied.

Therefore, as a result of completing these tasks I will expand my understanding of Machine Learning, concretely of NLP, using one of the most promising tools which is available nowadays. Furthermore, these trained algorithms will be published in an open source repository so that their performance can be boosted in the future.

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 provides a general overview of the project, bringing it closer to nowadays social and political context and it presents the objectives which should be acquired by its ending.

Chapter 2 describes the tools which are used to fulfil this project, including not only their origin and behaviour but also their application.

Chapter 3 expounds the structure of the model which is used in this project, providing also the historical context before it was published.

Chapter 4 shows the required procedure to train the algorithms according to the detection of propaganda, extracting information from the different datasets and defining the metrics that evaluate the result.

Chapter 5 exposes the results obtained by training the algorithms, and offers a didactic approach in order to comprehend the behaviour of the applied models.

Chapter 6 extracts a series of conclusions from the project, verifying whether the objectives are achieved and explaining the issues which might be faced, offering possible solutions for the future.

CHAPTER 2

Enabling Technologies

In this chapter the different technologies which are required will be explicated, considering the particular applications of them onto the project:

2.1 Python

Python is one of the most popular high-level programming languages nowadays. Since its first release, introduced in 1991 by Guido Van Rossum, it has evolved rapidly. In fact, according to the magazine IEEE Spectrum [5], it is the most useful programming tool for the IEEE members, being the Artificial Intelligence one of its main fields of use.

Whereas R, SQL or Matlab are still valuable for solving specific problems, Python is ahead of other generalized programming languages, including Java, C or C++, as it is shown in Figure 2.1. Moreover, as far as Machine Learning is concerned, its usage is even more substantial considering a number of reasons, including not only a large number of libraries (NumPy, scikit), frameworks (Tensorflow) and visualization tools (Matplotlib), which will all be introduced afterwards; but also its conciseness and ease to be read or the considerable support which can be found online.

Figure 2.1: IEEE Spectrum. Top Programming Languages in 2021

2.1.1 Pandas

Pandas is the largest open source Python library in terms of managing large amounts of data [6]. Indeed, its main use according to this project is not only the manipulation of the datasets where all the data is stored, selecting the pieces of information which are relevant to the aim of the project and discarding those which are not needed, but also the fit of these data into the requirements of the training algorithm. Consequently, its use in the field of NLP is key.

Specifically, Pandas introduces two basic structures: Series and Dataframes. Whereas Series are applied to 1-dimensional arrays, Dataframes include a great amount of facilities according multi-dimensional structures. Therefore, these Dataframes are a perfect fit in order to process data, including the selection of certain columns.

Furthermore, Pandas also provides tools in order to eliminate certain pieces of information which are not needed, including rows which content no data or truncate considerably long phrases which the algorithm can not process.

2.1.2 NumPy

Pandas has been developed on top of another Python package, NumPy, which target are arrays of multiple dimensions [7]. In terms of this project, it is mainly used when performing mathematical operations to certain values of the data which must be processed.

Therefore, this package must be used at certain tasks, including the normalization of the values of the labels, considering they might not be integers and there must correspond to a certain value (normally, 0 or 1, as the aim of the project is to distinguish just propaganda from neutral content).

Moreover, NumPy is needed when extracting a dictionary that includes a classification of all the words which the algorithm has detected when training. Concretely, there will be various words that appear several times on a certain dataset, considering that it is not useful to handle several values of a label (normally, propaganda) to the same term.

2.1.3 Matplotlib

Matplotlib is a library which is designed in order to create elegant and visual plots in Python [8]. It allows to adjust the size, axis, title, legend or select the number of samples which are included in a visualization, among many other tasks. As far as this project is concerned, it provides a more intuitive approach to the results which have been obtained, by visualizing not only the value of most relevant words of a dataset but also a graphical representations of these values.

2.1.4 Scikit-learn

This library, built on the previously explained NumPy and Matplotlib, contains several tools which are helpful in terms of the Machine Learning field [9], including classification, regression or preprocessing data. In fact, it has been used for several tasks, including the split of the data into the three main sets which are needed, that is, training, development and testing.

In addition, Scikit-learn is useful when comparing the results which have been obtained among the training of different datasets, even with the results which other researchers have been obtained in the past by training these datasets using other techniques.

2.1.5 glob, os

Glob and os are two Python modules which are used to manipulate paths. Whereas glob joins all the data which is contained among the different files which are located in a certain path, os can be used to split the name of the path which is needed. These two modules are used when reading a dataset where all the samples are distributed among a great amount of files.

2.2 Natural Language Processing (NLP)

NLP is a field which encompasses linguistics, Computer Science and Machine Learning, by programming algorithms which are able to process a wide amount of natural language data. This area includes performing different tasks, including not only the classification of texts according to an specific feature, but also the answering of questions, summarizing or even translating fragments of text.

Nevertheless, this project is focused on the first of the above-mentioned tasks, by training a series of algorithms in order that they can classify whether a phrase is propaganda or not. With that objective, we have been used the most recent technology in order to obtain the best result possible, that is, the Transformer.

2.2.1 Transformers

Transformers is a deep learning model, mainly used in the field of NLP, which is completely based on self-attention, that is, associating different positions of a sequence as a means to compute a representation of it. It supports integration among the three most used deep learning libraries, PyTorch, Tensorflow and JAX.

In terms of its NLP applications, Transformers is generally introduced in language models. Nevertheless, there are not only text-based models available in Transformers but also image, audio or multimodal ones. In this project, the importance of Transformers is key, considering that these datasets have already been used to train algorithms, however the expected results should be considerably superior in terms of efficiency and accuracy. Furthermore, a deep explanation about the architecture of this powerful tool can be seen in Chapter 3.2.

2.2.2 SHAP

Shapley additive explanations (SHAP) consists of a visual approach in order to describe the output of a Machine Learning model, allowing the user to have a clear understanding of how a model is trained. In terms of NLP, it associates every word which is contained on a trained model to a value that represents the amount of an specific feature which that phrase is introducing to the text.

According to this project, these shap values can be not only represented according to its importance, but also stored in a dictionary, which is very useful in terms of taking advantage of the knowledge which is already acquired by training.

2.2.3 GSITK

GSItk is a Python library which works on top of the previously explicated Scikit-learn, NumPy and Pandas which facilitates the development process on NLP projects [10] [11]. It is capable of treat datasets, classifiers and evaluation techniques, therefore its facilities for this project are considerable.

Concretely, this library has been implemented when preprocessing the pieces of text which form every dataset, by eliminating not necessary spaces between words or the uppercase characters, as this algorithm is focus on the content of every phrase.

CHAPTER 3

Architecture

In this chapter, the design phase of this project is covered, as well as the implementation details involving the architecture of the Transformer. However, in order to have a clear understanding of these architecture, a brief introduction to the evolution of machine learning and NLP during the 20th century will be discussed in the first place [12].

3.1 Historical context

As soon as in the beginnings of the previous century, the Russian Andréi Markov introduced his works on stochastic processes, including the concept which is referred nowadays as the Markov Chain. A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. This mathematician applied his concept to a dataset formed by 20,000 letters, in order to predict the future elements of the chain, just by using the last past one.

Several years later, in 1948, a reference work for the Telecommunications field was published, that is, *The Mathematical Theory of Communication*, written by Claude Shannon. As a result, Markov´s Chains were used in order to model communications with 3 main

elements: encoder, transmitter and decoder.

Afterwards, in 1950 Alan Turing´s article *Computing Machinery and Intelligence* condenses the knowledge acquired when building the famous Turing Machine which decrypted the messages from the German Enigma in the Second World War. In fact, this was the first implementation of artificial intelligence, although this concept was not considered until 1956.

On the field of language translation, a experiment carried out jointly by the Georgetown University and IBM involved automatic translation of several Russian sentences into English, by means of creating a program which run through a list of rules which are known in all languages. Nevertheless, this method seems to be not useful nowadays, as there may be billions of rule combinations around the net so as to make a list which includes all of them.

A few years later, in 1982, the Recurrent Neural Network (RNN) is introduced by John Hopfield. A RNN is a type or artificial neural network where there are connections among the different nodes which result in a temporal sequence, as it can bee seen in Figure 3.1. It memorizes the state of the sequence in order to process variable length inputs. Furthermore, in this years the Convolutional Neural Network (CNN), was introduced, which is mainly applied to analyze visual images and differs from the previous in the fact that it has a finite impulse response, whereas it is infinite in the case of the RNN.

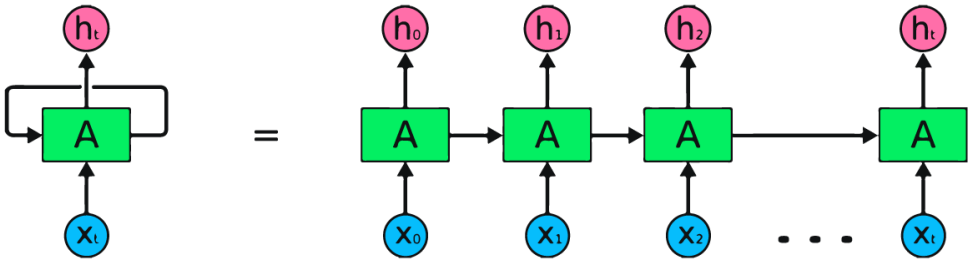


Figure 3.1: Architecture of the RNN

3.2 Transformers

Although RNNs have been extremely useful until the recent years, the publication in 2019 of the article *Attention is all you need* started a new perception in the development of the Artificial Intelligence, and more concretely, of NLP, with the first description of the Transformer [2]. Firstly, a general overview of the model is presented. This is intended

to offer a general view of the architecture and the main reasons for its applicability in this field. After that, each sub-module which composes the global architecture will be developed separately and in much more depth.

The Transformer divides the input data, weighting the importance of every part in the whole context. It differs from the traditional RNNs in the fact that Transformers does not demand to process the sequential data in its entry order. Therefore, it can predict context for words which are placed in every position of the sentence.

Precisely, the RNN mainly take into account the symbol positions of both input and output sequences, by matching positions into steps in computation. However, the main obstacle of this sequential method is that it is incompatible with a key feature when training long sequences: parallelization.

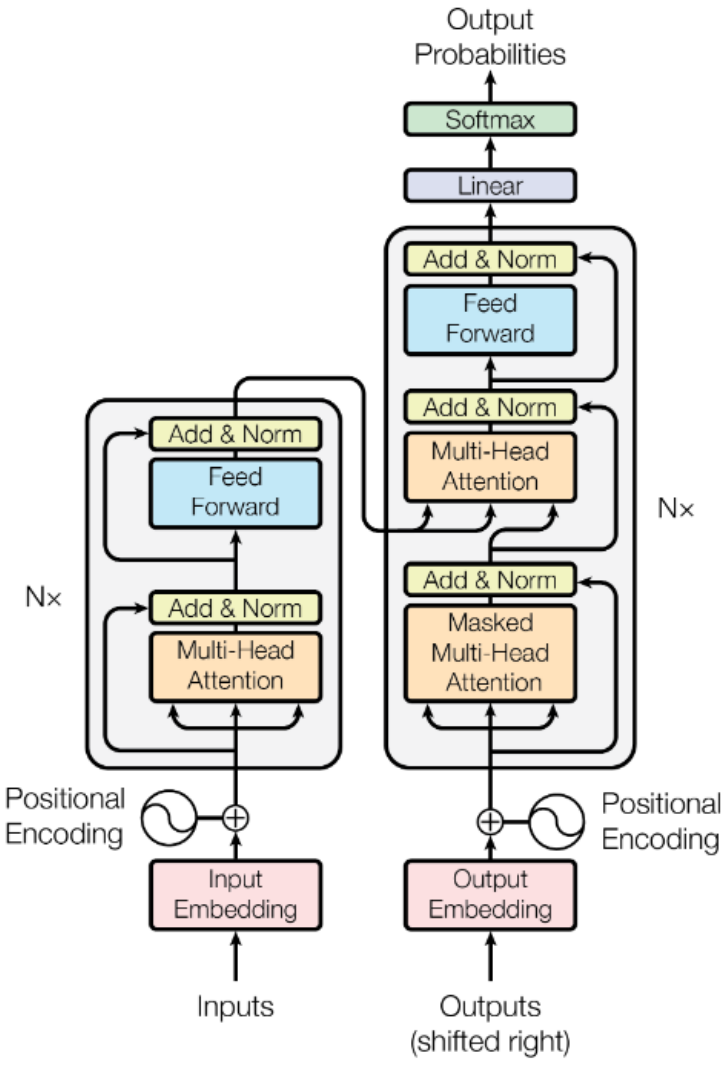


Figure 3.2: Model architecture of the Transformer [2]

As it is explained, the transformer model architecture avoids sequential RNN or convolution, depending just in the relationships which can be formed among every single feature of input and output. As far as it is concerned, it is the first deep learning model to have achieved this objective fully.

Similarly to the most valuable neural models, the Transformer is based on an encoder-decoder structure, where the first allocates the input symbols to a continuous sequence, while the second generates the output by means of this continuous sequence, taking into account that the model absorbs the previously generated symbols as an input in order to produce the next one.

Nevertheless, as seen in Figure 3.2, the essential characteristic of the Transformer architecture is that it includes stacked self-attention, as it has already been explained (in fact, both encoder and decoder are formed by N=6 layers). Specifically, each layer of the encoder is composed by two sub-layers, which correspond to a multi-head attention and a positioning mechanism. Furthermore, the decoder adds a third sub-layer which executes multi-head attention over the output of the encoder pile.

In the following sections, each one of these sub-layers which for the architecture of the Transformer will be developed. So as to fulfil this aim, it is worth mentioning that the original description of Transformers selected a 512-dimensional model ($d_{model} = 512$). Although this parameter can be changed when training, this precise number will be considered for the rest of this chapter.

3.2.1 Input embedding

In the first place, the input text is passed through a tokenizer, where it is normalised, that is, it is lower-cased and truncated into subparts, which are then converted into an integer representation. Afterwards, each input token is converted into a 512-dimensional vector by means of a learned embedding sub-layer. These vector contains just the meaning of the word, not the context of its position on the sequence.

For instance, as it is explained in [12], if the following sequence is selected as an input: *The black cat sat on the couch and the brown dog slept on the rug*, and the dot product of the 512-dimensional vectors which are obtained after tokenizing and embedding the words *black* and *brown* is performed, it will concluded that this vectors are really similar, as both represent a color. However, these vectors have not included the information regarding the rest of the words in the sentence, which will be added by the positional encoding function.

3.2.2 Positional encoding

There are a number of solutions in order to obtain the information of the position of a word. As it was originally explained in [2], in the Transformer model a couple of functions which consider even and odd numbers of the dimension of the model separately, which create different frequencies for each position and dimension of the model.

$$PE_{(pos2i)} = \sin \frac{pos}{1000^{\frac{2i}{d_{model}}}} \tag{3.1}$$

$$PE_{(pos2i+1)} = \cos \frac{pos}{1000^{\frac{2i}{d_{model}}}} \tag{3.2}$$

Therefore, a 512-dimensional vector is obtained for each position in the sentence, which can be added to the word embedding vector as both have the same number of dimensions (the dimension of the model of the Transformer, $d_{model} = 512$).

3.2.3 Multi-head attention

As it has already been explained, the main aspect of the Transformer model is attention. Moreover, in order to compute this parameter, the input sequence if divided into 8 *heads* in parallel in order to speed up the training. Therefore, this parallelization of the tasks is a crucial advancement in comparison to the traditional RNNs.

Afterwards, for each *head*, attention is computed separately and then concatenated. This parameter is calculated by means of the following equation, where each word vector has three representations, by means of a query (Q), a key (K) and a value (V) 64-dimensional vectors:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3.3}$$

3.2.4 Feed-forward neural network

After the multi-head attention task, the Transformer includes a feed-forward neural network (FNN). This also happens when implementing a usual RNN structure, where in the last layers a FNN, which is fed with the output of the recurrent structure, is generally introduced.

Precisely, the FFN which is included in this layer of the Transformer model has the following characteristics: it is a position-wise network, which contains two layers and can be interpreted as performing two convolutions of one kernel size.

3.2.5 Post-layer normalization

As it can be seen in Figure 3.2, each attention and feed-forward sub-layer is followed by a *Add & Norm* sub-layer, which represents the post-layer normalization which is performed to every vector in order to obtain real probabilities, which are computed by the following equation:

$$LayerNorm(v) = \gamma \frac{v - \mu}{\sigma} + \beta \tag{3.4}$$

where μ is the means deviation of the input v, σ its standard deviation, γ a scaling parameter and β a bias vector.

CHAPTER 4

Methods

In this chapter it is going to be analyzed the data which is used, the metrics which must be considered in order to evaluate the performance of the algorithms and the sequence of steps which should be taken when training one of these classifiers.

4.1 Data acquisition

In order to build an algorithm that can differ Propaganda from neutral content, a series of datasets must be used. A Dataset is an aggregation of tabulated data in a storage system with structured data. Indeed, it is represented by a matrix, where the rows represent the different inputs of data and the columns the variables which should be considered about this data.

In this project, not all these variables have been taken into account, for instance, author, place or date of the publication are not relevant information in order to select propagandistic or violent content. Concretely, four datasets have been used, where different topics have been approached although there is a common feature among all, the possible appearance of Propaganda. The datasets which will be developed are the following:

4.1.1 PTC

The PTC corpus [13] has been created as a result of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection held in the conference SemEval.

It selects an specific propaganda technique among up to 18 possible options, which include the following: loaded language, name-calling or labeling, repetition, exaggeration or minimization, doubt, appeal to fear/prejudice, flag-waving, causal oversimplification, slogans, appeal to authority, black-and-white fallacy or dictatorship, through-terminating *cliché*, whataboutism, reductio ad Hitlerum, red herring, bandwagon, obfuscation or intentional vagueness or confusion, and straw man.

Nevertheless, the aim of this project is the detection of propaganda, not the division of these contents into subtopics. Therefore, when training the classifier which makes use of this dataset, the feature which contains a label representing the type of propaganda will not be considered, selecting just the one which represents whether a content is indeed propagandistic or not.

Moreover, the data must be divided into the three sets in the following way. In the first place, the data is divided into two sets: train and development, which are required by Transformers in order to train the classifier. In addition to the previous two, a third set (test) is needed as to verify the performance of the algorithm and verify the obtained metrics.

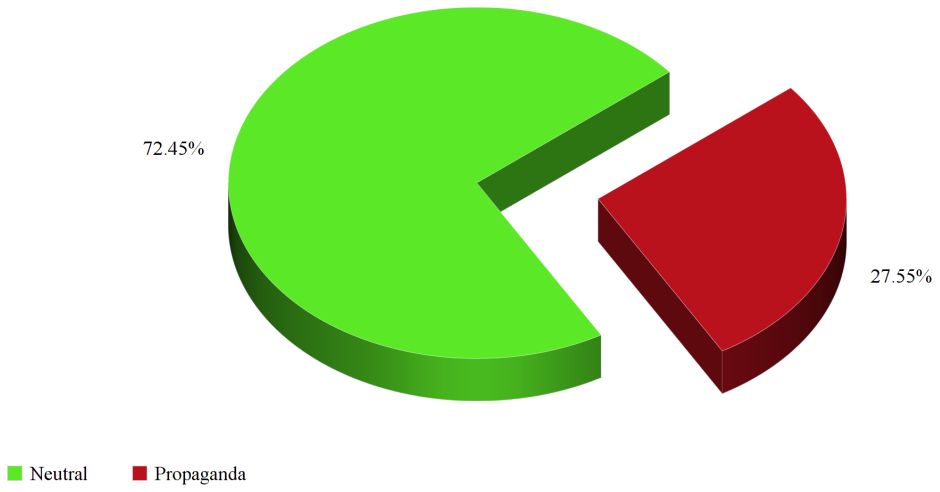


Figure 4.1: Distribution of PTC Dataset

4.1.2 NELA-GT

NELA-GT-2020 Dataset [14] is formed by 1.8M news articles, including tweets from 519 sources which are annotated for covering multiple dimensions of veracity. In particular, the dataset differs from the rest in the fact that it allocates the samples into three categories: Reliable, Mixed and Unreliable, therefore it measures their credibility.

On the other hand, the samples which form the rest of the datasets are just divided into two categories: Propaganda or Neutral content. Moreover, if a source is detected to have characteristics such as Conspiracy, Pseudoscience or Questionable, it will be associated to the Unreliable category.

The use cases of this data include two topics which have had a considerable impact on the society and there are a great number of accusation of using Propaganda: COVID-19 and the 2020 U.S. Presidential Election. Moreover, there is a third topic which is included in this dataset: Embedded Tweets. This field includes a several number tweets which have been added by the news publishers to their articles.

Furthermore, as this dataset includes a large amount of data, there have been selected a reasonable number of samples (concretely, 20.000) so that the training of the algorithm does not represent an excessive amount of time. Lastly, after selecting the samples, the data is split into three sets as it has already been explained previously.

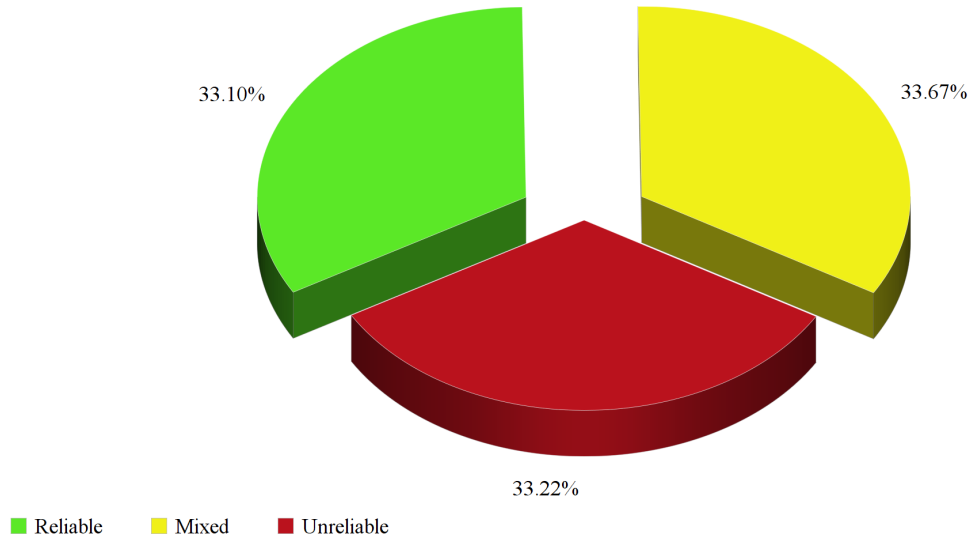


Figure 4.2: Distribution of NELA-GT Dataset

4.1.3 QProp

The Proppy corpus [15] is a dataset where the news are classified into Propaganda and Non-Propaganda. It is based on 52k articles labelled as either propagandistic or not. This classification, established by human experts, is assessed on the level of propagandistic content in an article based on different representations, from writing style and readability level to the presence of certain keywords.

The data has already been split into three different sets, i.e. training, developing and testing sets. Therefore, unlike the other three datasets, this division does not have to be done.

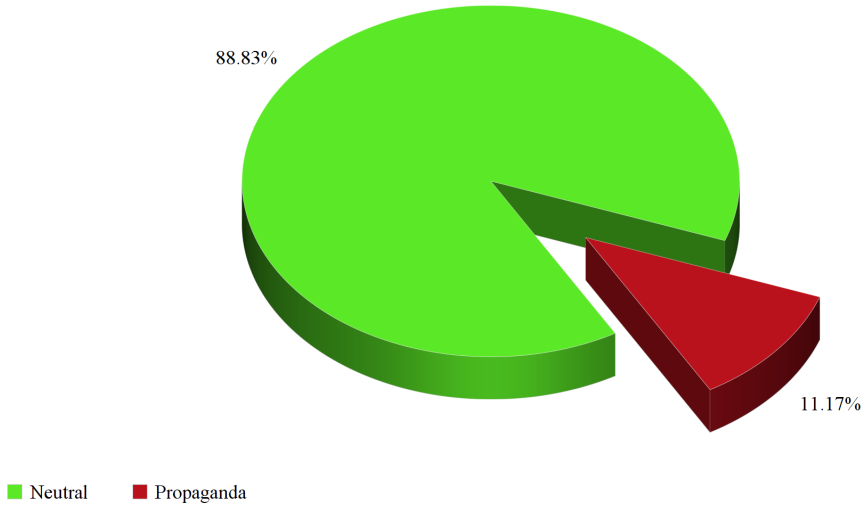


Figure 4.3: Distribution of QProp Dataset

4.1.4 Jacobs

Jacobs Dataset [16] is focused in the detection of extremism and terrorism, considering that cyber-recruiting in social media websites is one of the main means of expanding radicalism. This data, compiled by the Dark Web Project of the University of Arizona, has been collected from the western jihadist website *Ansar AlJihad Network* and labelled as Recruiting or Non-Recruiting, which represents whether the content can be used to attract new members or not. Therefore, although the content of the data differs from other datasets, which are closely related to news and social media, a Recruiting content can also by spotted as Propaganda.

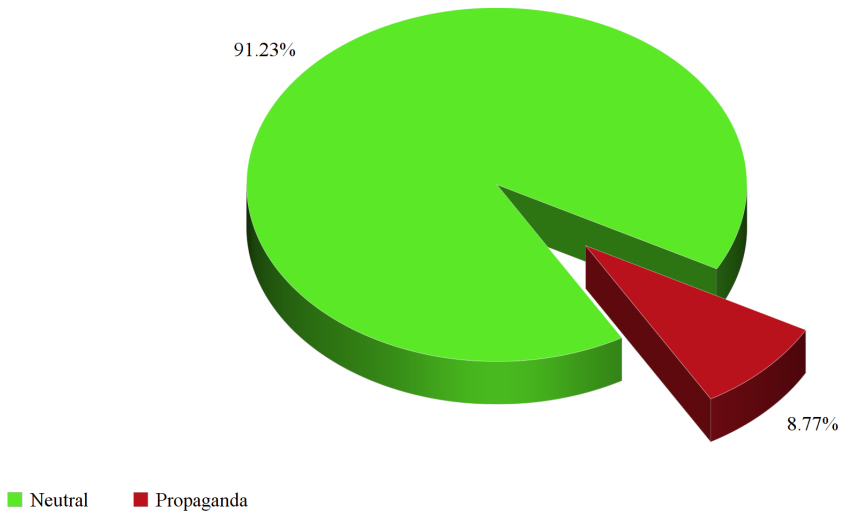


Figure 4.4: Distribution of Jacobs Dataset

4.2 Prediction Metrics

After understanding the differences among the datasets which are used in order to detect Propaganda, the metrics which are used when evaluating must be considered. With that aim, a confusion matrix is displayed in Figure 4.5, where a general representation of the data divided into four categories can be seen: the values which are selected as positive (true positives), the negative as negatives (true negatives), the positive values selected as negative (false positive) and finally the negatives selected as positives by the algorithm (false negatives).

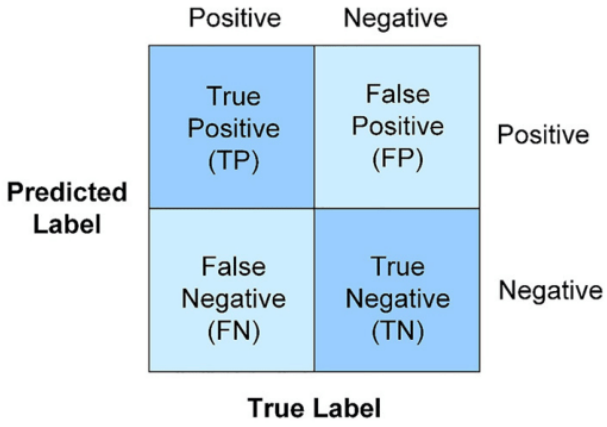


Figure 4.5: Representation of a Confusion Matrix

The metrics which are used by Transformers can be derived from the previously explained parameters of this matrix. Firstly, accuracy measures the correct values with respect to the total number:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Secondly, precision selects the amount of true positives out of the total number of positives:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Afterwards, recall compares the true positive values with the total number of real positive values:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Finally, f-score is obtained by means of the precision and the recall. This is the most useful metric when comparing the performance of different algorithms:

$$f - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

All this metrics will be computed by the Transformer when training each classifier. To find this data, go to Section 5.2.

4.3 Training Method

In order to train the different classifiers, a number of steps must be considered. In Figure 4.6 it can be seen the method that I have followed in order to train a propaganda classifier with Transformers. The steps are the following:

Figure 4.6: How to train a classifier with Transformers

First of all, the data which is needed must be extracted from every dataset. All of the datasets contain a great amount of information, regarding for instance author, place and date of publication of the text which is not needed in this project, where just the proper text and the label which represents whether the content is indeed propaganda is required. In addition, these labels may not be an integer, therefore there have to be approximated in order to assign the labels which are needed.

When the data is selected, it must be divided into three sets: train, development and test. The first two of them will be used in order to train the classifier, whereas the third will be used afterwards in order to verify the acquired knowledge and obtain the values of the metrics which are considered.

Later, the text must be preprocessed so that it is easier for the algorithm to train with it. In this stage, all the texts is lower-cased, trimmed and adapted to be used by Transformers. Additionally, the dataset are contained into Pandas Dataframes, which must be converted into a Dataset type to be used.

Once the previous steps have been completed, the words which are contained in the texts of the dataset are tokenized. Therefore, the text can be divided into small pieces which are converted into a numerical representation, as it is explained in Section 3.2.1.

Following the tokenization, prediction metrics which are going to be considered are imported. In this project, these metrics include accuracy, precision, recall and, most important, f-score. All this metrics are developed in the previous section.

After all the previous steps are finished, training can finally be done. In this project, it has been used a BERT sequence classification model as it is the one which produces the leading results [17]. Starting from this model, our propaganda classifier is built, and then tested by means of the testing set in order to obtain the corresponding value of the metrics.

Finally, the model can be uploaded into the Transformers website [18] in order to be used in the future.

CHAPTER 5

Results

In this chapter the results which are obtained when training the different classifiers will be described. However, instead of showing them directly it has been decided to execute in a gradual manner, first selecting a series of examples which are representative in order to understand how the different classifiers are instructed and then showing the actual results of training them with the whole set.

5.1 Selected examples

Although these results are relevant when comparing the performance among the four different classifiers, it is not an straightforward task to obtain explanations for the internal decisions of the model. In order to fulfil this aim, a more concrete approach must be considered, by selecting some relevant samples of the datasets and seeing if the classifier has a satisfactory result when predicting whether they are propaganda or not.

Nevertheless, rather than selecting a concrete text and introducing it ads an input to the model to verify if it predicts precisely its respective characteristic (whether it is propagandistic or not), SHAP will be used to provide a more visual approach to the issue. By applying this tool, it can be spotted not only if the fragment contains propaganda but

the specific words which are more relevant to this concern.

Therefore, four examples have been chosen, one from each dataset, which are representative in order to detect the words that the model has considered as more propagandistic. Consequently, the intention of this analysis is not the interpretation of the value which SHAP assigns for each word, but the relative comparison among these values in order to determine which of the words introduce a further uninformative approach.

5.1.1 Example in PTC

As the first example a significant topic has been introduced:

In so doing he would have intensified the demand of the people of God and the world for the documentation needed to determine who has told the truth.

In this case, the classifier can detect that the word *God* is related to religious texts, which are linked with subjectivism. The narratives where this topic is present can clearly not be defined as neutral, therefore, even it can be controversial to include religion in the set of Propaganda, this classifier has learned that this is the best solution. Furthermore, terms like *truth* are often used when someone wished to earn the respect of a group of people and this environment can be a perfect breeding ground for Propaganda.

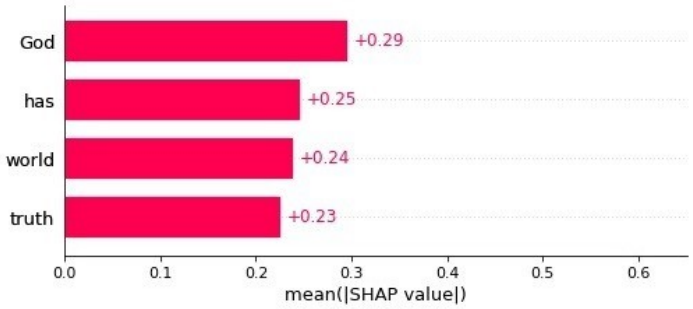


Figure 5.1: Most significant terms of the example included in the PTC Dataset

5.1.2 Example in NELA-GT

In this case the example addresses two significantly controversial topics: COVID-19 and the US Presidency:

Former Mayor Rudy Giuliani launched another conspiracy theory about the coronavirus, saying that former President Barack Obama gave funding to a Wuhan, China virology lab in 2017 from the U.S. budget. President Donald Trump took over the presidency Jan. 20,

2017 but the budget for that year was part of a battle between the Republicans running the House and Senate in 2016. At the time , Speaker Paul Ryan (R-WI) was in control of Congress and the Senate majority leader was Mitch McConnell (R-KY).

Therefore, when selecting the most relevant words, terms like *China* and *conspiracy* appear. In the first place, as this dataset contains articles which analyze the figure of the former US president, Donald Trump, *China* often appears in the propagandistic articles which have the intention to exalt feelings like the patriotism of the US. On the other hand, *conspiracy* is a clearly subjective word which has significantly been used by the supporters of Donald Trump, therefore it is included as Unreliable.

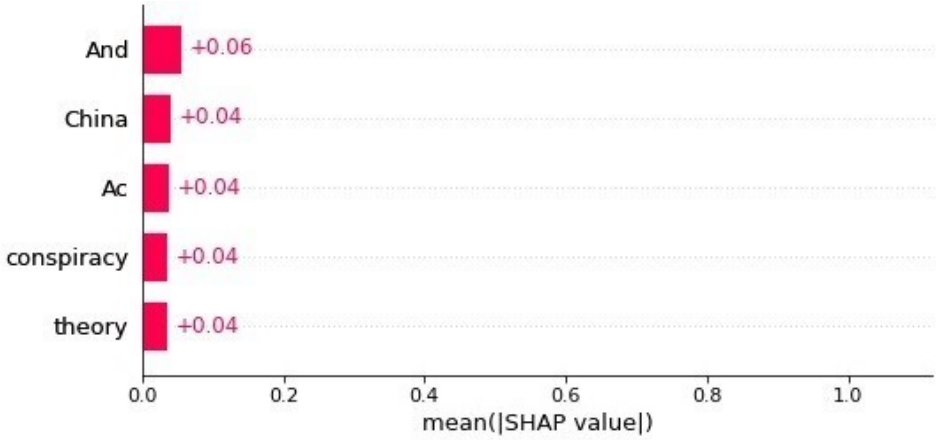


Figure 5.2: Most representative words of the example included in the NELA-GT Dataset

5.1.3 Example in QProp

In order to reach every possible situation, a significantly neutral example has been selected:

Two crashes, within six weeks of one another, caused the deaths of seven teenagers in the fall of 2007 and were among the many others that prompted the adoption of tougher teen-driving laws in 2008. Advocates at the state Capitol Thursday said the laws have saved the lives of other young drivers.

Actually, it comes from a news article in this case, where it is explained that seven adolescents died in 2007 as a result of several car-crashes. Therefore, considering that the fragment contains considerably objective information, the classifier has not included any significant word as Propaganda.

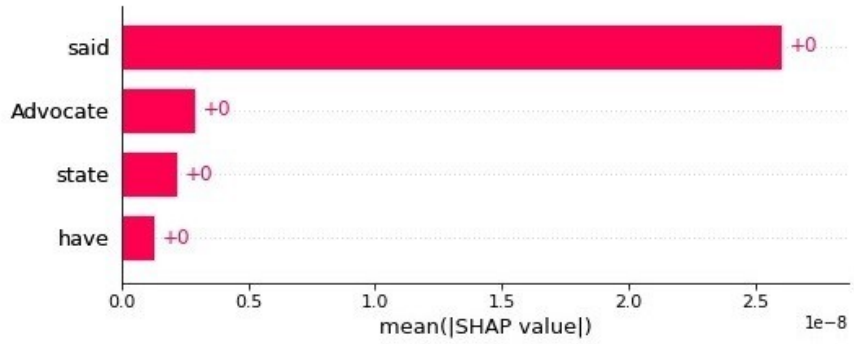


Figure 5.3: Most representative words of the example included in the QProp Dataset

5.1.4 Example in Jacobs

In the last dataset the following example has been selected:

Pope fuels controversy over apology to Muslims Sat, 09 May 2009 21:06:34 GMT Pope Benedict XVI (2nd R), accompanied by Jordan's Prince Ghazi bin Mohammed (R) Pope Benedict XVI has received mixed reactions from Muslim groups over his offensive remarks about Islam during his visit to Jordan. In 2006, the Pontiff had described certain Islamic teachings, for which he apologized to Muslim leaders at Amman's al-Hussein Mosque on Saturday.

Finally, when analyzing the last example, it can be appreciated that the two most relevant terms are *fuels* and *controversy*, which appear at the beginning of the fragment and can be clearly consider as propagandistic. Moreover, the first one is specially notorious considering this dataset includes a substantial amount of warlike topics, as its target is the radicalism and jihadism.

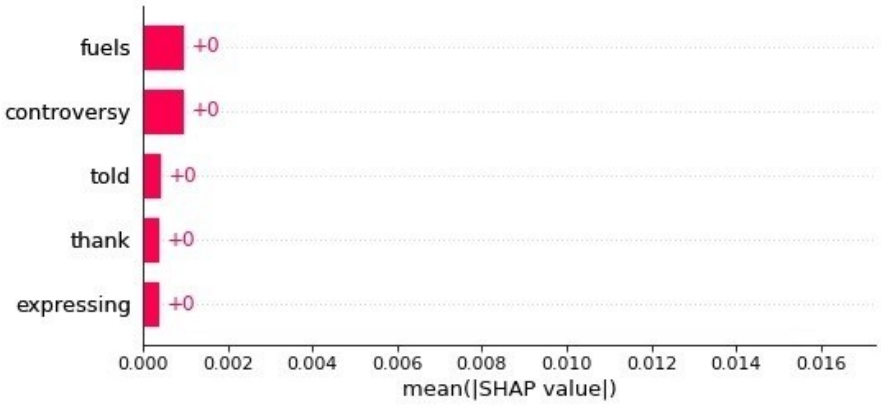


Figure 5.4: Most representative terms of the example included in the Jacobs Dataset

5.2 Classifier Results

Once the detailed perspective has been presented, the results that are obtained when training the classifier with the whole set of samples can be displayed. Therefore, we can not only observe values of the metrics that are obtained but also plot the representation of this data.

First of all, Table 5.1 represents the results that are obtained after executing the phases of training and testing the classifiers. It can be spotted that the overall performance is satisfying, just being it relatively poor in the case of the NELA-GT Dataset. This aspect might be explained by the fact that it is the only source of data which is divided into three different labels instead of two, which makes the task of the classifier slightly more challenging. Moreover, when analyzing the precision of the first classifier (QProp), it can be spotted that it is lower than the rest of the metric parameters. This circumstance is probably caused due to the excess of neutral samples in the dataset, in comparison to the ones which are labelled as Propaganda.

	QProp	PTC	NELA-GT	Jacobs
Accuracy	0.89	0.80	0.71	0.97
Precision	0.79	0.79	0.71	0.97
Recall	0.89	0.80	0.71	0.97
F-score	0.83	0.79	0.71	0.97

Table 5.1: Metrics obtained when training the different classifiers

Nevertheless, this table merely extracts an overview of the performance of the different classifiers. Therefore, each of them should be analyzed thoroughly in order to obtain a deeper view of the results. As a means to fulfil this goal, there have been extracted not only the most significant values of the dataset but also a representation of how the values which are assigned to the words are distributed.

Furthermore, it has to be considered that all of the values which are going to be displayed may not correspond to the objective of the classifiers, that is, the detection of propaganda. This situation is obtained considering that this classifiers are trained to classify large loads of data, therefore even the values assigned to every specific term may not be fully reasonable,

the aim is to predict propagandistic content in a satisfactory manner.

As explained in Section 1.2, one of the aims of this project is to summarize the acquired knowledge into a dictionary. Computing the SHAP values allows the user to copy them into a lexicon, so that they can be represented in a more effective way. Moreover, these dictionaries can be used in the future when training other models. Figure 5.5 shows a sample of this data, considering the great number of words which compose a dataset.


```
"the": -0.0016158604024288554,
"letter": 0.009929031650244724,
"also": 0.0201290762561257,
"contrasts": 0.06321174510958372,
"stance": -0.007855445990571752,
"of": 0.028256346049602143,
"islamic": -0.019214440998621286,
"countries": -0.029954321146942676,
"like": 0.010666570463217794,
"turkey": 0.013034366839565336,
"and": -0.016743097035214306,
"saudi": -0.02652661061802064,
"arabia": -0.035048313806328224,
"who": -0.06141841174758156,
```


Figure 5.5: Sample of the values stored in a dictionary

These values which are loaded in a dictionary can also be represented. Therefore, in the following sections the most significant outcomes will be displayed for every classifier, considering not only which words are more relevant but also how this values are distributed.

5.2.1 Results of PTC

In terms of the first dataset, the most significant terms which are obtained can be spotted below. By sketching these words, it can clearly be seen that terms including *ridiculous*, *horrified*, *absurd* or *bleak* are usually utilized in propagandistic articles, whereas *Congratulations* or *Number* appear to be objective.

Consequently, it can be asseverated that extreme feelings have a close relation with Propaganda. Whereas rather neutral articles circumscribe to inform or even offer an opinion but in a deferential manner, propagandistic articles generally make use of these feelings in order to attract the attention of the reader.

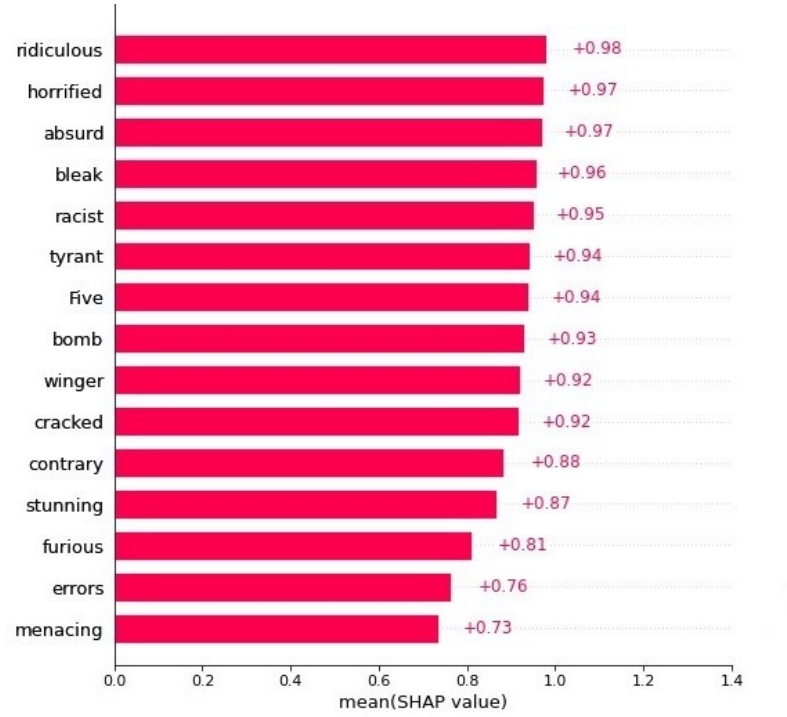


Figure 5.6: Most Propagandistic terms according to the PTC Dataset

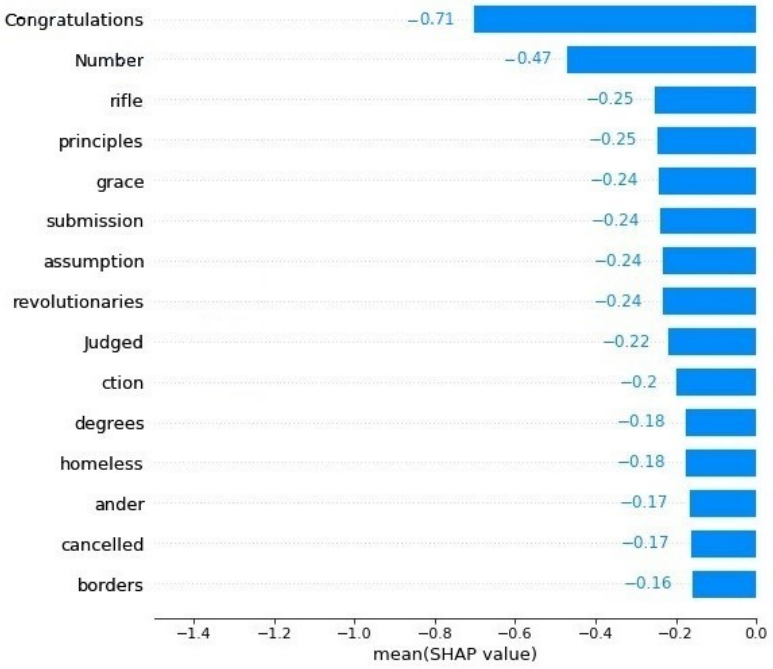


Figure 5.7: Least Propagandistic terms according to the PTC Dataset

As it will be seen in the following sections, the representation of the obtained values from all the classifiers are considerably similar. The data is concentrated in the origin and tends to follow the shape of a Gaussian with mean 0 and very low standard deviation, considering this concentration of values around the origin.

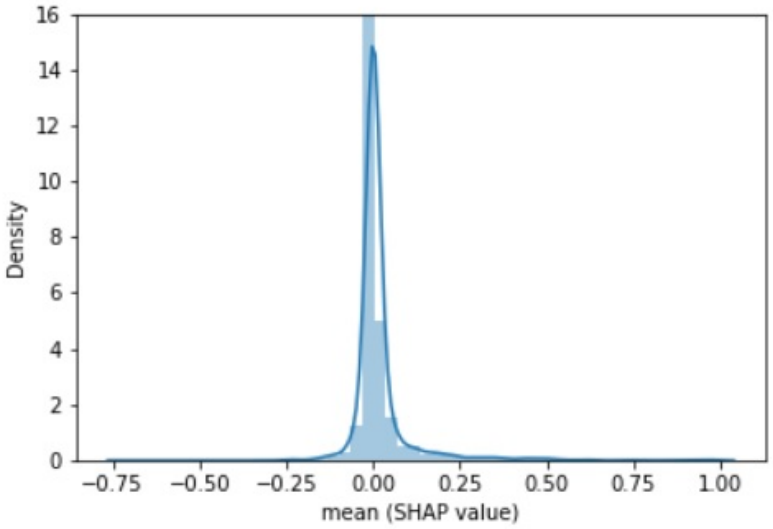


Figure 5.8: Representation of the distributed values obtained in the PTC Dataset

5.2.2 Results of NELA-GT

Unlike the rest of the classifiers, the one referred in this section has to distinguish the data into three categories: Unreliable, Mixed and Reliable, therefore the task is slightly more complex and the results are not as satisfying. By sketching the different terms which have been selected as representative for these three groups, any reasonable conclusions can be extracted.

Considering that the metrics which are obtained when training this classifier are slightly inferior to the rest, this issue may explain the apparent difficulty to extract clear data. Nevertheless, it has to be indicated that the aim of this these algorithms is the distinction of propagandistic pieces of information, not the classification of every word. This task has been done simply with the objective of showing how the classifiers operate, but it is not the final objective of the project.

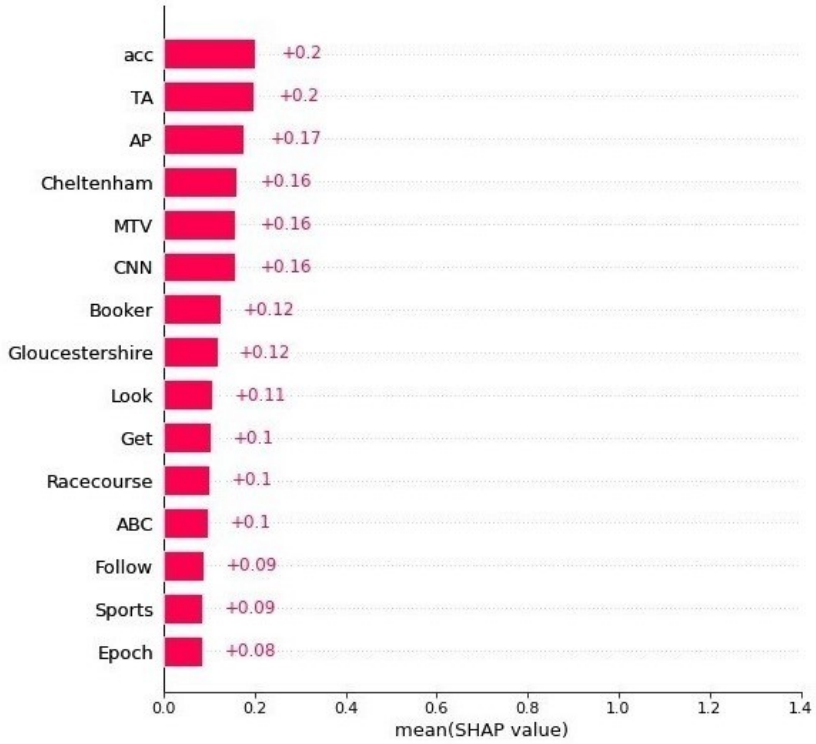


Figure 5.9: Most Unreliable terms according to the NELA-GT Dataset

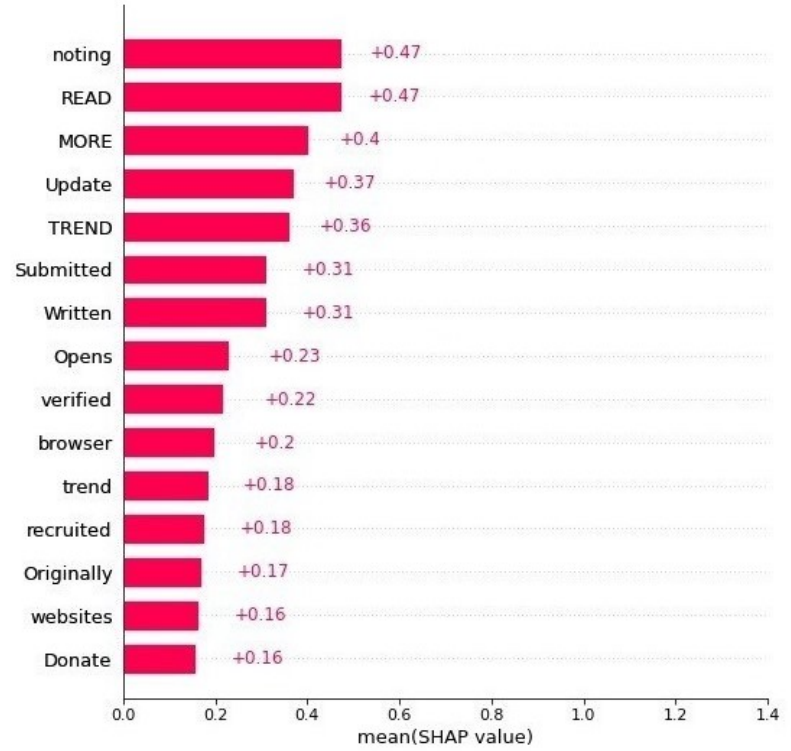


Figure 5.10: Most Mixed terms according to the NELA-GT Dataset

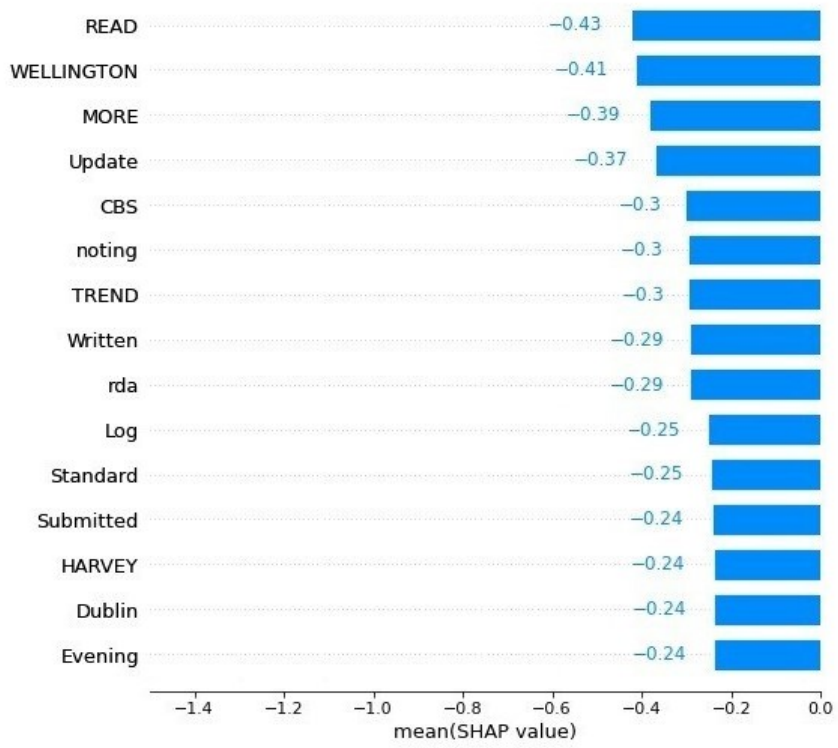


Figure 5.11: Most Reliable terms according to the NELA-GT Dataset

Regarding the representation of the data, it can be appreciated that it is mostly concentrated around the null value, in a similar way as the rest of the cases which are analyzed:

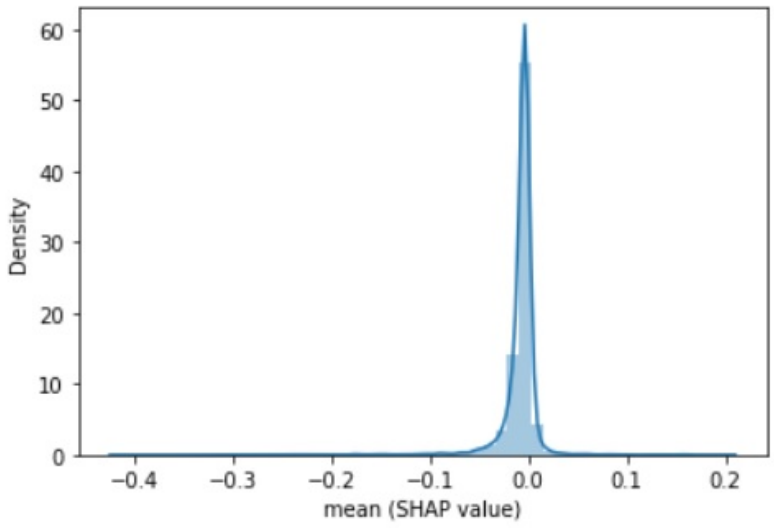


Figure 5.12: Representation of the distributed values obtained in the NELA-GT Dataset

5.2.3 Results of QProp

Extracting the values from this classifier has been the most problematic task out of the four. Even though a similar procedure has been implemented, the results show that the values that are assigned to the words are almost inappreciable, as they are extremely close to the origin. Nevertheless, as the aim when selecting the most representative values that the classifier has detected is to analyze this words, this procedure can still be developed.

Consequently, it can be spotted that a couple of terms which are linked to the religious topic (concretely to the jew) have been distinguished, *Judaism* and *Zionist*. Considering that the most possible motivation yo use this terms is to offend the followers of this religion, the use of this racism, aside from being hateful, can clearly be considered as propagandistic.

It is commonly known that in the USA there is a considerable amount of media which does support these ideas and this phenomena seems to be one of the main reasons for the necessity to start developing language classifiers as in this project.

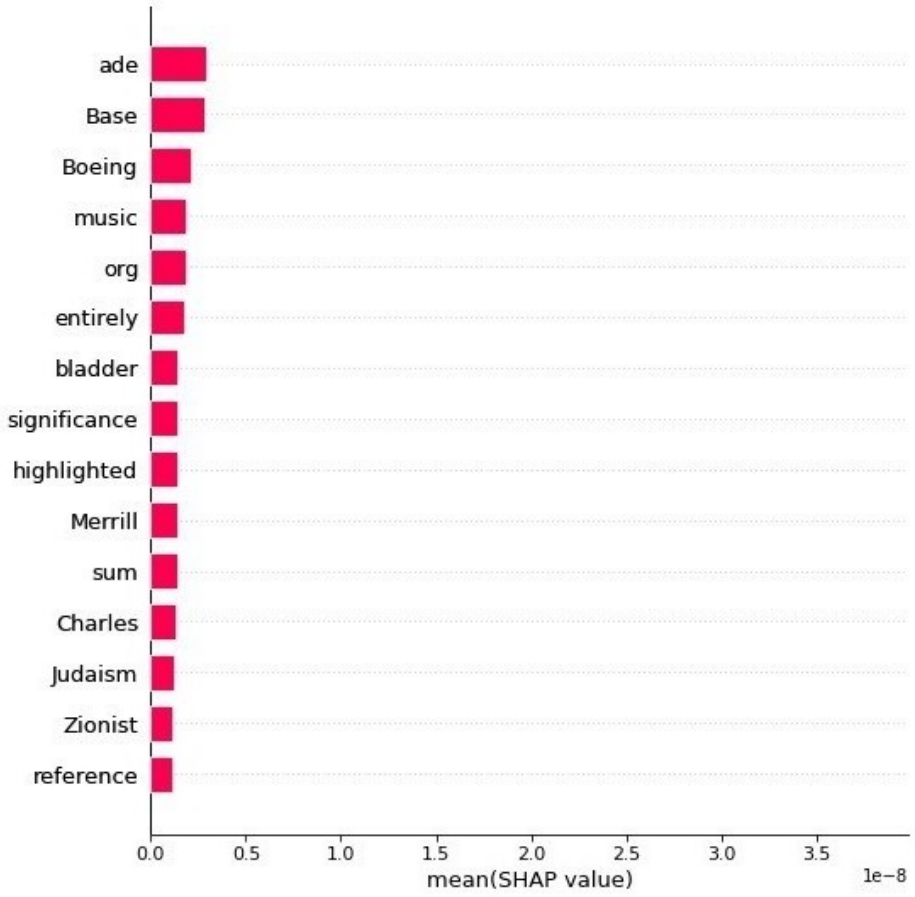


Figure 5.13: Most Propagandistic terms according to the QProp Dataset

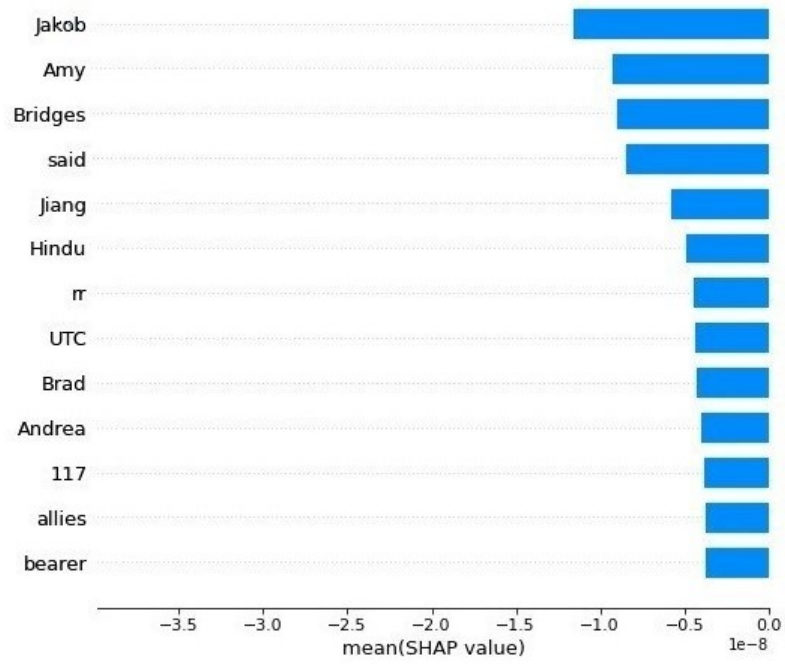


Figure 5.14: Least Propagandistic terms according to the QProp Dataset

In terms of the distribution of the data, even its values are small, it still follows the same pattern as the rest of the datasets, where the majority of the words are located in the positions close to 0. Moreover, it can be appreciated that there are more representative values in the neutral field. This phenomena is derived by the fact that this dataset has a considerably larger amount of samples labelled as Neutral, whereas there are very few propagandistic ones.

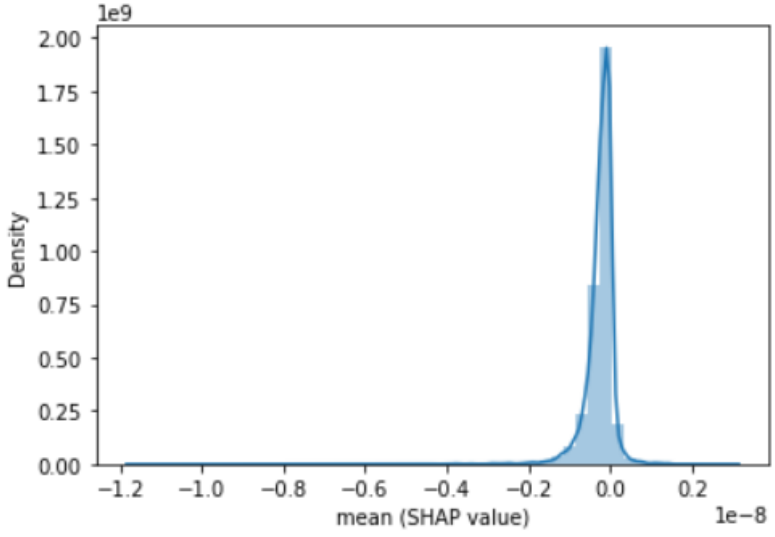


Figure 5.15: Representation of the distributed values obtained in the QProp Dataset

5.2.4 Results of Jacobs

Lastly, the results obtained by training a classifier with the Jacobs dataset are displayed. Considering that the data is related to religious radicalism, it is reasonable to obtain as Propaganda terms like *compassion*, *grace*, *Almighty*, *mercy* or *faithful*. This may happen due to the fact that the Muslim radicals use these terms in order to persuade new followers of their activity, which has to be softened so that it is more attractive to these possible new members.

On the other side, the terms which are detected as more neutral do not include religious topics. Even though some of them may be considered as violent, it must be reminded that these datasets is mainly composed by the texts that the jihadists post in the media in order to attract new followers to their cause, consequently this proceeding is the one which the classifier will detect as propagandistic.

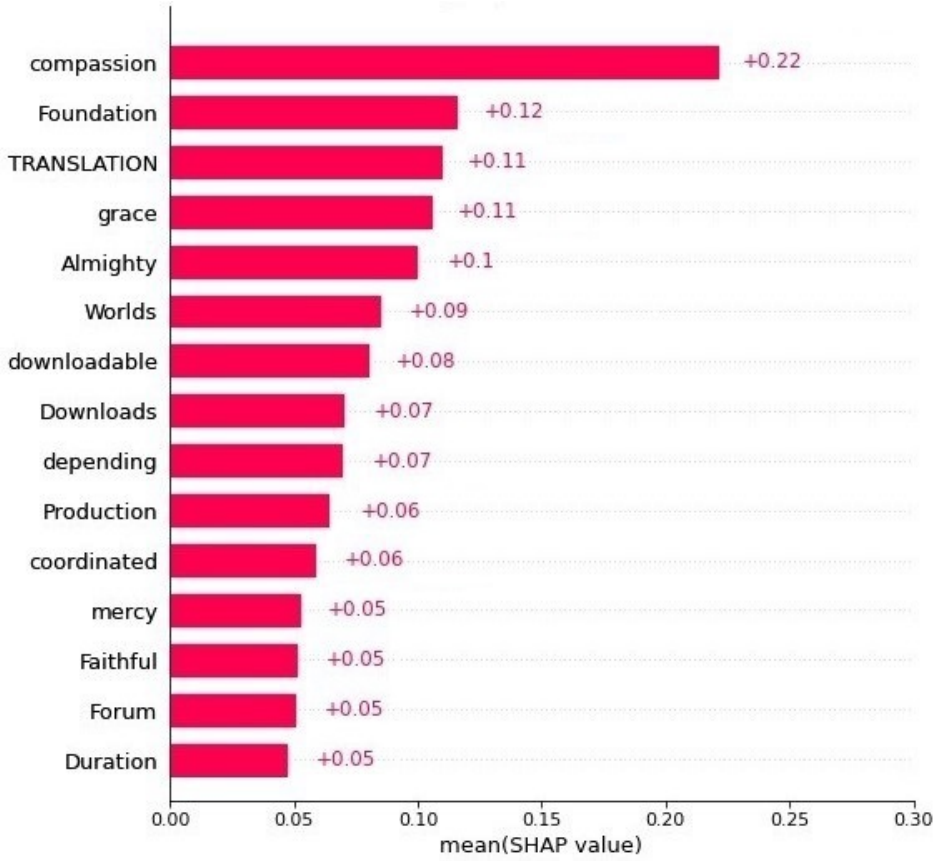


Figure 5.16: Most Propagandistic terms according to the Jacobs Dataset

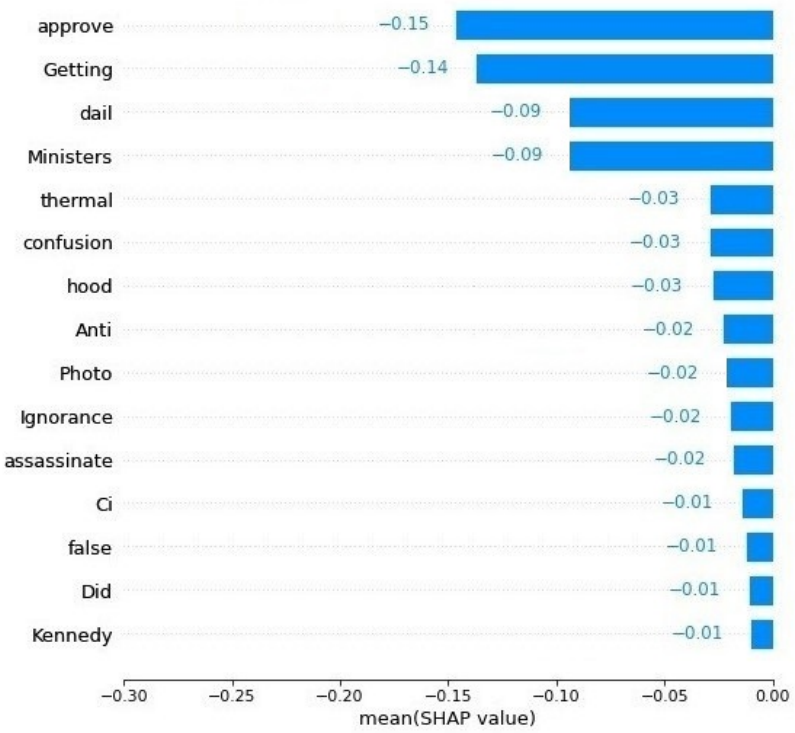


Figure 5.17: Least Propagandistic terms according to the Jacobs Dataset

Lastly, data is again distributed following a normal distribution, where the most of it is located in the center of the graph, sticking to a Gaussian with 0 mean and very low standard deviation, almost having the shape of a delta centered at the origin. However, this is not the case as in this plot the scale of the abscissa axis is wider than, for instance, in Figure 5.15. Therefore, selecting a wider scale changes the visual shape of the plot, although it still follows a Gaussian shape.

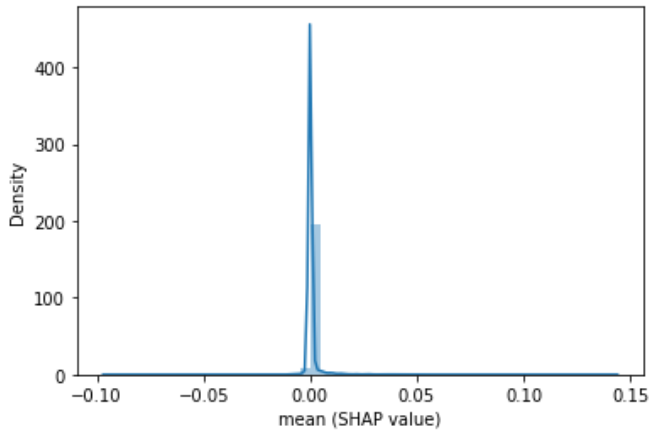


Figure 5.18: Representation of the distributed values obtained in the Jacobs Dataset

CHAPTER 6

Conclusions and future work

In this chapter the outcome that is extracted from this project will be described, discussing whether the initial goals have been achieved. Furthermore, the possible issues which have occurred have been faced and, finally, a possible future development of this project will be considered.

6.1 Conclusions

Since the detection of Propaganda is not objective, its implementation in a Machine Learning based algorithm can be a challenging task. This may be an argument to be used by the detractors of this technique, but what can not be omitted is the growth in recent years of several political tendencies which could have a negative impact for the society. It may be considered that the freedom of speech covers every single topic, but there must be some limitations to protect, at least, the basic human rights.

Moreover, in wartime the expansion of propagandistic news is even more considerable, therefore the detection of this content becomes additionally relevant. What makes this project attractive for the writer is that that it is not isolated, as its results will be applied to enhance the European Participation program, where it belongs. Consequently, as the

attacking side in the previously mentioned war is known to handle loads of propaganda, it is a relief to expect that the results which are obtained in this project could be used in a tremendously important task, to preserve peace in our continent.

Setting the focus in this project, the results are clearly satisfactory. In terms of metrics, all the classifiers have performed in the expected way, which confirms the capability of the Transformer to be used in the future not only in terms of NLP, but in all the fields of Artificial Intelligence.

In order to show the weaknesses as well, it has to be noticed that the integration of SHAP with Transformers was not as gratifying. SHAP values were extremely time-consuming to be extracted, therefore the implementation of this tool must be considered in the future when training an algorithm with Transformers, as the process of obtaining these values can be significantly faster.

All in all, executing this project has allowed myself to introduce in the world of Artificial Intelligence which, considering that it is not dominated by a great number of professionals, it offers a great perspective with lots of opportunities. Moreover, to master such a powerful tool as Transformers will be a considerable advantage for my future. Therefore, I am extremely grateful to my tutor for realising that this topic was the best option for this project.

6.2 Achieved goals

In Section 1.2, a series of objectives were presented. In this fragment it can be seen whether they were completed satisfactorily:

- Objective 1 has been completed, since the model was not only applied but also its architecture was thoroughly studied. Consequently, it has been demonstrated that Transformers is a significantly powerful tool with a considerable number of different applications.
- Objective 2 has been accomplished as the metrics which are obtained are satisfactory in all four algorithms. Therefore, it can be concluded that the implementation of Transformers in NLP classifiers is adequate.
- Objective 3 has been achieved, considering that by means of SHAP the amount of propaganda which the model assigns to each word has been extracted and analyzed. Moreover, the representations that are obtained have been studied not only from a

general view but also from a more particular perspective, both producing a reasonable outcome.

- Objective 4 has been fulfilled, since four intuitive dictionaries have been extracted to be used by other researchers. Furthermore, these documents can be joined in the future so as to produce a more powerful classifier.

Consequently, it can be concluded that the four aims were sufficiently realized.

6.3 Problems Faced

These are the main problems that have been faced during this project:

Lack of Knowledge

First of all, building four language classifiers by means of the most recent technology is not a straightforward task, considering the very little knowledge about Machine Learning, NLP or even the programming language Python which is given to us during the degree. Learning and adapting my mind to this field was a challenging task, specially at first. Nevertheless, GSI has a series of considerably powerful tutorials which helped in order to solve the aforementioned issue considerably soon.

Time Consumption

Secondly, it is known that training is significantly time-consuming, although when the whole process was understood and the steps which are explained in the Section 4.3 were followed, the task became significantly more accessible. However, passing the model through the module SHAP in order to visualize the values which are assigned for each word takes even more time, as SHAP is not such an optimized tool as the Transformer is.

Shorten Samples

In three of the four samples datasets the samples were too long for SHAP to detect every word on them, which resulted in an error. Therefore, the samples had to be trimmed in this datasets in order to perform this activity. As result, the values which are obtained are not extremely accurate, as the samples are trimmed by the number of characters, leaving sentences at the middle, which leads to a loss of meaning.

6.4 Future work

In this section it has been discussed the possible improvements that can be made to this project in order to get better results:

Data Division

Even though the results look satisfactory, as the metrics which are obtained as a result of using Transformers are one step ahead of other previous technologies, when facing the values assigned by SHAP the outcome is sightly different, as the most significant words which are obtained by the different classifiers are not always representative. Indeed, it is remarkable in the case of NELA-GT, where the classification of the data into three categories instead of just selecting Propaganda leads to indecipherable values where any possible conclusion is difficult to be drawn. Therefore, an aspect to be considered is adapting the original datasets so that the aim of the project, the detection of propaganda, is effortless for the classifier, by labelling the data into the to main categories which are analyzed: Propaganda and Neutral content.

Datasets Join

Another possible implementation for the future is to unify the datasets, so that a larger amount of samples is studied. Although the training of the whole set would take a considerable amount of time, theoretically the resultant classifier would have a promising performance. Moreover, considering that all four datasets have already been adapted to a similar format in order to be introduced in Transformers, this would not be such a challenging task.

Application in other fields

It seems reasonable to think that language classifiers are not only valid for detecting propagandistic content. Moreover, Transformers is able to operate not only with text but also video or audio, which results in a significant amount of opportunities in very different areas.

APPENDIX A

Impact of this project

This appendix reflects, quantitatively or qualitatively, on the possible impact, in terms of social, economic, environmental and ethical implications that can be extracted as a result of the fulfillment of this project.

A.1 Social impact

In this section it is going to be argued the most important impact of this project, the one referred to the society.

As it has been explained throughout the document, the detection of Propaganda is a significantly growing necessity for nowadays society. Radicalism, on its various forms and faces, is growing and starts to threat the political context, even in the European Union. This is reflected in the start of various project which are targeted in this issue, as the one in which this project is included, Participation.

Moreover, the presence of a war in our proximity makes Propaganda to be much more relevant, since to persuade with the best story of the happening events is even more important than the actual military conflict.

The growth of artificial intelligence is key in order to treat this situation, therefore the appearance of new technologies such as Transformers which are capable of enhancing the performance is considerably significant. Text classifiers can be used in the future not only by the media but also by governments or even by the actual society, which may want to just be informed by certain articles which have passed the filter of a classifier which may be similar to the ones which are developed in this project.

A.2 Economic impact

This technology can not only mean an improvement for the population in social terms, but also on the economic ones.

It seems clearly more optimum in terms of money and time for a certain means of communication to let a machine learning based classifier to select the news which are relevant and letting the employees just to analyze the results instead of making them to do the whole job.

Moreover, the use of Transformers means a significant reduction of the time which a traditional RNN based classifier would spend to train. Therefore, this means a reduction on the electricity bill which, especially in the current economical context, doe not have to be neglected.

A.3 Environmental impact

Training an artificial intelligence based classifier is not only expensive and time-consuming but it has also several environmental consequences.

A great amount of electricity needs to be produced, as these algorithms precise a significant amount of time to be trained, which is clearly negative for our planet, especially considering that a very little amount of this electricity comes from a renewable source. Therefore, the reduction of time as a result of the implementation of Transformers is positive in this area too.

Additionally, the execution of the entirety of the code which has been written to fulfill this project has been performed in a cluster instead of locally, which means a saving of resources that ultimately has a helpful consequence for the nature.

A.4 Ethical implications

Lastly, the automatic detection of Propaganda can derive on ethical issues.

It can clearly be appreciated this activity is not completely objective, as the definition of Propaganda is not a globally assumed fact. Nevertheless, it has to be considered the target of this project is just the implementation of the classifiers, not the division of the data which has already been done by some media experts when building the datasets.

Therefore, the only concerns of this project regarding this ethical aspect appear when analyzing the most relevant terms which have been selected by the classifier. When performing this task, I have intended to be as objective as possible, limiting myself to group this data onto a certain topic and extracting some relevant conclusions.

APPENDIX B

Economic budget

This appendix details an adequate budget to bring about the project, considering both physical and human resources, licenses and taxes.

B.1 Physical resources

As it has been explained previously, the entirety of the code is stored in a cluster which can be accessed from every laptop by means of a token. Therefore, even though in this section I will include the information regarding the hardware of my computer, as it is the one used to access the laptop, it has to be noticed that it can be accessed from almost any device.

Moreover, it has to be considered that the system where the code is executed needs a GPU, as this is the result of using the module PyTorch. Therefore, as in my laptop contains one, it will be included in the characteristics:

As a result, the hardware of my personal laptop includes the following elements:

- **CPU:** Intel(R) Core(TM) i7-8550U 1.80GHz
- **GPU:** NVIDIA GeForce MX150

- **RAM:** 16 GB
- **Storage:** 237 GB SSD, 1 TB HDD

All in all, the estimated cost of this hardware is around 900 €.

B.2 Human resources

This project have been done individually, therefore it has to be considered the salary of one single person.

Considering that the total number of hours which are required to finish a project where theoretically it comprehends a time equivalent to 12 ECTs, it can be concluded that 360 hours are required.

Assuming that this project is executed by means of a part-time shift, given by the investigation group GSI, which implicates a salary of 500 € per month, working for an average of 5 hours a day, this leads to an approximate salary of 2800 €.

B.3 Licenses

Both Transformers and all the modules which have been used are open source, therefore it can be concluded that any costs in this area have to be considered.

B.4 Taxes

The realization of the project does not derive in the payment of any tax, considering that the which corresponds to the acquisition of the physical resource is already included on its estimated budget. Nevertheless, different taxes must be considered in the case of selling the results to a third party, when the tasks of the country have to be included.

B.5 Total budget

As a result of the above-explained costs, an total budget of 3700€ can be estimated.

Bibliography

[1] Todd C Helmus, Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, and Zev Winkelman. *Russian social media influence: Understanding Russian propaganda in Eastern Europe*. Rand Corporation, 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] Participation. Extremism and radicalization: A new approach in research and preventive design, 2022. Accessed: 2 April 2022.

[4] Andrea Ceron and Vincenzo Memoli. Flames and debates: Do social media affect satisfaction with democracy? *Social indicators research*, 126(1):225–240, 2016.

[5] Stephen Cass. Top programming languages 2021: Python dominates as the de facto platform for new technologies, 2021. Accessed: 28 April 2022.

[6] Pandas. A fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the python programming language., 2022. Accessed: 01 June 2022.

[7] NumPy. Numerical calculus and data analysis, 2022. Accessed: 04 June 2022.

[8] Matplotlib. Visualization with Python, 2022. Accessed: 06 June 2022.

[9] Scikit-learn. Machine Learning in Python, 2022. Accessed: 07 June 2022.

[10] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246, 2017.

[11] 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.

[12] Denis Rothman. *Transformers for Natural Language Processing*. Packt, Birmingham, UK, 1st edition, 2021.

[13] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[14] Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR*, abs/2102.04567, 2021.

[15] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019.

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

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[18] Hugging Face Transformers. State-of-the-art machine learning for pytorch, tensorflow and jax., 2022. Accessed: 15 June 2022.

[19] SHAP. A game theoretic approach to explain the output of any machine learning model, 2022. Accessed: 08 June 2022.