UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

DEVELOPMENT AND EVALUATION OF A MENTAL HEALTH DETECTION SYSTEM ON SOCIAL MEDIA POSTS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

ANDREA LAGUNA LIANG

JUNIO 2023

TRABAJO DE FIN DE GRADO

Título:	Desarrollo y Evaluación de un Sistema de Detección para la Salud Mental utilizando técnicas de Procesamiento de Lenguage Natural y Aprendizaje Automático.
Título (inglés):	Development and Evaluation of a Mental Health Detection System on Social Media Posts using Natural Language Processing and Machine Learning Techniques
Autor:	Andrea Laguna Liang
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 MENTAL HEALTH DETECTION SYSTEM ON SOCIAL MEDIA POSTS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

Andrea Laguna Liang

Junio 2023

Resumen

La depresión es un trastorno mental frecuente que afecta a todos los ámbitos de la vida y puede conducir al suicidio. Sin embargo, existen tratamientos, y los programas de prevención son eficaces [67]. El objetivo de este trabajo es diseñar e implementar un sistema automático para la detección de la depresión en las redes sociales. Para ello, se utilizaron técnicas de aprendizaje máquina y procesamiento de lenguaje natural. Este trabajo aborda dos objetivos principales. El primero es determinar si la contextualización a través de las emociones y los tópicos puede ser utilizada para la detección de la depresión, mientras que el segundo objetivo es perfilar alternativas de modelos a través de un compromiso entre la calidad de la clasificación y la eficiencia energética.

Tras la exploración, se alcanzaron ambos objetivos. En relación con el primer objetivo, este trabajo muestra que la contextualización a través de la información de emociones y tópicos fue informativa para la detección de la depresión, a pesar de disminuir la F-score en algunos casos. Los resultados fueron particularmente informativos para el dataset DepSign, donde se encontraron ejemplos concretos de comportamientos que indicaban depresión. Ejemplos de esta información incluyen la indicación de depresión grave si se detecta un tópico que trate sobre medicamentos, o si están presentes ciertas palabras como *depresión, suicidio* y *deprimido/a* . Así, se demostró que se podían extraer conocimientos relevantes a partir de algoritmos de aprendizaje máquina y procesamiento de lenguaje natural.

Además, se identificó un candidato claro para el compromiso entre costes computacionales y calidad de la clasificación. Así, se demostró que en algunas aplicaciones pueden no ser necesarias las técnicas más avanzadas, sino que pueden utilizarse algunas técnicas preexistentes con un mejor equilibrio entre coste computacional y rendimiento. El ejemplo concreto encontrado en este trabajo fue también en el conjunto de datos DepSign, donde el uso de SIMON, un algoritmo basado en los word embeddings, en lugar de un modelo Transformer, redujo la F-score en sólo un 2%, mientras que el coste fue más de cien veces menor. Por último, la interpretabilidad resultó ser un componente clave para el análisis de este trabajo, sobre todo por tratarse su ámbito de aplicación de un campo médico.

Palabras clave: Aprendizaje automático, Procesamiento de Lenguaje Natural, de-

presión, BERTopic, word embeddings, Transformers, detección de emociones, redes sociales

Abstract

Depression is a common mental disorder that affects all areas of life, and can lead to suicide. However, treatments exist and prevention programs are effective [67]. The aim of this work is to design and implement an automatic system for the detection of depression on social media. To do so, Machine Learning (ML) and Natural Language Processing (NLP) techniques were used. There were two main objectives. The first one was to further determine whether contextualization through emotions and topics could be used for depression detection, while the second objective was to profile model alternatives through a trade-off between performance and energy efficiency.

After the exploration, both objectives were achieved. Related to the first objective, this work shows that contextualization through emotion and topic information was informative for the detection of depression, despite lowering the F-score in some cases. The results were particularly informative for the DepSign dataset, where concrete examples of behaviours indicating depression were found. Examples of this information included the indication of severe depression if the topic of medications was detected, or if certain words such as *depression, suicide* and *depressed* were present. Thus, it was shown that relevant knowledge could be extracted from ML and NLP algorithms.

Furthermore, a clear candidate for the trade-off between computational costs and performance was identified. Thus, it was shown that in some applications state-of-the-art techniques may not be needed, but rather that some pre-existing techniques with a better balance between computational cost and performance can be used. The concrete example found in this work was also in the DepSign dataset, where using SIMON, an algorithm based on word embeddings, rather than a Transformer model, reduced the F-score in only 2%, while the cost was more than a hundred times lower. Finally, interpretability was found to be a key component for the analysis of this work, specially given its concern with a medical field.

Keywords: Machine Learning, Natural Language Processing, Depression, BERTopic, word embeddings, Transformer models, Emotion Detection, social media

Agradecimientos

Es realmente complicado escribir agradecimientos, ya que difícilmente expresa esta palabra todo lo que han hecho otros por uno mismo. Detrás de todo el esfuerzo propio, se encuentran muchas personas, sin las cuales, no habría sido fructífero.

En primer lugar, me gustaría darle las gracias a Óscar, por esta gran oportunidad y su inestimable apoyo. Realmente no se podría pedir un mejor mentor. Gracias, también, a todos los compañeros del GSI, por hacer del departamento un lugar tan acogedor.

A mis amigos, gracias por todos los recuerdos y las risas que me llevo de esta etapa. Alisson, Alena, Tomás, Paula, Rafa y Giannina, gracias por hacer de la escuela un lugar especial. A Paloma, Lucía y Snorri, por estar siempre. Y a los Rhythm Peanuts, mis cacahuetes favoritos.

Gracias también a mis padres, Chi y Héctor, por todo. Por apoyarme cuando necesitaba apoyo, y por creer en mí cuando dudaba. Por guardarme las espaldas para que pueda mirar al futuro, gracias.

Por último, le dedico este trabajo a mi abuela Maribel, quien siempre ha sido un modelo a seguir para mí. Me habría encantado tenerte a mi lado en mi graduación. Allá donde estés, abuela, espero que estés orgullosa.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

Depression is a common mental disorder, which affects 3.8% of the global population, and has different prevalence is different demographic groups. It is more prevalent in adults older than 60 years, and more prevalent in women than men. This disease can affect all areas of life, and in severe cases, lead to suicide. However, there are effective treatments for depression, and prevention programs against depression are effective [67]. This disease is underdiagnosed and undertreated, but early detection and treatment can improve its outcome [28].

Where mental health used to be taboo, now lies a much greater acceptance of the public discussion of mental disorders, in no small part thanks to the rise of social media. This has allowed for a boom in discourse on said platforms about the topic, and a great abundance of new data to be generated daily. This work aims to utilize this data to generate knowledge that could help improve the treatment people suffering from depression receive.

Artificial Intelligence is thus an invaluable tool for this types of applications, but as everything, it comes with a cost. And, in the case of AI, the cost is to be paid by the environment. Because of the vast amount of energy intense calculations that must be made

to train these algorithms, training a Transformer model (which will be explained later on) was estimated to generate up to 626 K pounds of CO2 [59], which is approximately 284 tonnes. Then, the cost of usage must be considered, as well as the great amount of storage space is needed for these data-hungry algorithms, which carries its own cost in terms of physical resources.

1.2 Project goals

The objective of this work is to use Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyse language associated with depression, generating automatic systems capable of classifying texts as either depressive or not. This is to be done on text extracted from social media platforms. The aim is to include different context information, and analyze what knowledge can be extracted from it. This is to be done in both Spanish and in English. As previously mentioned, the information extracted from the experiments run could later be used by automatic systems to perform screening on social media, thus extracting demographic information about the prevalence of depression. In addition, the costs in terms of energy and time will be considered and analyzed. The objectives can be listed as follows:

1. Collect and study datasets pertaining depression in both Spanish and in English, as well as the appropriate resources for their analysis, which includes embedding models and emotion lexicons, among others.
2. Determine whether contextualization through emotion and topic information can be used for depression detection.
3. Attempt to profile model alternatives by means of a trade-off between detection performance and energy efficiency.

1.3 Structure of this document

In this section, the structure for this document will be detailed.

Chapter 1. In this chapter, the general context has been presented, as well as the problem to be tackled.

Chapter 2 will present the tools and technologies which were used in this implementation, as well as previous studies that have used similar techniques and tools for similar

applications to this work.

Chapter 3 will describe the specific models and how they will be used in this work.

Chapter 4. This section will describe the implementation of the models previously described, as well as the resources used for the analyses, the design of the experiments and their results.

Chapter 5. In this section, final conclusion of the work will be presented.

Chapter A and chapter B. Finally, in these sections, the impact of this project and the economic budget will be discussed.

CHAPTER 2

Enabling Technologies and Related Work

In this chapter, the different enabling technologies and related work for this project will be presented. For the enabling technologies, the programming language which was used, Python, will be presented, alongside with the main libraries used and the development environment. Then, the tools related to Natural Language Processing will be presented. Finally, the technologies related to Machine Learning will be described. The related work section will explore previous research concerning the detection of depression on social media, as well as the use of word embeddings, Transformers, topic modelling and emotion detection for detection of mental health conditions.

2.1 Enabling technologies

2.1.1 Python

The programming language used for the implementation of this project was Python. In the following section, the main libraries used will be described. Firstly, the library **Pandas [46]** is an open source library for real world data analysis in Python. It employs the DataFrame object for manipulation of data [3]. For this application, it has been extensively used for

the manipulation of datasets. **Matplotlib [30]**, on the other hand, is a library for creating visualizations in Python. In this work, this library was used for the generation of plots. It was used alongside the extension **Seaborn [65]**, a statistical data visualization library based on matplotlib, for the generation of graphs.

Finally, the environment used for this development was **Jupyter Notebook**, a web based development environment. In addition to allowing code execution, it provides rich media representation of the results. It also allows the use of markup language, making it possible for code to be annotated for easier interpretation [2].

2.1.2 Machine Learning

The ML algorithms used during the development of this project were implemented by the Python library Sci-kit learn [48]. **Sci-kit learn** is an open source toolbox for predictive data analysis, built on matplotlib, NumPy and SciPy.

The specific **classifiers** that were used in this work, as implemented by the aforementioned Sci-kit learn library, and chosen due to the nature of the datasets, are as follows, in no particular order. **Linear SVC**, or Linear Support Vector Classification, is an implementation of a Support Vector Machine. Support Vector Machines are a series of supervised learning methods which can be used for classification. These can be both binary and multiclass classification. Linear SVC is an implementation of a support vector machine with a linear kernel. In Sci-kit learn it is implemented with the method LinearSVC. The **polynomial SVC**, on the other hand, is also a support vector machine. In this case, it s implemented using the method SVC with the kernel= "rbf" [5].

The **KNeighbors Classifier** implements a neighbors-based classification, which means that rather than "learning", the algorithm performs a "instance-based learning", where instances of the training data are stored, and class is determined by the nearest neighbors of the data points [6]. The method KNeighborsClassifier was used for its implementation. Finally, the **Random Forest Classifier** was also used, implemented by the method RandomForestClassifier. Random forests generate several decision tree classifiers in several sub-samples of the dataset. Then, these results are averaged in order to improve classification and control over-fitting [7]. These classifiers are constructed by the introduction of randomness [4].

2.1.3 Natural Language Processing

Libraries for NLP were used for the development of this project. **NLTK** [16] is one of the main tools available for Python to work with human language. Published in 2009, it includes many functionalities such as tokenization, stemming and tagging. This platform also has many pre-trained models and corpora that can be accessed, including grammars for many languages, sentiment lexicons and corpora.

GSITK [11, 13] is a library built on top of scikit-learn for development of projects based on NLP and ML. Among others, it is also built upon pandas and numpy. Specifically, the feature extractor SIMON [13] was used in this work. A further description of this feature extractor can be found in Chapter 3.

2.2 Related work

2.2.1 Detection of depression on social media using NLP methods

A study by Park et al. [47] found that the language used in social media could provide valuable information of the mental life of the person who wrote it, finding that their method could be used to complement traditional methods. This is based on the notion that language contains relevant information about the psychological status and individual personality. In addition, De Choudhury et al. [23] found that social media contains useful information for the detection of depression. Also, De Choudhury et al. [22] showed that it was possible to discern mental health discourse on social media from discourse pertaining mental illness, specifically, containing suicide risk and ideation. Finally, Coppersmith et al. [20] found that simple natural language processing methods were capable of providing information about mental health and mental disorders. In particular, depression as well as bipolar disorder, post-traumatic stress disorder and seasonal affective disorder were analysed based on texts extracted from the social media platform Twitter. Thus, it can be seen that many efforts to use social media platforms as tools to detect depression and other mental illnesses have been made.

2.2.2 Vectorization

When processing natural language one of the main challenges is the representation of texts fed to learning models. Given the nature of such models, it is necessary to transform texts

into numerical representations. To do so, there are several different approaches to perform this mapping [36]. From the simple to the more complex, these vectorizers tackle the task in different ways. In this work, we explore the use of two approaches: word embeddings and Transformers.

2.2.2.1 Word embeddings

As mentioned, in order to be able to analyse texts with ML algorithms, these must be converted into vectors of numbers. Word embeddings, unlike previous techniques, allow the text to be represented as a continuous vector (as opposed to a sparse vector) [32].

Among other characteristics, these types of word vector representations are capable of capturing both semantic and syntactic regularities. These regularities are captured by the offsets between vectors, similar words tending to have similar vectors. So much so, that a particular relationship between words will have an associated offset. For example, the relationship between the vector representations of pairs of the plural and singular of a noun (apples/apple) is close to that of an unrelated noun (cars/car). Thus, algebraic operations performed on the word vectors reveal meaning. For example, the operation of subtracting the vector representation for "Man" to "King", and adding the vector for "Woman", will yield a vector most closely related to the vector for "Queen" [39]. There are many implementations of word embedding models [38].

Word embedding models have been previously been applied to the detection of mental illness, specifically anorexia and depression, by Trotzek et al. [61]. Similarly to the approach followed in this work, pre-trained word embedding models were also used in several implementations which have word embeddings as an input. As such, GloVe Embeddings were used, as well as a custom trained fastText Embeddings. Similarly, Pérez et al. [52] also used word embeddings to tackle the problem of depression on texts extracted from the social media Reddit. In their paper, however, rather than a classification, the severity of depression symptoms was estimated.

2.2.2.2 Transformers and topic modelling

The Transformer architecture [62], uses a encoder-decoder network architecture based only on layers with multi-headed self-attention mechanisms. This architecture originally showed to be superior to others for translation tasks, and it was also proven that it can be successfully generalized to others tasks. These Transformer models are trained on large datasets (generating the so called pre-trained models), and can be later fine-tuned for specific tasks,

resulting in improved performance in many tasks, including language modelling and sentiment analysis. This attention mechanism allows for a greater contextual understanding, and thus allowing for better predictions [58]. The aim of attention mechanisms is to find long-term dependencies in phrases [26].

Suri et al. [60] aimed to detect depressive tendencies on the social media platform Twitter. To do so, they used a Transformer model to process the textual features of their texts. This information was then combined with information regarding the online behaviour and interaction of users. This architecture improved the results of their baseline by a 12%.

For the extraction of topics in this work, the model **BERTopic** was used [27]. This model was proposed to uncover themes and narratives in texts in an unsupervised manner. It uses a pre-trained transformer-based language model to generate word embeddings, for their capabilities to represent texts in the vector space in a way that makes it possible to obtain their semantic similarity. After the documents have been transformed, the dimensionality of the embeddings is reduced to optimize the clustering that is done afterwards. Finally, using a custom class-based variation of TF-IDF, topic representations are extracted from the clusters of documents. In this work, this model will be used to extract contextual topic information from the texts. This model has been previously applied on the area of mental health by Baird et al. [15]. In their paper, BERTopic was applied to texts extracted from Twitter about telehealth for either mental health or substance abuse.

In addition, BERTopic has also been applied by Alhaj et al. [8] to detect cognitive distortions on the Arabic content of Twitter. Their implementation used the previously mentioned word embeddings (implemented by word2vec), together with the topic distribution generated by BERTopic, to generate a contextual topic embedding (CTE). This CTE aimed to both keep the semantic information and the contextual topic representation by concatenating the vectors produced by both of them. The classification performance of the CTE was bench-marked against just a word2vec embedding, and was found to improve the results.

Another work, by Sarkar et al. [56] aimed to predict depression and anxiety on the social media platform Reddit with a multi-task learning approach. A combination of word embedding features (pre-trained BERT model) and topic modelling features (LDA and BERTopic) were used. Their work concluded that this combination of features can be leveraged for domain specific tasks.

2.2.3 Emotion detection and emotion lexicons

Emotion Detection or Sentiment Analysis is a field of study that covers the emotion detection and recognition from text. There are two types of such analyses. If positive, negative, or neutral feelings are studied, this consist of Sentiment analysis, which aims to determine polarity. Emotion analysis, on the other hand, can detect types of feelings such as happiness, sadness, etc. Emotion models may be dimensional (that is, representing emotion based on its valence, arousal or power) or categorical (where emotions are discrete, such as anger or happiness) [29, 43].

Word-affect association lexicons, also known as emotion lexicons, are a compilation of words associated with the affect that they convey (which includes emotions, sentiments, etc.). These emotion lexicons may be generated through manual annotation or automatically. Emotion lexicons have many applications, among which the study of health disorders is most relevant for the case at hand. Lexicon-based emotion analyses have important advantages, such as being interpretable and having a low carbon footprint, which makes it a popular technique for real-world applications [41]. Previously, Li et al. [35] have generated and used a domain specific emotion lexicon for the detection of depression, obtaining better results than with general purpose emotion lexicons.

Another study by De Choudhury et al. [21] used texts extracted from the social media platform Facebook in order to implement prediction systems capable of detecting postpartum depression based on the user emotional and linguistic expression, as well as their activity and interactions. Contrary to the implementation of this work, though, the emotions considered were only "positive affect" and "negative affect", which would be considered as sentiment analysis. Their paper found that mothers from the cohort suffering from postpartum depression experienced higher levels of negative affect, and lower levels of positive affect, compared with their non-depressed counterparts, although in a less statistically significant way than other previous work on the social media platform Twitter. The authors proposed that the nature of the social media platform chosen could have affected the emotional expression of users. One of such works on the social media platform Twitter was done by De Choudhury et al. [23], where the detection of depression was studied, and again, positive and negative affect were considered, as well as activation and dominance as determined by the ANEW lexicon.

CHAPTER 3

Model description

In the following paragraphs, an overview of the models used in this implementation will be given. The first model that will be described is a feature extractor named **SIMON** [12], which uses embedding-based textual representations to generate a fixed length vector, proposing an improvement from regular word-embedding applications. This improvement consists of the following: by using a domain specific lexicon, the model then represents the words from the input text as a projection to said domain specific lexicon. This projection is done through the semantic similarity of words as given by the word embedding model [12].

In the original application of this model, as described by Araque et al. [12], an opinion lexicon was used to generate the domain specific lexicon. However, in this application, the domain specific lexicon was generated using the frequency with which words appeared in the texts used for training the model, which corresponded to the training split of the datasets, as was done in the method *FreqSelect* [9]. In this way, the information encoded in the generated representations contain implicit signals with regards to the objective classes. As such, the SIMON method extracts distributed representations exploiting both a word embedding model and a domain lexicon. In this application, two embedding models were explored (that is, one per language), and will be further described in the following sections.

Transformer Models [66] were also used to obtain the vector representation for these

texts. For the DepSign dataset, the multilingual xlm-roberta-base [19] was used, whilst the Spanish PlanTL-GOB-ES/roberta-large-bne model [25] was used for the Spanish dataset. These transformer models were used without fine tuning, in order to generate a vector which could be used as characteristics for a classifier.

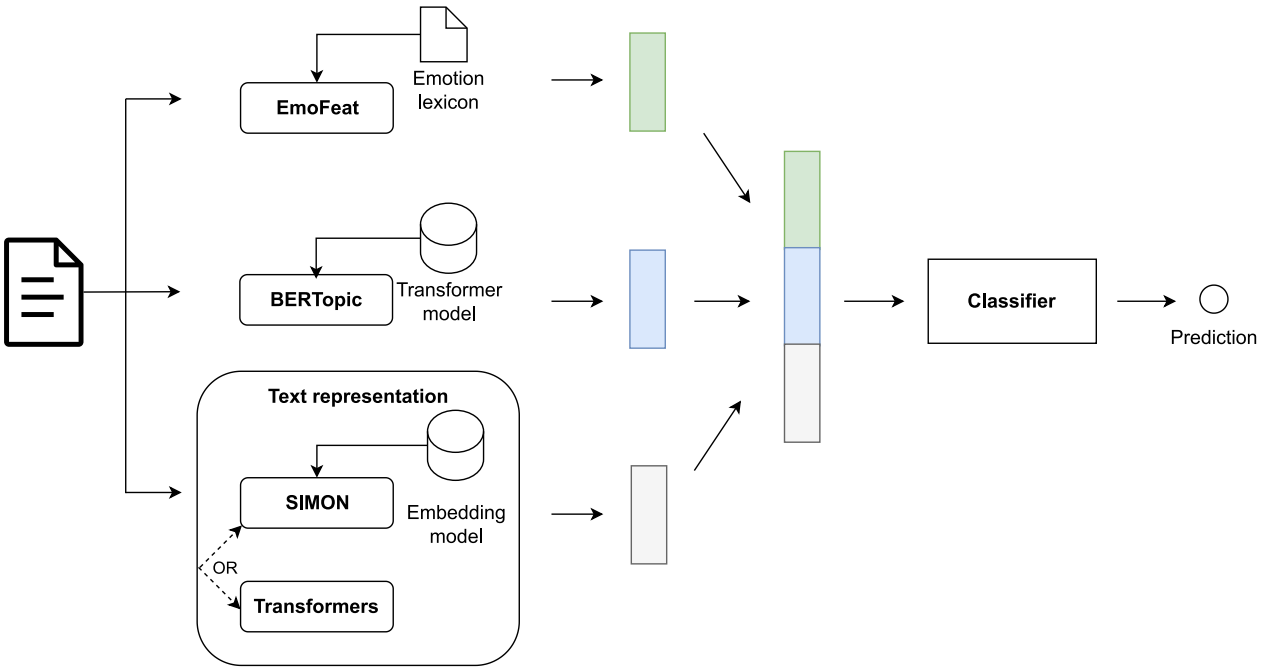


Figure 3.1: General architecture used in this work.

Information pertaining emotions was extracted using **EmoFeat** [9]. EmoFeat used an emotion lexicon for each language, to perform a statistical summary of the annotated values for the words contained in a text, thus reducing the matrix generated into a vector. In this particular application, both the maximum and mean emotion were calculated for each emotion, for every text. As such, the vectors generated for the texts in English had length 16 (as they were 8 annotated emotions), while for the Spanish texts the final length was of 12. The emotion lexicons used are further described in the next section.

In order to extract topic information to include as additional characteristics for classification, **BERTopic** was used. This is a topic modelling and visualization library based on Hugging Face Transformers and c-TF-IDF [27]. It was used to extract topics from both datasets. In the case of the DepSign dataset, originally 291 topics were extracted by BERTopic. The number of topics was then manually reduced to 128 and 64 from the original topic model, and all three of these cases were then considered for evaluation, to see how the size of the topic information vector could affect the classification. In the case of the Spanish dataset, only 21 topics were detected, so no reduction was carried out. Fine tuning of the model was carried out on the training split of the datasets, and the refined model was

used to generate the additional topic characteristic information for all texts in the datasets. For the vectorization, unigrams, bigrams and trigrams were considered. The graph 3.1 is meant to show a simplified structure of the implementation, featuring the models that were just described.

CHAPTER 4

Evaluation

4.1 Introduction

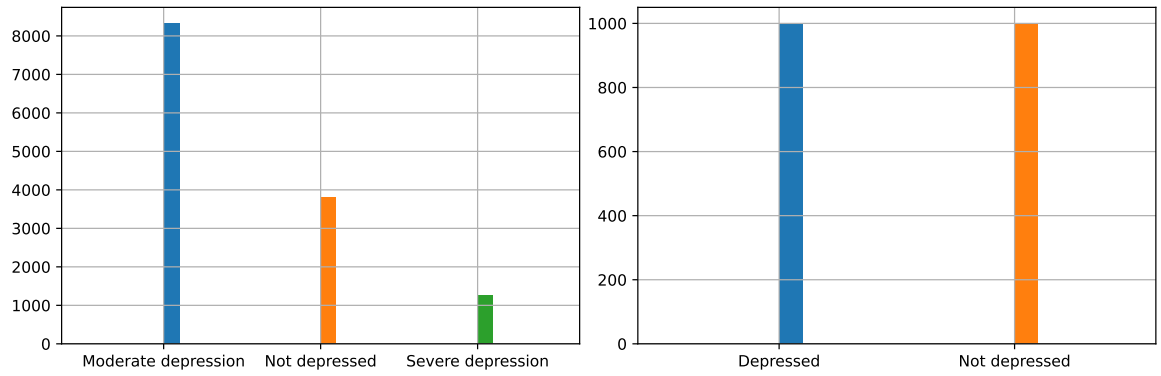
In this chapter, the datasets and resources used in this work will be described. As such, dataset distribution, number of classes and size will be discussed, as well as how they were generated. The embedding models chosen will be listed, and the emotion lexicons, analyzed. Then, the design of the experiments performed in this work will be described. Finally, the results for both of the datasets considered will be analyzed in terms of F-scores, computational costs and individual feature importance.

4.2 Dataset and resources

Two datasets consisting of texts from social media platforms were used for this work. The first dataset is an English dataset made available for the competition "Detecting Signs of Depression from Social Media Text-LT-EDI@ACL 2022" [24]. The dataset was created by Kayalvizhi and Thenmozhi [33], S et al. [55]. In this dataset, the texts were extracted from thematically relevant subreddits (forums) from the social media platform Reddit, and manually annotated by two domain experts with the following labels: "not depression",

"moderate depression" and "severe depression". These classes are unbalanced (4.1a), and the total number of texts used was of 13,387. From now on this dataset will be referred to as **"DepSign"**.

The second dataset is a Spanish dataset comprised in its entirety of manually selected depressive texts from the social media platform Twitter, provided by Leis et al. [34]. In order to have both examples of depressive and non-depressive texts, the Spanish dataset was mixed with random tweets that are mostly in Spanish retrieved from Pérez et al. [51]. This dataset was made to be balanced (4.1b), and the total amount of texts was of 2,000, and, similarly, will be referred to as **"Spanish dataset"** from now on. A summary of the dataset characteristics just mentioned can be seen on Table 4.1.



(a) Class distribution of the DepSign dataset (b) Class distribution of the Spanish dataset

Figure 4.1: Class distribution of the datasets

Table 4.1: Characteristics of the datasets

	DepSign dataset	Spanish dataset
Number of classes	3	2
Number of texts	13387	2000
Class distribution	Unbalanced	Balanced

For the application of SIMON to the DepSign dataset, the original word2vec [38] embedding model was used, while the one used for the Spanish dataset was the GloVe embeddings from SBWC [17, 50].

In addition to datasets, emotion lexicons were used, both in Spanish and in English. The Spanish lexicon was retrieved from Rangel et al. [53], Sidorov et al. [57] and has a length of 1909 words, while the English lexicon was obtained from Mohammad [40], Mohammad and Kiritchenko [42] and has a length of 16861 words. The English emotion lexicon covered 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), while the Spanish emotion lexicon covered 6 (anger, disgust, fear, joy, sadness, surprise).

4.3 Design

For both datasets, a test-train split approach was followed, with 33% of the dataset being used for testing. The metrics on which the evaluation is performed is the macro-averaged F1-score, although other metrics such as the classification report were also calculated, and will be used for a finer analysis of the results. Different feature combinations were considered, and their results compared, for both datasets. The feature combinations used can be found in Table 4.2. The objective of assessing these combinations was to consider the effect these features could bear on the classification metrics, as well as other factors, such as the computational costs, which will be described in the results section.

This further begs the question as to why these feature combinations were considered. Firstly, both SIMON and Transformers, however differently, extract the semantic information from the texts. In addition to this, due to the nature of depression, it made sense to use information on the emotions showed in theses texts. Finally, the topics of these texts were considered, as this information, which is prone to being interpreted by humans, could be useful in detecting certain topics of conversation which correlate with depressive symptoms, and, as such, could serve as an indication of the presence of the disease, even for a human observer. The reader will find further information of these models in the previous chapter 3.

Table 4.2: Combination of features used on both of the datasets.

Feature extractor	Emotions	Topics
SIMON	No	No
SIMON	No	Yes
SIMON	Yes	No
SIMON	Yes	Yes
Transformers	No	No
Transformers	No	Yes
Transformers	Yes	No
Transformers	Yes	Yes

Table 4.3

4.4 Results

As previously mentioned, to analyse the results obtained from this set of experiments, several things were considered. Firstly, the value of the macro-averaged F-score will be compared for both datasets, and all the combination of features. Then, computational costs will be considered and weighed in conjunction with F-score values. Finally, the relevance of individual features will be analysed.

4.4.1 Results of the DepSign dataset

As can be seen in Table 4.4, which presents all the results of the experiments performed on the DepSign dataset, the number of topics used for the classification has an effect on its outcome as measured by the **macro-averaged F-score**. In order to simplify the interpretation of these results, the following Figure 4.2 shows the effect on such metrics of the number of topics for different combination of features, for all the classifiers used. The vertical axis shows the macro-averaged F-score, while the horizontal axis shows the number of topics used for that specific classification. Each of the subgraphs shows a different scenario in regards to the features used, as specified in their respective titles. It can be observed that

			Dataset	DepSign			
			Classifier	Forest of randomized trees	KNeighbors Classifier	Linear SVC	Polynomial SVC
Feature extraction	Emotions	Topics	Number of topics				
Unigram baseline	NO	NO	0	69.39	68.58	51.07	53.80
SIMON	NO	NO	0	70.68	68.11	51.33	59.86
		YES	64	64.18	**68.73**	53.09	62.04
			128	62.65	68.61	55.68	62.67
			All(291)	61.18	68.64	56.91	62.64
	YES	NO	0	71.11	68.34	52.16	64.48
		YES	64	65.89	67.98	54.64	65.51
			128	64.69	67.66	55.53	65.42
			All(291)	62.33	67.81	57.94	65.63
Transformers	NO	NO	0	73.10	67.96	62.50	68.44
		YES	64	70.64	67.88	64.92	34.24
			128	69.16	67.87	58.87	34.24
			All(291)	68.68	67.81	59.94	34.24
	YES	NO	0	**73.35**	67.69	60.47	**68.48**
		YES	64	70.34	67.61	64.42	34.22
			128	69.51	67.68	**65.73**	34.22
			All(291)	69.04	67.62	52.53	34.22

Table 4.4

adding topic information can both improve or worsen the classification metric, depending on the classifier used, or even the number of topics considered.

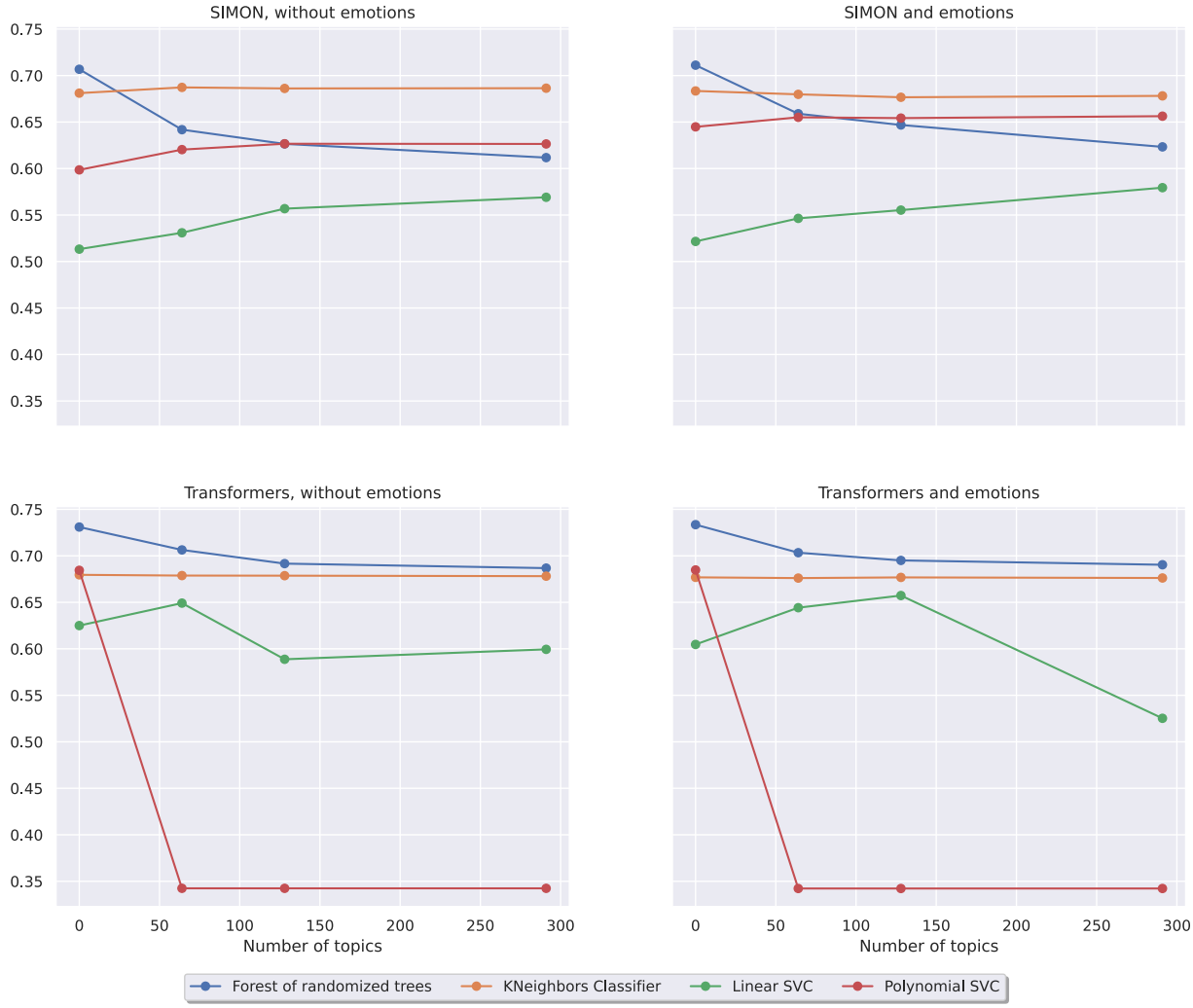


Figure 4.2: Effect of the number of topics used on the macro-averaged F1-score of the DepSign dataset.

In regards to the best outcomes according to the macro-averaged F-score, the Table 4.4 shows that the best results are obtained by the use of Transformers, together with information regarding emotions, but without topics, and the usage of the forest of randomized trees classifier, with a value of **73.35%**. However, it is worth noting that using this same classifier, the feature extractor SIMON, together with emotions, comes quite close to this results with a metric of **71.11%** F-score. It must be noted that in the paper where the DepSign dataset was created, a baseline analysis was made. In this analysis, the highest

metric was of a weighted average F1-score of 0.647 [33, 55]. Since the classification report for all the experiments in this work had been calculated, the weighted average F1-score for the best classifier was retrieved, with a value of 0.79.

Using the library CodeCarbon [18], it was possible to estimate the **computational costs** of using different features, both in terms of time and energy usage. The normalized results (that is, per text processed) can be found in Table 4.5. The Transformer, SIMON and Emotion features process either the train or test set as-is. The "all topics" feature considers the fit-transform of the train set, and the transform of the test set. The "reduced topics" consider the reduction of the number of topics of the train set, and the transform of the test set, as they are calculated based on the model with all the topics. As seen, the usage of Transformers over SIMON has a higher cost both in terms of time and energy of two orders of magnitude. Since the potential applications of the detection of depression on social media could imply the analysis of large amounts of texts, it is worth considering if this considerable increase of costs are justifiable with a little over 2% of improvement in F-score. In addition, the size of the vectors are significantly different. For the DepSign dataset, Transformer models generate vectors of size 768, while SIMON does of size 49. When the possible applications of this implementation are considered, it must be noted that this difference is size could mean a significant difference in memory needed.

Model	Normalized training time (s)	Normalized prediction time (s)	Normalized training energy (kW)	Normalized prediction energy (kW)
Transformers	5.75E-02	5.83E-02	2.28E-06	2.32E-06
SIMON	3.98E-04	3.98E-04	1.49E-08	1.48E-08
Emotions	**1.27E-04**	**1.37E-04**	**3.91E-09**	**4.21E-09**
All topics	1.50E-02	1.59E-02	3.35E-07	3.40E-07
128 topics	1.89E-03	1.09E-02	4.05E-08	2.35E-07
64 topics	1.61E-03	1.13E-02	3.46E-08	2.40E-07

Table 4.5: Computational costs in terms of energy and time consumption for the DepSign dataset. These costs have been normalized, meaning that the values displayed on the table are the time consumed or energy used per text, for the different models.

In order to better understand the information displayed in table 4.5, two graphs were plotted. Figure 4.3 shows the normalized time plotted against the normalized consumed energy, in milliseconds and milliwatts respectively. As can be seen in the graph, the Trans-

former model has a much higher cost both in terms of energy and time as the other models. The extraction of topics consumes roughly five times less time and six times less energy than the use of the Transformer models, and the consumption of time and energy of the extraction of emotions and the SIMON model is of a much smaller dimension, thus making the representation overlap them near the zero. The tags "Topics(reduce to N topics)" represent the process of taking the trained topic model, and reducing the number of extracted topics to either 128 or 64 topics. While it takes a small amount of time and energy to achieve this, it can be observed from the graph that it reduces the costs of calculating the topics of texts, without significant difference between 128 and 64 topics. Thus, depending on the application, the reduction of the number of topics considered may be a way of reducing the costs of the process.

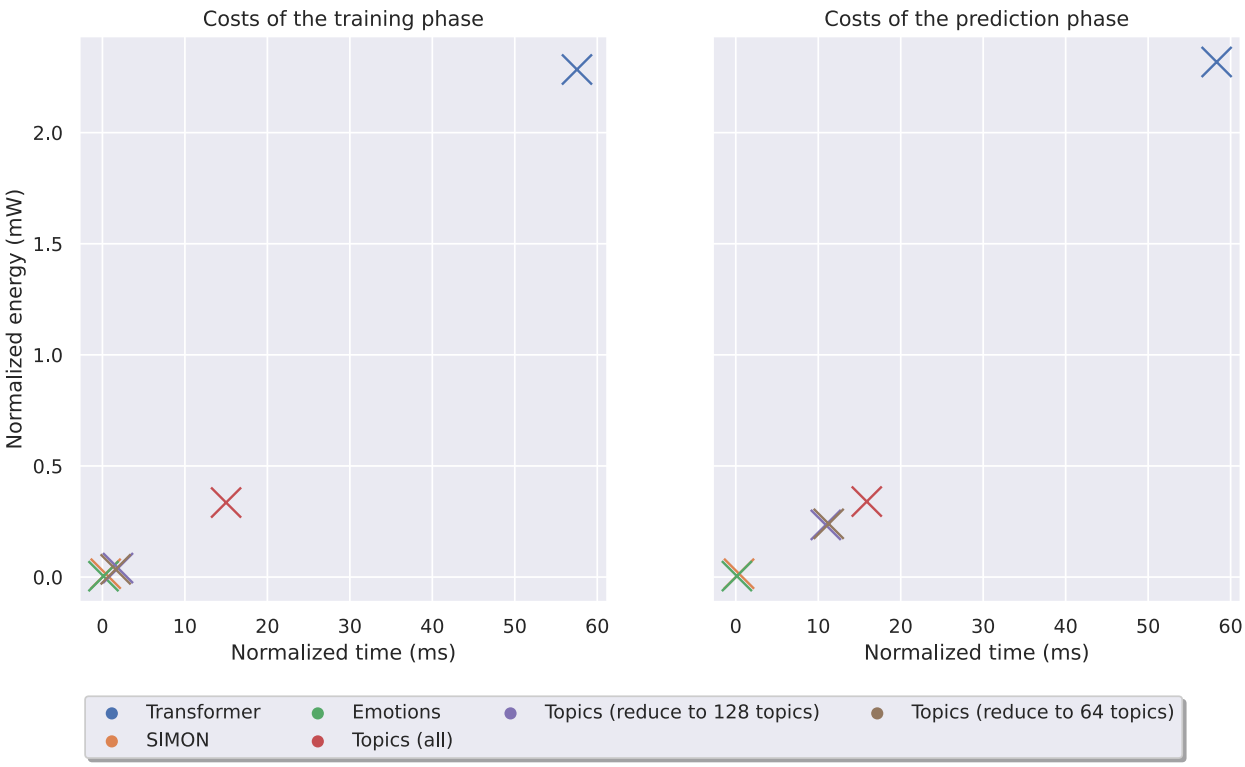


Figure 4.3: Comparison of the costs as given by the normalized time and duration of the analysis, for both the training and prediction phases, for each model, in the DepSign dataset. The horizontal axis represents the normalized time (per text processed) in milliseconds, while the vertical axis represents the normalized energy (per text) in milliwatts. The colors of the data points, as can be seen on the legend on the inferior side of the image, represents the models. The graph on the left represents the training phase, and the on the right the prediction phase.

In Figure 4.4, the normalized cost in terms of energy was plotted against the best macro-averaged F-score that was obtained for the different combinations of features. This best F-score was taken without distinguishing the classifier that obtained it. As was mentioned, different number of topics were analysed for the DepSign dataset. However in order to simplify the visual representation, only the results for all the topics were plotted. On this graph it can be seen even more blatantly that the use of SIMON over Transformer models greatly decreases the energy consumed, while the F-score is reduced only in 2%, as mentioned. It can also be seen that the usage of Transformers and topic information produces the worse outcome in terms of F-score, as well as costs. In terms of F-score, topics decrease this metric and increase the cost, which might indicate the benefit of abstaining from their use. However, the analysis of the individual feature importance will analyse whether this is the case. In addition, the data points for the Transformer model and SIMON are neatly clustered in two groups.

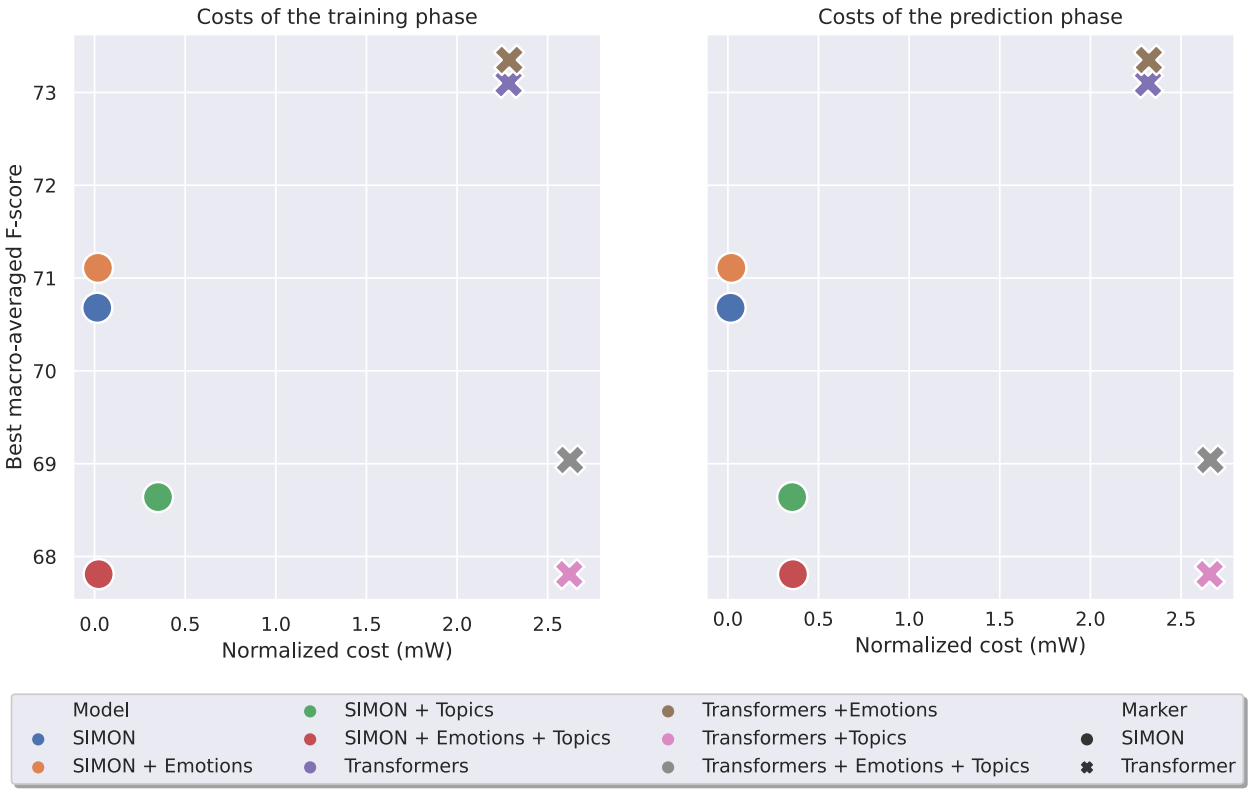


Figure 4.4: Comparison of the normalized cost in terms of energy consumed per text processed in milliwatts plotted against the best F-score obtained by that feature combination. The round data points represent those feature combinations that use SIMON, while the crosses represent those that use Transformers.

In order to analyze the **individual feature importance**, Figure 4.5, Figure 4.6 and Figure 4.7 were generated using SHAP. These plots represent the classification criteria for the classes "severe depression", "moderate depression" and "not depressed" from the DepSign dataset. The SHAP analysis was performed on the classifier forest of randomized trees. It must be noted that this feature combination is not the best in term of classification metrics. However, the combination of features it employs is the most informative, which is why it will be used for the following analysis. The classifier forest of random trees was chosen for this analysis, but others could have been used. For analyzing the feature importance of the text representation module, SIMON was selected, since computing such analyses using larger Transformers models incurs in much higher computational costs.

The first class to be analyzed is **"severe depression"**, as represented in Figure 4.5. The horizontal axis is the SHAP value, or the impact on the model output. Thus, all the positive values on the X-axis, which displays SHAP value, indicate that that feature in particular has an effect on the classification of the text as "severe depression", while the negative values of the x-axis indicate that the text does not belong to the aforementioned class. The colors of the graph, as can be seen on the right border of the image, indicate the original feature value, with red being high and blue being low [1]. Each instance is represented as a dot on the feature row, thus displaying density. Finally, the tags that appear on the left column of the image, which are generated by SHAP, were manually colored. Thus, the texts in blue indicate that that feature belongs to topic information, and the ones in green, that the feature represents information about emotions. The semantic information tags were left black. The tags are listed from more relevant to less relevant to the classifier.

With all this information, an example to illustrate the interpretation can be made. For instance, a high value of the blue category {*number mg, side effects, sertraline, adderall*}, which represents a topic regarding medications (sertraline is an antidepressant, while Adderall is used for the treatment of ADHD and narcolepsy), has a very strong effect in favor of classifying this text as "severe depression". In this case, a "high value" means a high probability as computed by BERTopic. It is interesting that topics features are so relevant for the classification task, as it can be seen on Figure 4.2 that the addition of topics can actually hinder the performance of the classifier. Despite reducing the macro-averaged F-score, the classifier is using the information from these characteristics in an assiduous manner.

Analysing this graph, it is possible to see which topics indicate the presence of severe depression. Interestingly, compared to the other two classes in this dataset, the SHAP

[1] `https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html`

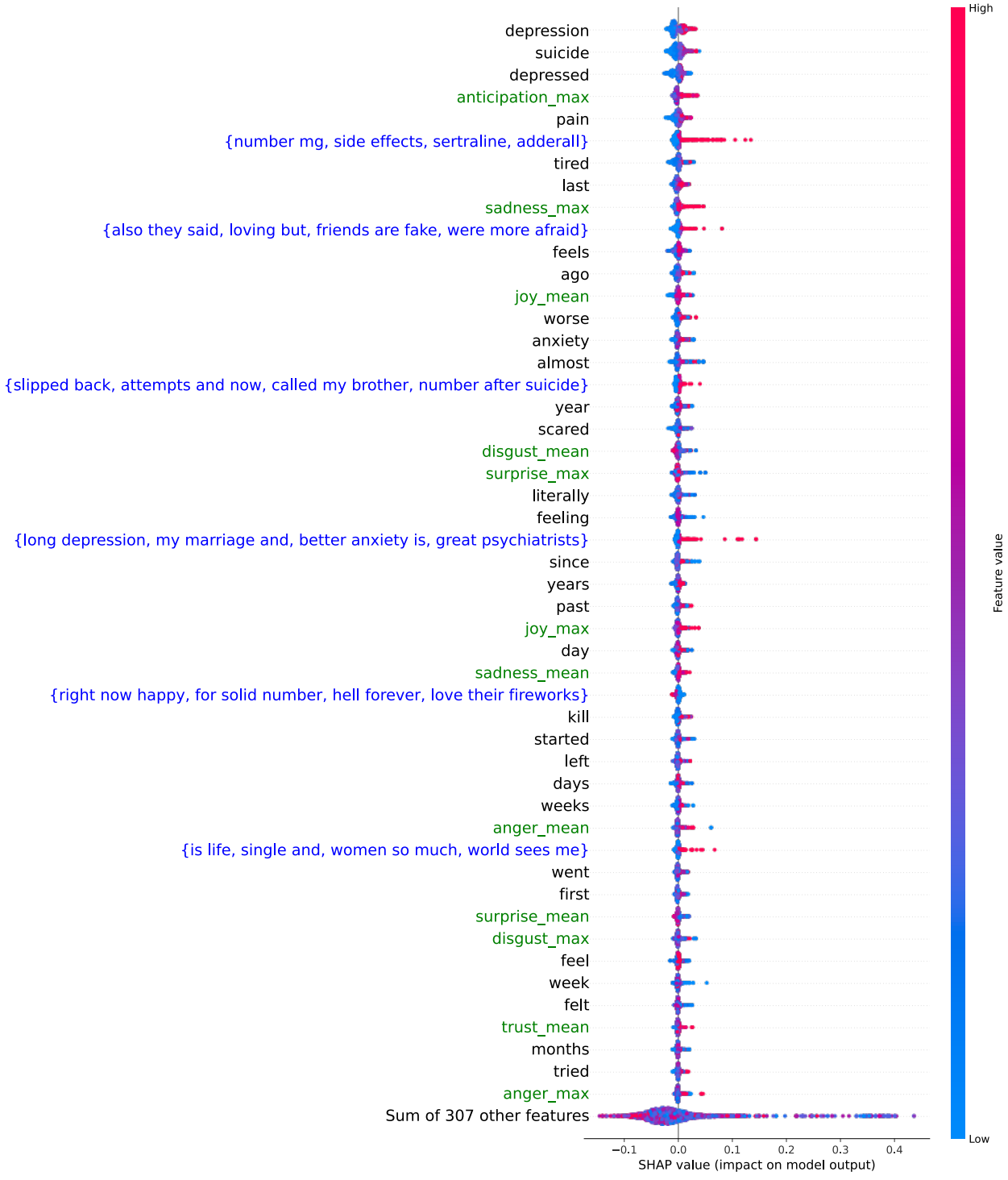


Figure 4.5: SHAP representation of the Random Forest algorithm application on the DepSign dataset, for the class **"severe depression"**, considering the features SIMON, emotions and topics. The black tags represent semantic information, while the blue tags represent topic information, and the green ones, emotions.

representation for "severe depression" has considered many more topics among the ones shown. These topics include themes such as medications, as well as suicide, fear, anxiety, feeling one´s friends are not genuine, and psychiatrists, among others. These themes are not surprising to find, since they correlate well with signs for depression [44], but having concrete themes which people talk about could be useful to gather signs of alert that could be useful for online monitoring of depression on social media.

The importance of emotions must also be noted, being some of the most important features. The presence of positive emotions such as anticipation, could indicate ironic language or noisy instances in the annotation. However, it could also be due of the mood disorders that can be caused by depression, which include increased impulsivity or increased participation in high-risk activities [44]. In the case of emotions, a higher value represents more of that emotion, while a lower value shows less (or no) presence of that emotion. Thus, when the presence of an emotion is discussed, it means that the value for that emotion is higher rather than lower.

Because of this observation, a manual analysis of the texts in the DepSign dataset [33, 55] was performed. The presence of sadness is apparent throughout the whole dataset, but the presence of joy is of more interest for this analysis. Different examples were found. Firstly, a person describing having a good day, despite their illness:

> *"good day : So I definitely have my demons that I fight everyday, I have anxiety and depression and I also think I'm bipolar. But today. Today I'm happy to be alive. That's all that matters right now and it feels so damn good! I know I'll have many more bad days but if I can remember this feeling I'll be ok."*

Another example case is of a person describing their improvement after beginning therapy: *"I have recently started therapy and medicine, and it's been amazing. I feel great. Happier than I think I've ever been, and for the first time I don't want to actively die all the time."* Other people also described improving with their medication, although sometimes it stopped working after some time. Another example of text with words that could have been understood by EmoFeat as having "joy", but in reality expresses the opposite, is as follows: *"Life is overrated imo. : My life is not precious and important. When people say life is precious and important, I just did not understand that and become baffled."*. Another case similar to the previous one, that also seems to convey some sense of anticipation is: *"Fantasizing about suicide is the ONLY nice thought I have now. Happy memories, people I've loved, achievements I've had, books I've read and loved, I know they "must have happened", but it's like I can't access that stuff."* Many cases of drug abuse were found, which would be

an example of the aforementioned high-risk activities. Specifically, a text describing drug abuse and high feelings of anger is as follows:

> *"I'm the worst person while under the influence. I become angry and hateful. So it's either I isolate myself because I'm a delicate flower that needs to be protected or I damage my reputation by drinking to numb the pain and being a d*ck to everyone around me. [...] Either way it feels like literal HELL I'm weak and not strong enough for this life everything is so f*cked up and I hate everyone and everything."*

After performing this manual analysis, a more thorough computational analysis of the texts was made. To this purpose, the vectors containing the information pertaining emotions in each of the texts were ranked according to different emotions. Then, the texts themselves were recovered. Finally, the comparison between the words in the text and in the emotion lexicon was made in order to see which words could be causing the effects on the emotion detection, and extracting the ones with a high value for the emotion being analyzed. The preprocessed texts were used for this analysis to replicate the experiments done previously, but the original texts are displayed for legibility.

For instance, an analysis was performed on the emotion sadness_max. The following text was recovered as an example, and then the five words with the most importance for the sadness emotion were found. These were: *depressed, suicides, depression, slipping, feel.* The text with these word in bold can be seen in the following quote. Four of these words can be seen in the semantic information given by SHAP.

> *"Anyone **feel** their reason/rationality **slipping**?: This is hard to describe, but I feel like the **depression** is overall getting worse because I'm less able to distinguish between "my voice" and "the **depression** voice." It wasn't long ago that I would get negative thoughts (my **friends** tolerate me, I'll never reach this goal, etc) and be able to take a second and go ok, this isn't me talking, it's the **depression**. Like I could point out an irrational belief to myself. Now I'm not sure if my thoughts are rational or not, like the "**depressed** me" and "me" have merged. I genuinely don't know if it's irrational to think that I'll never be in a relationship, thinking that some **suicides** can be considered rational vs the result of a sick brain, or thinking that my **depression** will only get worse. I thought I would always have reason (barring something like dementia), but even that is questionable at this point and it's scary as hell."*

Another example can be extracted for the emotion joy_mean. The words most relevant for the "joy" emotion are as follows: *happy, goofy, symptoms, life, developing* . They can be seen in bold in the text that follows. As was theorized in the manual analysis, the text itself does not express joy, but the words used are extremely positive, to refer to events that no longer are.

> *"Desperately need help but can never seem to get results : I'm at a complete loss and have no idea where to look. My whole* ***life*** *people have perceived me as the funny,* ***goofy*** *extrovert. To be fair it's a role I've given to myself, growing up I would always try to dismiss anything I found negative with a joke. Being the funny one allowed me to hide from reality and kept me* ***happy****. This worked fine throughout elementary school and made me a pretty* ***happy*** *kid, yet in middle school things started taking a turn for the worse. For as I can remember I've struggled with ADHD and anger issues. After middle school and throughout high school I started* ***developing*** ***symptoms*** *of anxiety, depression, insomnia, and borderline bipolarism."*

Other relevant emotions like trust_mean are triggered by words like *accept, improve, develop, plan, future*. Some of these same words, for instance, also have a high value for anticipation. Anticipation_max can be exemplified by the following texts, where the trigger words for this emotion are in bold.

> *"What is a painless and fast way to kill your self? : I'm just* ***wondering****, because I consider committing suicide in the near* ***future****."*

> *"41m stuck in a* ***dead*** *end job. i dont have any skills to get a* ***new*** *job and i have social anxiety. : i suffer from severe depression and social anxiety. i dont see the light at the end of tunnel. all i* ***see*** *is darkness and it is consuming me. dont know how long i can take this. i wish i can quit my job but i have a* ***mortgage****."*

In the case of this last text, "number", which is introduced by the text preprocessing, also has meaning for the emotion lexicon. In these two examples it can be seen that although the emotion analysis assigns high value of anticipation for these texts, they hardly have the positive component one would expect of this emotion.

Similarly, the semantic information has high importance, with the words *depression, suicide* and *depressed* being the three more relevant features. This could indicate that depressed people tend to be quite direct when speaking about their condition online, and

thus can indicate what type of online communities should be monitored if one wanted to be able to apply these systems to detect depressed individuals. As was mentioned, these words are also very relevant for the sadness_max feature. However, the word "scared" does appear as relevant in the classification, but "fear" is not as relevant as the other emotions. Finally, it is important to point out that there are many words that relate to time, the pass of time, or the past. These are: *last, ago, year, since, years, past, day, days, weeks, week, months*. This pattern of speaking of the past has been seen in the previous examples.

Despite being identified as relevant by the classifier, it has been observed that adding contextualization information does not mean that classification will always improve. So much so that, as can be seen in 4.4, sometimes adding this information can cause the macro-averaged F-score to drop below the unigram baseline´s values.

The SHAP graph for the **"moderate depression" class** 4.6 was also obtained. Most notably, the three topics which we can observe indicate against the belonging of a text to this class. That is, a high presence of these topics indicates against the class. Two of them, {*number mg, side effects, sertraline, adderall*} and {*also they said, loving but, friends are fake, were more afraid*} appeared in Figure 4.5 as positively indicating severe depression. The third one, {*teacher, because either, bridge program, she is sick*}, appears in Figure 4.7 as positively indicating the absence of depression.

Regarding emotion analysis, it can be seen that some emotions that helped classify a text as "severe depression", work in the opposite way for the classification of "moderate depression". This is the case for anticipation_max, sadness_max, anger_max, joy_max, trust_max. Other emotions like fear_mean and fear_max also go against the classification as "moderate depression", while others such as disgust_max and surprise_max are not conclusive for classification. One emotion that does favor the classification as "moderate depression" is anger_mean.

An example of anger_mean is extracted using the computational method: *"i knew it. : you all don't **care**, i **hate** you all **f*cking people**."* This text shows anger directed towards other people, while the example *"I **hate** myself and **want** to die : [removed]"* shows anger towards oneself.

In this class the word *depression* also appears as the first word in the tags list, with its presence (high value) indicating mostly against classification as moderate depression, and a low value appearing both for and against classification. Thus, the presence of this word indicates against 'moderate depression", but its absence is not informative. Both different words as well as others that had previously appeared in the graph for "severe depression" can be seen on the graph, but their effect seems more indecisive.

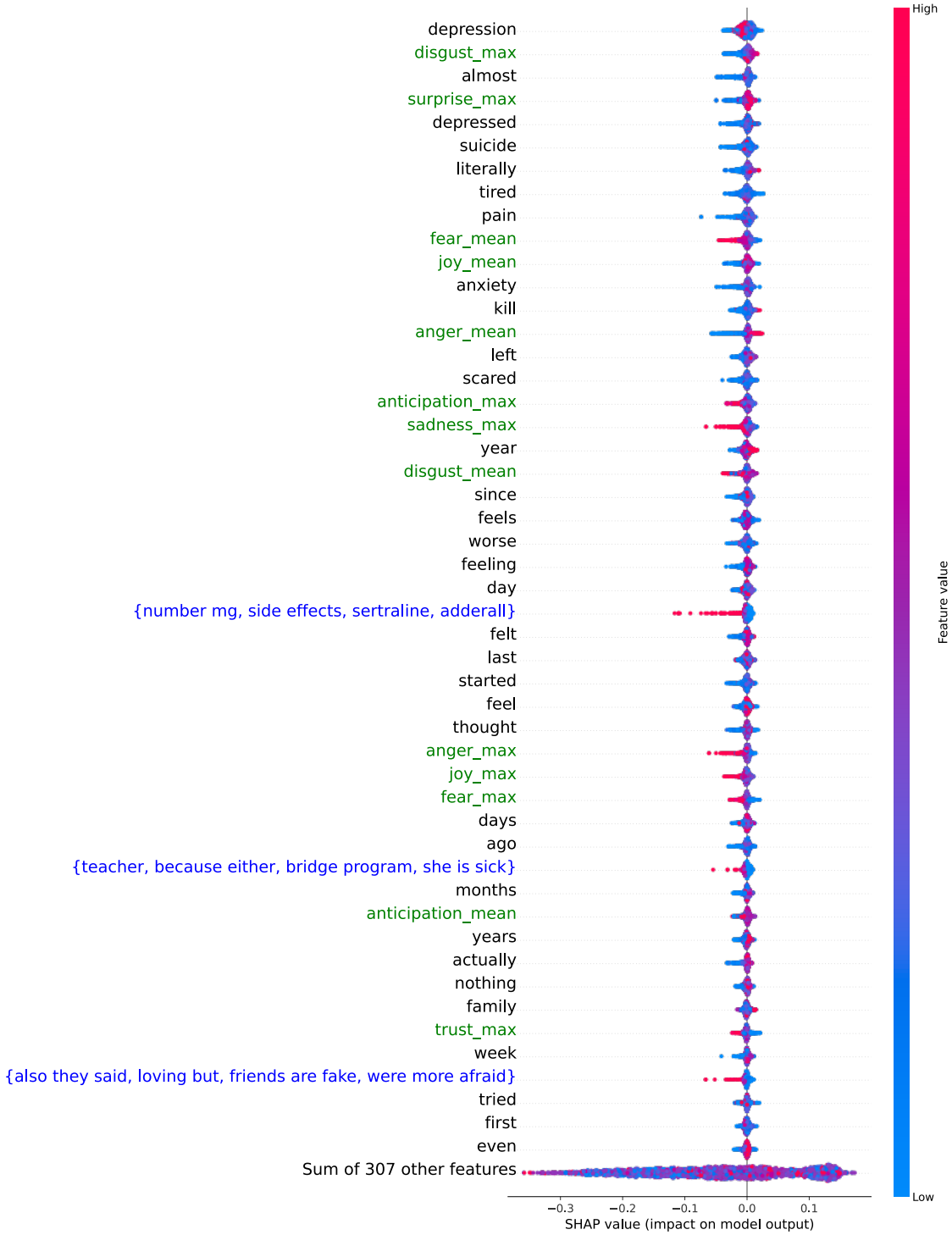


Figure 4.6: SHAP representation of the Random Forest algorithm application on the DepSign dataset, considering the features SIMON, emotions and topics, for the class **"moderate depression"**. The black tags represent semantic information, while the blue tags represent topic information, and the green ones, emotions.

Finally, the SHAP graph for the **"not depressed" class** was also obtained. Analyzing the topics extracted, two of them can be observed in this graph. The first one that positively indicates "not depression" is {*teacher, because either, bridge program, she is sick*}, as was mentioned. The second one is {*right now happy, for solid number, hell forever, love their fireworks*}.

Again, regarding the emotions, joy_mean and disgust_max indicate against "not depression". A few of the emotions are not conclusive, but fear_mean, fear_max, disgust_mean and trust_max indicate against depression. Anger_mean indicates depression, as was seen in the two previous SHAP graphs. Fear_mean, fear_max, disgust_mean, anger_max and trust_max indicated the class "not depression". A few example of these are as follows. For the disgust_mean emotion, there is the following example: *"I **live** through temporary pleasure : I just go day to day, that's it. I **lust** for pleasures that deprive me once I'm gone because I cannot have what **others** around me do"*. For the emotion anger_max, there is the text: *"**Fighting** **Parents** : What do I do if my **parents** are **fighting** and are on the brink of divorce?"*

Regarding semantic information, as has been seen, the word "kill" indicates positively the classification for both the class "severe depression" and "moderate depression", but in this case, its presence indicates against the class "not depression". Other words such as *suicide, depressed* and *depression* also tend to indicate against the class. Some other words, such as *pain, tired*, that were indicative of classification in the class "severe depression" appear, but are quite inconclusive.

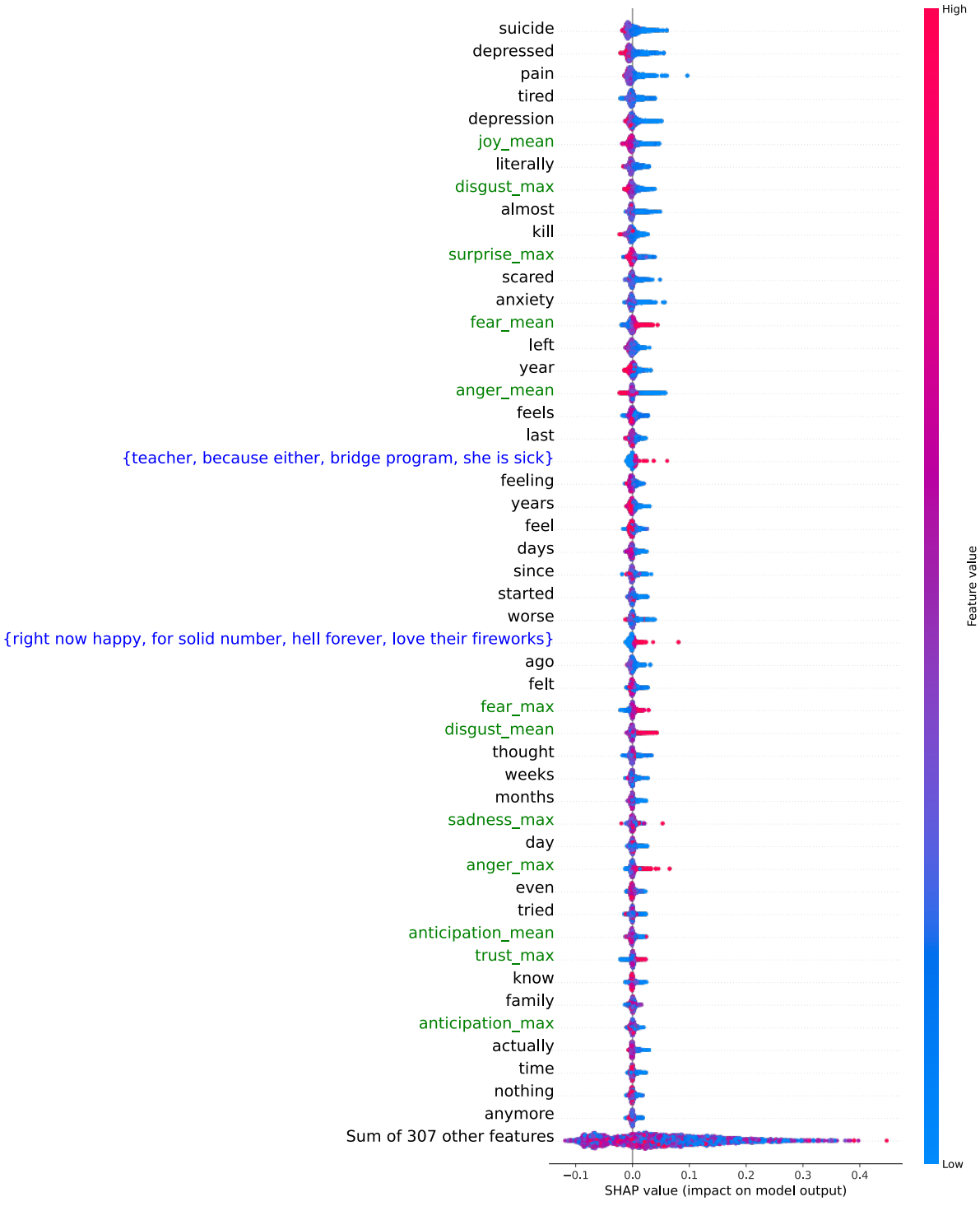


Figure 4.7: SHAP representation of the Random Forest algorithm application on the DepSign dataset, considering the features SIMON, emotions and topics, for the class **"not depression"**. The black tags represent semantic information, while the blue tags represent topic information, and the green ones, emotions.

4.4.2 Results of the Spanish dataset

In the case of the dataset, in Spanish, the **macro-averaged F-score** results can be found in the table 4.6. The best classifier is Linear SVC, using Transformers without emotion or topic information, with a macro-averaged F-score of **93.32%**. Contrary to the case of the DepSign dataset, the feature extraction done by the Transformer models provides much higher macro-averaged F-scores than with SIMON, the best of them being roughly a 10% better than the best of the scores provided by SIMON. It can be observed that for certain combinations of features and classifiers, the unigram baseline (as was also the case in the DepSign dataset) can sometimes provide better classification metrics that the ones provided by SIMON. However, because of the simplicity of the baseline, it would be expected for the results to be less generalizable.

			Dataset	**Spanish**			
			Classifier	**Forest of randomized trees**	**KNeighbors Classifier**	**Linear SVC**	**Polynomial SVC**
Feature extraction	**Emotions Emotions**	**Topics Topics**	**Number of topics**				
Unigram baseline	**NO**	**NO**	**0**	83.18	79.59	82.87	80.42
SIMON	**NO**	**NO**	**0**	82.27	77.11	80.75	80.57
		YES	**All (21)**	83.30	79.08	82.42	82.57
	YES	**NO**	**0**	82.42	75.28	79.84	79.24
		YES	**All (21)**	83.31	77.40	82.73	80.61
Transformers	**NO**	**NO**	**0**	90.14	**89.67**	**93.32**	**83.57**
		YES	**All (21)**	**90.60**	88.29	90.76	83.33
	YES	**NO**	**0**	89.99	89.37	93.02	**83.57**
		YES	**All (21)**	90.60	88.45	90.30	83.33

Table 4.6

The **computational costs** for this dataset can be found on Table 4.7. As was previously mentioned, the Transformer, SIMON and Emotion features process both the train or test set as-is. The "all topics" feature considers the fit-transform of the train set, and the

transform of the test set. In this case, the normalized training time, training energy and prediction energy are three orders of magnitude larger for the Transformer models, while the normalized prediction energy is two orders of magnitude greater. It is possible that the amount of time and energy required are not linear during the execution of these models, and because this dataset has a much smaller size, the costs are less distributed among the samples. Again, this increase in costs must be weighed against the improvement of the classification capabilities of the method. In this case, however, the much greater improvement may tip the scales in favor of using the better system, if a higher classification performance is required.

Model	Normalized training time (s)	Normalized prediction time (s)	Normalized training energy (kW)	Normalized prediction energy (kW)
Transformers	1.02E-01	9.40E-02	3.99E-06	3.69E-06
SIMON	1.45E-04	1.40E-04	4.62E-09	4.79E-09
Emotions	**4.69E-05**	**6.13E-05**	**1.05E-09**	**1.56E-09**
All topics	2.17E-02	1.79E-02	4.57E-07	3.77E-07

Table 4.7: Computational costs for the Spanish dataset.

As was done previously, Figure 4.8 shows the costs in terms of energy and time. As was the case with the DepSign dataset, the distributions of costs is similar, with the Transformer model having a much larger cost in terms of energy and time, while SIMON and emotions overlap near the zero.

Figure 4.9 shows an even more dramatic separation between the results of the Transformer model and SIMON, with the results crowded on opposing sides of the graph. So much so, that the data point showing the *SIMON + Topics* feature combination is hidden behind the *SIMON + Emotion + Topics* combination.

Another relevant thing to note about the computational costs is that, same as it happened in the DepSign dataset, the energy required to extract emotions is lower than all other methods used. This is congruent with the implementation of EmoFeat, as it is a straightforward method. This is favourable in the case of the DepSign dataset, since adding these methods with little extra cost consistently improve the results obtained. In the case of the Spanish dataset, however, this is not the case, since adding emotion information actually reduces the F-score. Since the Spanish emotion lexicon is of a much smaller size

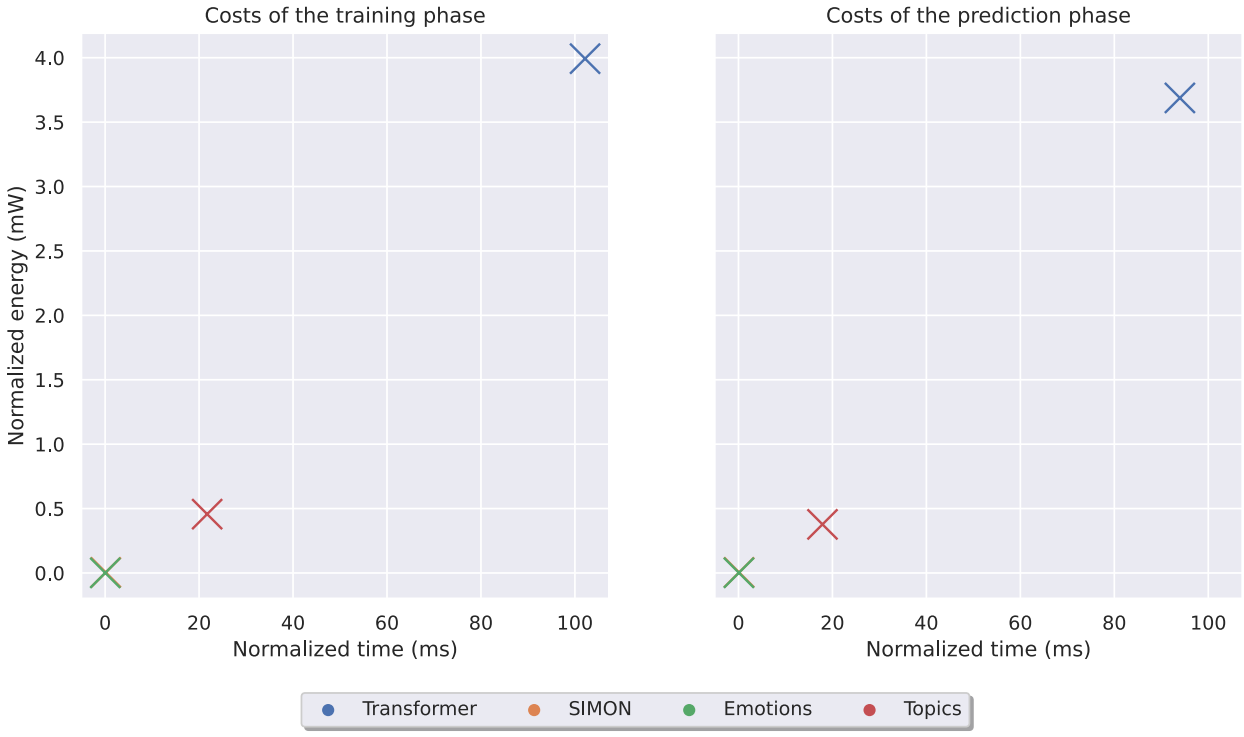


Figure 4.8: Comparison of the costs as given by the normalized time and duration of the analysis, for both the training and prediction phases, for each model, in the Spanish dataset. The horizontal axis represents the normalized time (per text processed) in milliseconds, while the vertical axis represents the normalized energy (per text) in milliwatts. The colors of the data points, as can be seen on the legend on the inferior side of the image, represent the models. The graph on the left represents the training phase, and the on the right the prediction phase.

than the DepSign dataset, this difference in results may be caused by the Spanish emotion lexicon having less coverage.

Finally, as was with the DepSign dataset, the difference in size of the vectors generated by SIMON and the Transformer model is even greater, with SIMON generating vectors of size 48, while the ones by the transformer are of size 1024. Again, if large datasets are considered, this difference in size could become significant.

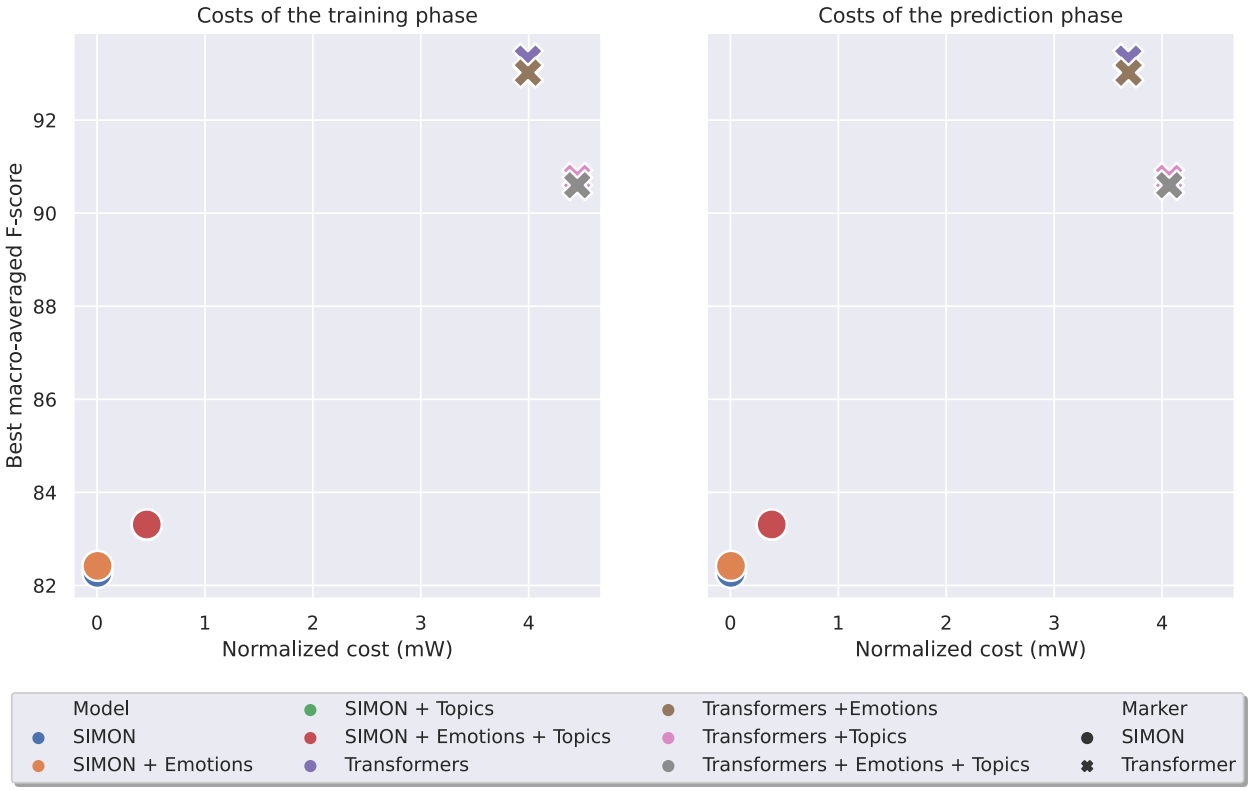


Figure 4.9: Comparison of the costs as given by the normalized time and duration of the analysis, for both the training and prediction phases, for each model, in the Spanish dataset.

Figure 4.10 is, as previously, generated using SHAP, and allows for the analysis of **individual feature importance**. However, as the dataset was in Spanish, the labels for this image have been translated to English, trying to preserve the details implied in the Spanish words. As such, where there was no direct word for word translation, more words may have been used, or clarification between square brackets included. Again, topics are colored in blue, while emotions are colored in green, and semantic information remains in black. The forest of randomized trees algorithm was also used for this diagram.

Because this dataset is a binary classification, the positive SHAP values can be interpreted as indicating depression, while the negative ones indicate the absence of it. When analysing this dataset, the first thing that can be seen is that in this case, the emotions seem to have less importance for classification, with only "sadness_max" appearing in the tags displayed, and indicating the presence of depression. Again, the reduced importance of emotions in the Spanish dataset compared to the DepSign dataset could be due to a decreased coverage of the emotion lexicon (see Sect. 4.2).

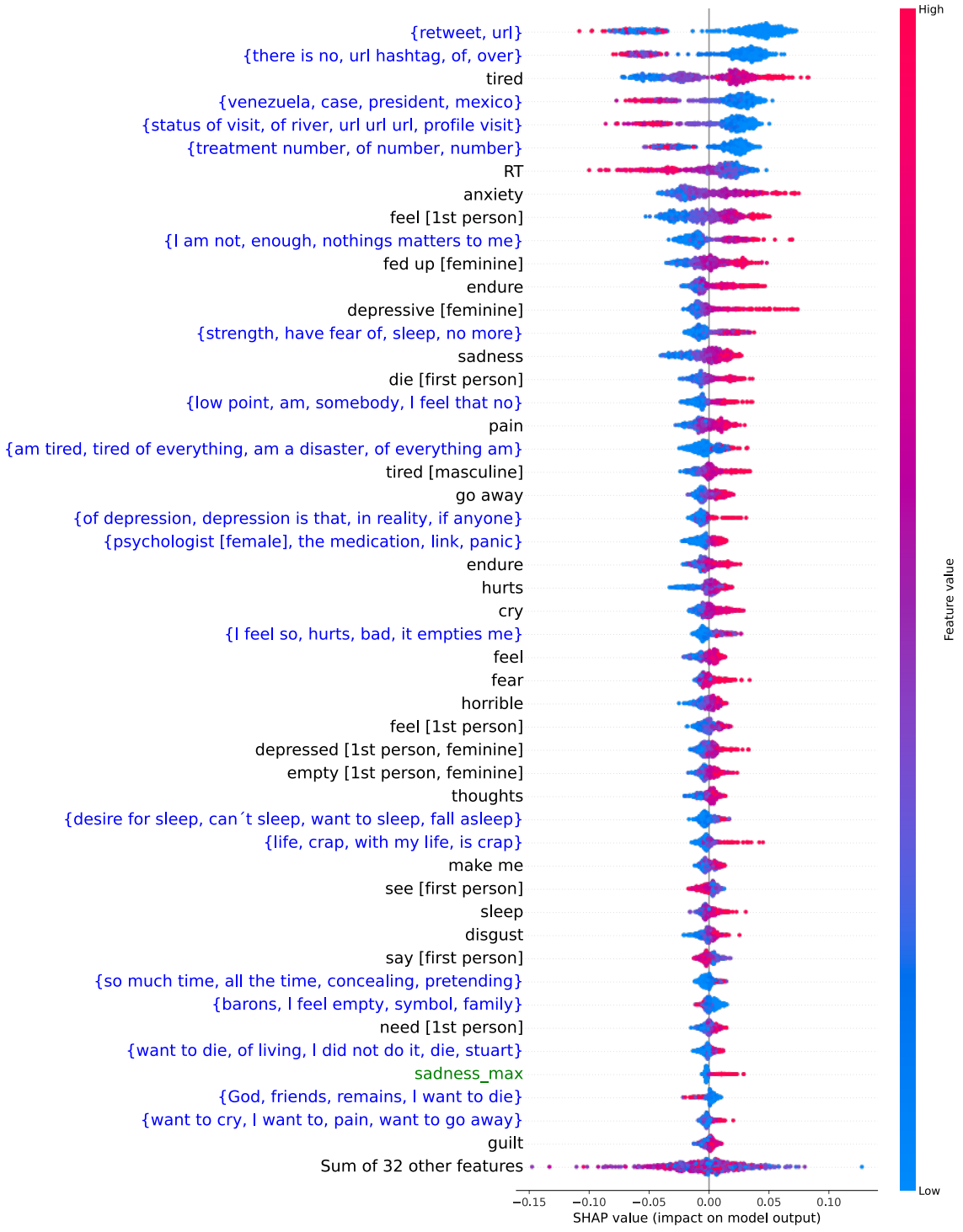


Figure 4.10: Translated SHAP representation of the Random Forest algorithm application on the Spanish dataset, considering the features SIMON, emotions and topics. The black tags represent semantic information, while the blue tags represent topic information, and the green ones, emotions.

The first two items on the graph ({*retweet, url*}, {*there is no, url hashtag, of, over*}) are topics that convey information about behaviours and language pertaining to this specific social media. The presence of these topic indicates that it is likely that the tweet is not depressive, but their absence does not give information about the classification of the text. Still, it gives information on the usage of links to outside content, hashtags and retweeting that could be useful for an initial filtering.

Both topic information and semantic information indicate as favourable for depression classification themes that are known to be signs of depression, such as problems with sleep, tiredness, and hopelessness [44]. While emotion contextual information is used to some degree for the classification of this dataset, the contextual information provided by topic information is much more relevant. These topics are distributed differently depending on the class, as can be seen in Figure 4.11. The labels on this image are not translated. Instead, the translated version can be found on Table 4.8, which has also been edited for legibility. The topic number is displayed in both Figure 4.11 and Table 4.8. The bars displayed represent the normalized frequency of the topic in each class, starting from the bottom. As such, the lowest orange bar is topic 0, the blue bar on top of it is topic 1, and so forth.

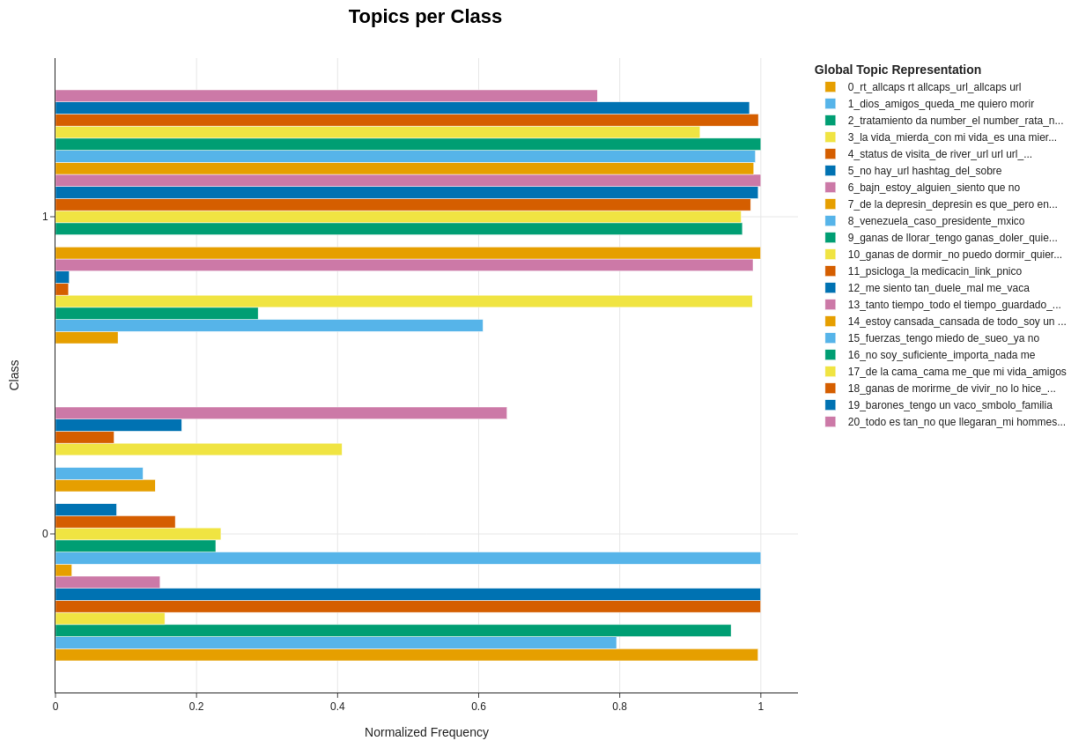


Figure 4.11: Distribution of topics per class for the Spanish dataset, where 1 indicates the class “depression” and 0 the class “not depression”.

Topic number	English translation
0	{retweet, url}
1	{God, friends, remains, I want to die}
2	{treatment number, of number, number}
3	{life, crap, with my life, is crap}
4	{status of visit, of river, url url url, profile visit}
5	{there is no, url hashtag, of, over}
6	{low point, am, somebody, I feel that no}
7	{of depression, depression is that, in reality, if anyone}
8	{venezuela, case, president, mexico}
9	{want to cry, I want to, pain, want to go away}
10	{desire for sleep, can´t sleep, want to sleep, fall asleep}
11	{psychologist [female], the medication, link, panic}
12	{I feel so, hurts, bad, it empties me}
13	{so much time, all the time, concealing, pretending}
14	{am tired, tired of everything, am a disaster, of everything am}
15	{strength, have fear of, sleep, no more}
16	{I am not, enough, nothings matters to me}
17	{from bed, bed I, that my life, friends}
18	{want to die, of living, I did not do it, die, stuart}
19	{barons, I feel empty, symbol, family}
20	{everything is so, they wouldn't arrive, my men are, my room stinks}

Table 4.8: Translated topics for the Spanish dataset. The topic number corresponds to the numbers displayed on Figure 4.11.

It can be easily seen that the distribution of topics is different in both classes. In the "depression" class, topic 8 is not present, which relates to political themes, while topics 4 and 5 are barely present. Topic zero, which translates to {*retweet, url*} has a much lower frequency, which can indicate differences between the online behaviours of people with and without depression, since not depressed people seem to retweet more and use more hyperlinks. This conclusion is similar to the one extracted analysing the SHAP graph.

Likewise, there are topics that do not appear in the non-depressed class: topic 13 ({*so much time, all the time, concealing, pretending*}) and topic 16 ({*I am not, enough, nothings matters to me*}). Both of these topics are well correlated with depressive symptoms, although are not exclusive to depression. Similarly, topics 3, 6, 7, and those between 9 and 19, are lower in the not depression class. This makes sense, since most of these are depression related, with themes such as wanting to cry, being in pain, wanting to sleep, discussion of psychologists and medication, panic, being tired, not feeling enough, and so forth.

Despite this, it can be seen that both topics 1 and 20 are high in both depression and not depression classes, which could indicate themes that are not exclusive to depression. Also, as can be seen on Table 4.8, some of the topics do not seem to be clearly defined (like 19 or 20) which makes them hard to interpret, and could also be introducing noise in the classification. In addition, the complete absence of texts without certain themes in a class may indicate that there are not enough texts in the dataset, since themes like politics are not exclusive to non-depressed people. As such, when applied to real world cases, this could cause the classifier to automatically classify any texts related to politics as not depressive. Similarly, topic 13 and 16, which cover themes such as pretending or concealing oneself, as well as feelings of not being worthy or of apathy, are related to depression, but not necessarily exclusive, but are only present in the depression class. For this reason, the topic information extracted from this dataset could benefit from a larger dataset.

CHAPTER 5

Conclusions and future work

After the development of this work, all the objectives have been achieved. The first Objective 1 of collecting and studying datasets and different resources was explained at length in the Chapter 3 and Chapter 4, in Section 4.2 .

Addressing Objective 2, as has been mentioned during the analysis of the results for both datasets (Section 4.4), contextualization through emotion and topic information has proven to be informative in the detection of depression, being capable of generating information useful beyond the current task. This information has been shown to relate well to signs that indicate depression, but is further capable of providing concrete examples of behaviours that may be indicative of its presence. This information has been specially relevant in the DepSign dataset, where both topics and emotions have been considered by the classifier as being relevant to classification. Many concrete examples have been extracted and examined, giving real world examples of characteristics related to depression. For instance, the topic covering medications, or the presence of words such as *depression, suicide* and *depressed*, strongly indicate severe depression, showing that there are concrete examples than can be learned from interpreting results obtained from ML algorithms. In the Spanish dataset, however, emotions have been less informative. Comparing dataset and emotion lexicon size, though, it can be theorized that a lower emotion lexicon coverage might have led to worse emotion information, and likewise, having less texts might have made it harder for the

BERTopic library to extract enough topic information, although insight has been extracted from the topics.

Addressing the final Objective 3, we have identified a clear candidate application in the experiments carried out in which the trade-off between computational costs and classification performance is appropriate. Thus, while it is undeniable that state of the art techniques provide a unique opportunity for applications in many fields, it must be noted that, for others, such advances techniques may not be necessary, as pre-existing techniques may offer a better balance between computational cost and results for suitable scenarios. In this work, such a candidate application has been found for the DepSign dataset. Thus, unless the application requires the highest possible F-score, the best classifier that was obtained in this work, a forest of randomized trees with a features obtained using Transformers and emotions (without topics) could be substituted by the features extracted from SIMON with emotions. This way, the F-score would only be reduced by roughly a 2%, but costs would be more than a hundred times smaller.

Finally, it is worth mentioning that the interpretability of classifiers is of great value as the analyses performed in this work depend on the interpretability of the different learners. This allows us to gain insights into the classification process and extract aggregated knowledge. As such, not only can this information be applied for further uses, but the functioning of the algorithm can be better assessed and understood. This is of great importance, specially in any areas related to health, as the implementation of these methods will require of the understanding and acceptance of clinicians.

There are some limitations to this work that must be noted. Firstly, the datasets used model depression as a discrete category, but the depression condition is far more complex, including major depression and persistent depressive disorder, among others [44]. This nuisance may affect the generalization capabilities of the developed learning models, as it simplifies the original problem. Finally, simple learning algorithms were applied to all models, in order to preserve their interpretability and thoroughly study the effect of the computed features, avoiding the side effect of more complex methods. However, it is likely that more complex learning algorithms would have outperformed the ones chosen. Instead, this work addresses the study of depressive language and thus we have not aimed to maximize the classification performance.

Despite these limitations, we consider that the models described in this work offer useful insight. For future work, in order to surpass these limitations, the usage of more nuanced datasets that make a distinction between types of depression is proposed. This would require clinical diagnosis by trained professionals. This interdisciplinary approach of cooperation

with mental health professionals could improve the outcomes of this experiments, as well as the quality of the knowledge extracted. In addition, in this work we have considered depression as an static label. However, this disease develops over a period of time. For this reason, further analysis of post history, that is, taking time into account, could be used to detect early signs of the disorder. Furthermore, additional testing combining the text representations provided by SIMON and Transformer models could be used to analyze whether the combination of both methods could provide additional information.

APPENDIX A

Impact of this project

In this appendix, the impact caused by this project will be assessed. Thus, the social, economic, environmental and ethical aspects will be analysed.

A.1 Social impact

As was introduced previously, depression is a widespread mental disorder that causes a great deal of suffering both to the person suffering it, as well as to those close to them. In addition, the costs of depression, both direct and indirect, take up the resources of society. By addressing issues related to this matter, this project aims to aid in providing means to better treat those who need it, and thus reduce the personal and societal costs caused by depression.

A.2 Economic impact

Early detection and treatment of depression can not only improve the outcome of the illness, but can also greatly reduce the costs associated with it, including treatment and medication,

but mainly, the indirect costs, which include medical leaves and reduced productivity. It is estimated that the cost of depression in Spain approaches the 1% of GDP [45]. If detection systems similar to the one suggested in this work were to be implemented, it would be possible to reduce this costs by providing better access to treatment.

A.3 Environmental impact

It is undeniable that the ever growing complexity of algorithm, as well as the increasing amounts of data available, are driving the energy consumed in natural language and machine learning algorithms. This work has tried to be conscientious of this matter, by proposing alternatives in order to decrease the energy consumed when performing the calculations needed for this implementation.

A.4 Ethical impact

On the ethical impact of this work, it is important to use these technologies in order to improve quality of life of people with mental health conditions. Their use for purposes of discriminating against or otherwise going against their well-being must be prohibited by law.

APPENDIX B

Economic budget

In this chapter, the economic budget of this project will be estimated, including the physical and human resources that have been used.

B.1 Physical resources

The physical resources used for this project include the use of a GPU Nvidia Titan X, as well as access to a remote server with 128 GB of RAM and 1T of disk space. It is estimated that this service would have cost around 400€ had it been hosted on the web.

B.2 Physical resources

To determine the approximate cost in human labor this project would have taken, the amount of hours it took for completion was estimated. This came to be estimated to around 350 hours. The gross expected salary for a recent engineer graduate is approximately 1460€ a month, which would be around 8.7€ an hour. Thus, the total approximate cost of this project for the hours taken by the student would be 3045€. To this cost, the salary of the

supervisor should be added.

Bibliography

[1] adjusttext - automatic label placement for matplotlib. URL https://github.com/Phlya/adjustText.

[2] The Jupyter Notebook x2014; Jupyter Notebook 6.5.4 documentation — jupyter-notebook.readthedocs.io. https://jupyter-notebook.readthedocs.io/en/stable/notebook.html. [Accessed 07-May-2023].

[3] pandas - Python Data Analysis Library — pandas.pydata.org. https://pandas.pydata.org/about/. [Accessed 07-May-2023].

[4] 1.11. Ensemble methods — scikit-learn.org. https://scikit-learn.org/stable/modules/ensemble.html#forest, . [Accessed 08-May-2023].

[5] 1.4. Support Vector Machines — scikit-learn.org. https://scikit-learn.org/stable/modules/svm.html#svm-classification, . [Accessed 08-May-2023].

[6] 1.6. Nearest Neighbors — scikit-learn.org. https://scikit-learn.org/stable/modules/neighbors.html#classification, . [Accessed 08-May-2023].

[7] sklearn.ensemble.RandomForestClassifier — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn-ensemble-randomforestclassifier, . [Accessed 08-May-2023].

[8] Fatima Alhaj, Ali Al-Haj, Ahmad Sharieh, and Riad Jabri. Improving arabic cognitive distortion classification in twitter using bertopic. *International Journal of Advanced Computer Science and Applications*, 13, 02 2022. doi: 10.14569/IJACSA.2022.0130199.

[9] Oscar Araque and Carlos A. Iglesias. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access*, 8:17877–17891, 2020. doi: 10.1109/ACCESS.2020.2967219.

[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. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.02.002. URL http://www.sciencedirect.com/science/article/pii/S0957417417300751.

[11] 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, 7 2017. ISSN 09574174. doi: 10.1016/j.eswa.2017.02.002.

[12] 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. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.12.005. URL http://www.sciencedirect.com/science/article/pii/S0950705118305926.

[13] 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, 2 2019. ISSN 09507051. doi: 10.1016/j.knosys.2018.12.005.

[14] Oscar Araque, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. Gsitk: A sentiment analysis framework for agile replication and development. *SoftwareX*, 17, 1 2022. ISSN 23527110. doi: 10.1016/j.softx.2021.100921.

[15] Aaron Baird, Yusen Xia, and Yichen Cheng. Consumer perceptions of telehealth for mental health or substance abuse: a twitter-based topic modeling analysis. *JAMIA Open*, 5, 4 2022. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooac028.

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

[17] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. URL https://crscardellino.github.io/SBWCE/.

[18] CodeCarbon, 2020. https://mlco2.github.io/codecarbon/ and https://github.com/mlco2/codecarbon.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

[20] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3207. URL https://aclanthology.org/W14-3207.

[21] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. pages 626–638. ACM, 2 2014. ISBN 9781450325400. doi: 10.1145/2531602.2531675.

[22] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110, 2016.

[23] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7:128–137, 8 2021. ISSN 2334-0770. doi: 10.1609/icwsm.v7i1.14432.

[24] Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Jerin Mahibha C, and Kayalvizhi Sampath. Findings of the shared task on detecting signs of depression from social media. Association for Computational Linguistics, 5 2022. URL `https://competitions.codalab.org/competitions/36410`.

[25] Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68, 2022. ISSN 1135-5948. doi: 10.26342/2022-68-3. URL `https://upcommons.upc.edu/handle/2117/367156`.

[26] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

[27] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[28] Aron Halfin. Depression: the benefits of early and appropriate treatment. *The American journal of managed care*, 13:S92–7, 11 2007. ISSN 1936-2692.

[29] S Hardik, Dhruvi Gosai, and Himangini Gohil. A review on a emotion detection and recognization from text using natural language processing. 04 2018.

[30] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[31] Oscar Araque Iborra. A distributional semantics perspective of lexical resources for affect analysis: An application to extremist narratives, 2020.

[32] Qilu Jiao and Shunyao Zhang. A brief survey of word embedding and its recent development. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 1697–1701, 2021. doi: 10.1109/IAEAC50856.2021.9390956.

[33] Sampath Kayalvizhi and Durairaj Thenmozhi. Data set creation and empirical analysis for detecting signs of depression from social media postings. In Lekshmi Kalinathan, Priyadharsini R., Madheswari Kanmani, and Manisha S., editors, *Computational Intelligence in Data Science*, pages 136–151, Cham, 2022. Springer International Publishing. ISBN 978-3-031-16364-7.

[34] Angela Leis, Francesco Ronzano, Miguel A. Mayer, Laura I. Furlong, and Ferran Sanz. Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis. *Journal of Medical Internet Research*, 21, 6 2019. ISSN 14388871. doi: 10.2196/14199.

[35] Ran Li, Yuanfei Zhang, Lihua Yin, Zhe Sun, Zheng Lin, Peng Fu, Weiping Wang, and Gang Shi. Emomix+: An approach of depression detection based on emotion lexicon for mobile application. *Security and Communication Networks*, 2022:1–12, 1 2022. ISSN 1939-0122. doi: 10.1155/2022/1208846.

[36] Yang Liu and Meng Zhang. Neural network methods for natural language processing, 2018.

[37] 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. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*, 2013.

[39] Tomas 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, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090.

[40] Saif M Mohammad. #emotional tweets. pages 246–255. URL http://www.ark.cs.cmu.edu/GeoText.

[41] Saif M. Mohammad. Best practices in the creation and use of emotion lexicons. 10 2022.

[42] Saif M Mohammad and Svetlana Kiritchenko. Submitted to the special issue on semantic analysis in social media, computational intelligence. using hashtags to capture fine emotion categories from tweets.

[43] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11:81, 12 2021. ISSN 1869-5450. doi: 10.1007/s13278-021-00776-6.

[44] National Institute of Mental Health. Depression, 4 2023. URL https://www.nimh.nih.gov/health/topics/depression.

[45] Médicos y Pacientes. El coste de la depresión en españa se acerca al 1% del PIB. https://www.medicosypacientes.com/articulo/el-coste-de-la-depresion-en-espana-se-acerca-al-1-del-pib, June 2016. Accessed: 2023-5-7.

[46] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[47] Gregory Park, H. Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108, 11 2014. doi: 10.1037/pspp0000020.

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

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

[50] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[51] Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.785.

[52] Anxo Pérez, Javier Parapar, and Álvaro Barreiro. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, 132:102380, 10 2022. ISSN 09333657. doi: 10.1016/j.artmed.2022.102380.

[53] Ismael Díaz Rangel, Grigori Sidorov, and Sergio Suárez Guerra. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, 29:31–46, 2014. ISSN 07185758. doi: 10.7764/onomazein.29.5. Citar para SEL lexicon emociones en español.

[54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. Association for Computational Linguistics, 2 2019. URL http://arxiv.org/abs/1908.10084.

[55] Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*,

pages 331–338, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.51. URL https://aclanthology.org/2022.ltedi-1.51.

[56] Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. Predicting depression and anxiety on reddit: a multi-task learning approach. pages 427–435. IEEE, 11 2022. ISBN 978-1-6654-5661-6. doi: 10.1109/ASONAM55673.2022.10068655.

[57] Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. *Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets*, volume 7629 LNAI, pages 1–14. 2013. doi: 10.1007/978-3-642-37807-2_1. URL http://link.springer.com/10.1007/978-3-642-37807-2_1. Cited By :81.

[58] Sushant Singh and Ausif Mahmood. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3077350.

[59] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.

[60] Manan Suri, Nalin Semwal, Divya Chaudhary, Ian Gorton, and Bijendra Kumar. I don't feel so good! detecting depressive tendencies using transformer-based multimodal frameworks. In *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, MLNLP '22, page 360–365, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399067. doi: 10.1145/3578741.3578817. URL https://doi.org/10.1145/3578741.3578817.

[61] Marcel Trotzek, Sven Koitka, and Christoph Friedrich. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. 09 2018.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[63] Vincent Warmerdam, Thomas Kober, and Rachael Tatman. Going beyond t-sne: Exposing whatlies in text embeddings. pages 52–60. Association for Computational Linguistics, 11 2020. doi: 10.18653/v1/2020.nlposs-1.8. URL https://www.aclweb.org/anthology/2020.nlposs-1.8.

[64] Vincent Warmerdam, Thomas Kober, and Rachael Tatman. Going beyond T-SNE: Exposing whatlies in text embeddings. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlposs-1.8. URL https://www.aclweb.org/anthology/2020.nlposs-1.8.

[65] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

[66] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[67] World Health Organization. Depressive disorder (depression), 3 2023. URL https://www.who.int/news-room/fact-sheets/detail/depression.