Detecting Mental Disorders in Social Media using a Multichannel Representation

by:

Mario Ezra Aragón Saenzpardo

A dissertation submitted in partial fulfillment
of the requirements for the degree of:

DOCTOR OF SCIENCE IN COMPUTER SCIENCE

at

Instituto Nacional de Astrofísica, Óptica y Electrónica
2022
Tonantzintla, Puebla

Supervised by:
Dr. Manuel Montes y Gómez INAOE
Dr. Adrián Pastor López Monroy, CIMAT

ABSTRACT

Millions of people around the world are affected by one or more mental disorders that interfere with their thinking and behavior. Timely detection of these issues is challenging but crucial since it could open the possibility to offer help to people before the illness gets worse. One alternative to accomplish this is to monitor how people express themselves, that is for example what and how they write, or even a step further, what emotions they express in their social media communications.

Over the last few years, studies related to the detection of mental disorders in social media have been increasing. The latter because the awareness created by health campaigns that emphasize the commonness of these disorders among all of us, that also has motivated the creation of new datasets, many of them extracted from social media platforms. In this study, we aim to contribute with the analysis of three major mental disorders that are hitting the world: Anorexia, Self-harm, and Depression. To this end, we propose a novel model that, first, extracts three different views, or information channels, from the posts shared by users: thematic interests, writing style, and emotions. Then, it fusions the information from each channel by using a gated multi-modal unit a module that learns the relations between channels. We evaluate the feasibility of our approach in the aforementioned tasks, first by comparing its output against traditional and modern strategies, and later against the best contestants in the eRisk evaluation forum, a workshop that explores issues related to the evaluation of methodologies and practical applications of topics related to health and safety for early risk detection on the internet. In both evaluations, our approach outperforms all of its competitors. Through an exhaustive analysis section, we provide evidence of what is being captured by each information channel, then highlight the importance and robustness of a more holistic view in critical classification tasks.

In our evaluation, we use recent public data sets for three important mental disorders:

Depression, Anorexia, and Self-harm. The obtained results suggest that the presence and variability of emotions, style, and thematic information captured by the proposed representations, allow us to highlight important information about social media users suffering from these mental disorders. Furthermore, the fusion of these representations can boost the classification performance. Moreover, these representations provide better interpretability to the results.

RESUMEN

Millones de personas en todo el mundo se ven afectadas por uno o varios trastornos mentales que interfieren en su pensamiento y comportamiento. La detección oportuna de estos problemas es un reto, pero es crucial, ya que podría abrir la posibilidad de ofrecer ayuda a las personas antes de que la enfermedad empeore. Una alternativa para lograrlo es vigilar cómo se expresan las personas, es decir, por ejemplo, qué y cómo escriben, o incluso un paso más allá, qué emociones expresan en sus comunicaciones en las redes sociales.

En los últimos años han aumentado los estudios relacionados con la detección de trastornos mentales en los medios sociales. Esto último se debe a la concienciación creada por las campañas sanitarias que hacen hincapié en lo común de estos trastornos entre todos nosotros, lo que también ha motivado la creación de nuevos conjuntos de datos, muchos de ellos extraídos de las plataformas de medios sociales. En este estudio, pretendemos contribuir con el análisis de tres de los principales trastornos mentales que azotan al mundo: Anorexia, Autolesiones y Depresión. Para ello, proponemos un modelo novedoso que, en primer lugar, extrae tres puntos de vista diferentes, o canales de información, de las publicaciones compartidas por los usuarios: intereses temáticos, estilo de escritura y emociones. A continuación, fusiona la información de cada canal utilizando una unidad multimodal un módulo que aprende las relaciones entre canales. Evaluamos la viabilidad de nuestro enfoque en las tareas mencionadas, primero comparando su resultado con estrategias tradicionales y modernas, y después con los mejores concursantes del foro de evaluación eRisk, un taller que explora cuestiones relacionadas con la evaluación de metodologías y aplicaciones prácticas de temas relacionados con la salud y la seguridad para la detección temprana de riesgos en Internet. En ambas evaluaciones, nuestro enfoque supera a todos sus competidores. A través de una sección de análisis exhaustivo, aportamos pruebas de lo que capta cada canal de información, y luego destacamos la importancia y solidez de una visión más holística en las

tareas de clasificación críticas.

En nuestra evaluación, utilizamos conjuntos de datos públicos recientes para tres importantes trastornos mentales: Depresión, Anorexia y Autolesiones. Los resultados obtenidos sugieren que la presencia y la variabilidad de las emociones, el estilo y la información temática capturada por las representaciones propuestas, permiten destacar información importante sobre los usuarios de los medios sociales que sufren estos trastornos mentales. Además, la fusión de estas representaciones puede potenciar el rendimiento.También estas representaciones abren la posibilidad de añadir cierta interpretabilidad a los resultados.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyó otorgado a través de la beca no. 654803. Así como al INAOE por todas las facilidades prestadas durante mi estancia académica.

A mis asesores, Dr. Manuel Montes y Gómez y Dr. Adrián Pastor López Monroy quienes con su conocimiento, experiencia y buen carácter me acompañaron a lo largo de mis estudios y me prepararon para alcanzar esta meta.

A mis sinodales, Dr. Luis Villaseñor Pineda, Dr. Hugo Jair Escalante Balderas, Dr. Aurelio López López, Dr. Saúl Pomares Hernández y Dra. Rada Mihalcea por sus observaciones y comentarios.

A mi familia, amigos y personas cercanas a mí por su apoyo constante e incondicional. Finalmente, pero no menos importante, a mis compañeros del INAOE y todos aquellos que creyeron en mí apoyándome y animándome.

DEDICATORIA

Para mi maravillosa esposa Monica Irasema Delgado Navarro quien me apoyo y alentó para alcanzar esta nueva meta y poder continuar mis sueños. Gracias por creer en mí y motivarme todos los días.

A mi madre quien siempre me ha apoyado y trabajado duro para poder darme lo que necesito en mis estudios y en mi día a día.

A mis amigos que siempre me han alentado a alcanzar mis metas y creer en mí.

A mis asesores que me apoyaron en esta etapa de mi vida, son fuente de admiración, conocimiento y respeto, me han motivado a crecer día con día.

A todos ellos dedico esta tesis, pues son quienes me han dado todo su apoyo incondicional.

CONTENTS

Abstract i

Resumen iii

Agradecimientos v

Dedicatoria vii

I Introduction 1

1 Introduction 3

1.1 Problem Statement 5

1.1.1 Multi-Channel Learning 5

1.2 Hypothesis 6

1.3 Main Objective 7

1.3.1 Specific Objectives 7

1.4 Contributions Resume 7

1.5 Document Outline 8

II Theoretical Background Review 9

2 Background Concepts 11

2.1 Text Representation . . . 11

2.1.1 Bag of Words . . . 12

2.1.2 Word Embeddings . . . 13

2.2 Text Classification . . . 15

2.2.1 Feature Selection . . . 16

2.2.2 Classifiers . . . 16

2.3 Deep Learning . . . 19

2.3.1 Convolutional Neural Networks . . . 20

2.3.2 Recurrent Neural Network . . . 22

2.3.3 Attention Models . . . 26

2.3.4 Transformers . . . 27

2.4 Gated Multimodal Units . . . 30

2.5 Clustering and Affinity Propagation . . . 31

2.6 Evaluation metrics . . . 33

2.6.1 Precision . . . 33

2.6.2 Recall . . . 34

2.6.3 F1 score . . . 34

2.7 Mental Disorders . . . 34

3 Related Work 37

3.1 General Approaches . . . 37

3.2 Ensemble Approaches . . . 39

3.2.1 Discussion . . . 40

3.2.2 Interactions with healthcare practitioners . . . 41

III Contributions **42**

4 Detecting traces of Mental Disorders through Emotional Patterns **44**

4.1 From Sentiments to Emotions . . . 44

4.2 Can emotions have shades? A sub-emotion based representation . . . 45

4.2.1 Generating Sub-emotions . . . 46

4.2.2 Analysis of the novel sub-emotions . . . 48

4.2.3 Converting text to sub-emotions sequences . . . 50

4.3 Using BoSE to identify mental disorders . . . 52

4.3.1 BoSE definition . . . 53

4.3.2 Experiments and results . . . 53

4.4 Learning Sequential Information from Sub-Emotions (Δ-BoSE) . . . 59

4.5 Deep Learning for emotion patterns . . . 62

4.5.1 Adapting Convolutional Neural Network and BoSE . . . 63

4.5.2 Adapting Recurrent Neural Network and BoSE . . . 65

4.5.3 Adapting Attention to the Sub-Emotions . . . 66

5 What they say and how they say it?: Thematic and style representations **72**

5.1 Thematic Embeddings . . . 72

5.2 Proposed Style Embeddings . . . 73

5.3 What does each individual channel capture? . . . 74

5.4 Experimental Settings . . . 75

5.4.1 Pre-processing . . . 75

5.4.2 Classification & Predictions . . . 76

5.4.3 Baselines . . . 77

5.5 Evaluation . . . 77

5.5.1 Completeness and diversity analysis . . . 78

6 Learning a Multi-channel Representation **80**

6.1 A dynamic channel fusion approach 80
6.2 Evaluation 82
6.3 Comparison against the eRisk participants 82
6.4 Analysis of the Method 85
6.4.1 Contribution of each information channel 85
6.4.2 Qualitative analysis of each channel 87
6.4.3 On the predicted posts' probabilities 89
6.5 Multi-Modal BERT (MMBT) Experiments 91
6.5.1 Combining information using multiple BERTs 91
6.5.2 Contribution of each information channel 94

IV General Conclusions **97**

7 Conclusions and Future work **99**
7.1 Academic Production 102

LIST OF FIGURES

1.1 Multi-channel Representation 6

2.1 Example of a BoW Histogram 12
2.2 Word2Vec CBOW and Skip-gram methods 14
2.3 Text Classification General Process 15
2.4 SVM hyperplane example 17
2.5 Artificial Neural Network architecture 19
2.6 Example of a node with 3 inputs, x1, x2 and x3. 20
2.7 Example CNN architecture 22
2.8 A simple example of an RNN unit 23
2.9 General Diagram of the GRU cell 24
2.10 General Diagram of the LSTM network 26
2.11 Attention over the context in the sequence 27
2.12 Transformer Model Architecture (Vaswani et al. 2017) 28
2.13 BERT input representation (Devlin et al. 2019) 29
2.14 Multimodal bitransformer architecture (Kiela et al. 2019). 30
2.15 Overview of GMU module 31
2.16 Clustering algorithms 32

4.1 Distribution of emotions at the depression task 45
4.2 Distribution of emotions at the anorexia task 45
4.3 Distribution of emotions at the self-harm task 46
4.4 Obtain word representation for EmoLEX. 47
4.5 Obtained sub-emotions 48
4.6 Examples of words grouped in different sub-emotions 50
4.7 Determination of the the vocabulary words and sub-emotions vectors. . . . 50
4.8 Replace each word in the vocabulary with its closest sub-emotion. 51
4.9 Masking the users' documents, replacing word sequences for sequences of sub-emotions. 52
4.10 Construction of the Δ-BoSE representation 60
4.11 Comparison of the emotional signals between control and mental-disorder groups 62
4.12 Diagram of the Convolutional Neural Network Model 64
4.13 Diagram of the Recurrent Neural Network Model 66
4.14 Diagram of our bi-GRU model with Attention 68
4.15 Examples of weighted sequences of sub-emotions with different contexts . 71

5.1 Diagram of the generation of style embeddings 73
5.2 Similarities of several word pairs using the style and the original embeddings. 75

6.1 Diagram of the multi-channel representation 81
6.2 Boxplot of the F1 scores 84
6.3 Average proportion of GMU unit activations for the channels over the test set 85
6.4 Saliency obtained with the different type of channels for the positive class . 87
6.5 n of the posts' predictions 89
6.6 Output of the different posts of being a positive class 90
6.7 General diagram of the Multimodal BERT with vectors of three channels . . 91
6.8 General diagram of the BERT-CNN model 92

6.9 General diagram of the BERT-3CNN model 93
6.10 General diagram of the BERT-GMU model 93
6.11 Average z_i value for the three mental disorders over the test set instances. . 95

LIST OF TABLES

4.1 Size of the vocabulary for each emotion presented in the lexical resources . 49

4.2 Data sets used for experimentation 54

4.3 F1 results over the positive class in three eRisk's tasks. 56

4.4 Examples of relevant sub-emotions for depression detection 57

4.5 Examples of relevant sub-emotions for anorexia detection 58

4.6 Examples of relevant sub-emotions for self-harm detection 58

4.7 F_1, *Precision* and *Recall* results over the positive class. 61

4.8 Depression data set 2020 62

4.9 CNN results 65

4.10 RNN results 67

4.11 Attention results 69

4.12 Mean and variance value for the attention weights in the sub-emotions. . . . 69

4.13 Significance results 70

5.1 Examples of the closest words for the channels 76

5.2 F1 results over the positive class in three eRisk's tasks 78

5.3 MPF and CFD results in the three tasks, measured over the positive class . . 79

6.1 F1 results over the positive class in three eRisk's tasks 83

6.2 F_1, $Precision$ and $Recall$ results over the positive class 84

6.3 Posts with highest z_i value for each mental disorder 86

6.4 Words with the highest saliency for each task and each channel 88

6.5 F1 results over the positive class in three eRisk's tasks 94

6.6 Posts with highest z_i value for each channel over the depression task. 96

Part I

Introduction

CHAPTER 1

INTRODUCTION

A mental disorder is a disease that causes different disturbances in the thinking and behavior of the affected person. These interferences could vary from mild to severe and result in an inability to live ordinary demands or routines in daily life. The mental disorder may be related to a particular event that generated excessive stress on the affected person or a series of different stressful events (World Health Organization 2019). For example, some of the causes that affect people are environmental stress, genetic factors, or different difficult life situations.

Common mental disorders such as depression, anorexia, dementia, post-traumatic stress disorder (PTSD), or schizophrenia affect millions of people around the world (Kessler et al. 2017). Most people believe that mental disorders are uncommon or only happen to people with specific personal profiles, when in fact, they are prevalent and very familiar (Mathers and Loncar 2006). Many families think they are not prepared to face the fact that some loved one has a mental problem. The idea of having a mental disorder causes emotional and physical damage that could make people feel fear for the idea of being vulnerable to criticism, judgment, or wrong opinions.

The National Institute of Mental Health made a study where they found that young people are more affected by mental disorders (Merikangas et al. 2010). This study explains that one in every five young people is affected by at least one of them. The authors also discovered that the percentage of people suffering from a mental disorder is higher than other frequent primary physical conditions, such as diabetes or asthma. A different study made by the Canadian Association of College and University Student Services (CACUSS) found that the number of students reporting being in anguish is increasing in comparison with

previous years (Group 2016). The authors also observe that one in every five students suffer from depression, feel anxious, or are dealing with other mental disorder. The students also said that their health was poor, and 13% had considered suicide at least once. These findings present an alarming rise in mental disorders, and the numbers of suicide are increasing.

As we previously mentioned, mental disorders affect people worldwide, and our country is not an exception. In 2018 a study of mental disorders in Mexico reveals that 17% of people in the country have at least one mental disorder and one in four will suffer a mental disorder at least once in their life (Renteria-Rodriguez 2018). Nowadays, of the people that are affected, just one in every five get treatment. Mental disorders increase in countries that have gone through events of generalized violence or natural disasters (Renteria-Rodriguez 2018).

Nowadays, we live in a developed world, and for many people, their social life does not always occur in their surroundings or immediate environment. In many cases, it takes place in a virtual world created by social media platforms like Facebook, Twitter, or Reddit. In other words social media became a vital link for people that live far from their loved ones, such as family members and friends. This reality presents great opportunities which, if properly addressed, could contribute to the understanding of *what* and *how* we communicate. In this regard, the goal of this study is to analyze their social media documents [1], via the automatic identification of emotional patterns, writing style, and thematic content to detect the presence of signs of mental disorders (Chikersal et al. 2020; Guntuku et al. 2017; Pestian et al. 2010). For example, with the emotional information, we can capture expressions that reveal symptoms of people that have some psychological distress. With the style information, the usage of passive voice, questions, and personal expressions used by users with mental disorders. Finally, with the thematic information, we can understand the context of the surrounding words, phrases, and objects.

Many typical analyses run on the information shared by users and only considers the thematic aspect of the content, simply ignoring important patterns that may be beyond the topics. Thus, this work hypothesizes that there are other dimensions of communication that provide insightful information to characterize users beyond topics. For example, the writing style or even the emotions conveyed in the text. Accordingly, the goal of this study is to present a novel approach that exploits all these different views and obtains a more holistic representation of the user, which we name as multi-channel representation. For this purpose, we define a *channel* as a different property or view from the same modality (Qianli et al. 2017). In this work, we use the text modality and three channels that will separately focus on different aspects of the user. The first one is the thematic information, the second one is the expressed emotions, and the third one is the author's writing style. The intuition of

[1] In this work, we refer as "document" to the concatenation of the posts of each user.

our approach is that people that present some mental disorder tend to express differently, at different dimensions, regarding the control group. For example, they tend to repeatedly bring up topics related to prior traumas or even sentimental relationships, but also at the same time to communicate particular emotions such as anger and disgust. For this thesis, we study how all these different communication aspects can be captured in different channels and combined to offer a more complete view of the user. Providing evidence that although each channel is different they complement each other. Interestingly, all these components in some way are related to how we humans analyze not only the content of a message but also the manner how it is expressed.

1.1 Problem Statement

Previous studies focused on the detection of mental disorders like depression, anorexia, or self-harm suggest that these symptoms are detectable in online environments. Most of the works focus on the usage of dictionaries related to mental disorders, sentiment analysis looking for the polarity of the post, or counting the frequency of the words. The performance of these approaches is still modest, suggesting the challenge of the problem. This performance presents an opportunity for the exploration of new techniques to extract different types of information from the user's posts, and then using a model that learns to automatically combine them and create a better representation. With this new representation, we can improve the detection of different mental disorders and provided some insights with the interpretability of our methods.

1.1.1 Multi-Channel Learning

In the real world, the information usually is presented in different modalities that help to learn a new combined representation (Duong et al. 2017). For example, videos that contain audio, images, and text (subtitles). Inspired in the multi-modal learning, we proposed a multi-channel learning for the text modality. Our method, yields that the complementarity in the types of information is important to get a better picture and understanding of the posts written by the users.

For this thesis, as we previously defined, *channel* is a different property or view from the same modality. For example, in (Qianli et al. 2017) they divided the 3D skeleton sequences into different channels and then learned to combine the information of the channels. For our work, we used the text modality present in the documents of the users. Some examples of channels that we used are the thematic aspects, the emotions presented, and the writing style of the author. Multi-channel Learning creates a representation that combines two or more of

these channels, discovering the relationship between different channels.

Figure 1.1 shows the process of extracting the different types of information (channels) from the documents, and a model that learns how to automatically combine the channels in a single representation.

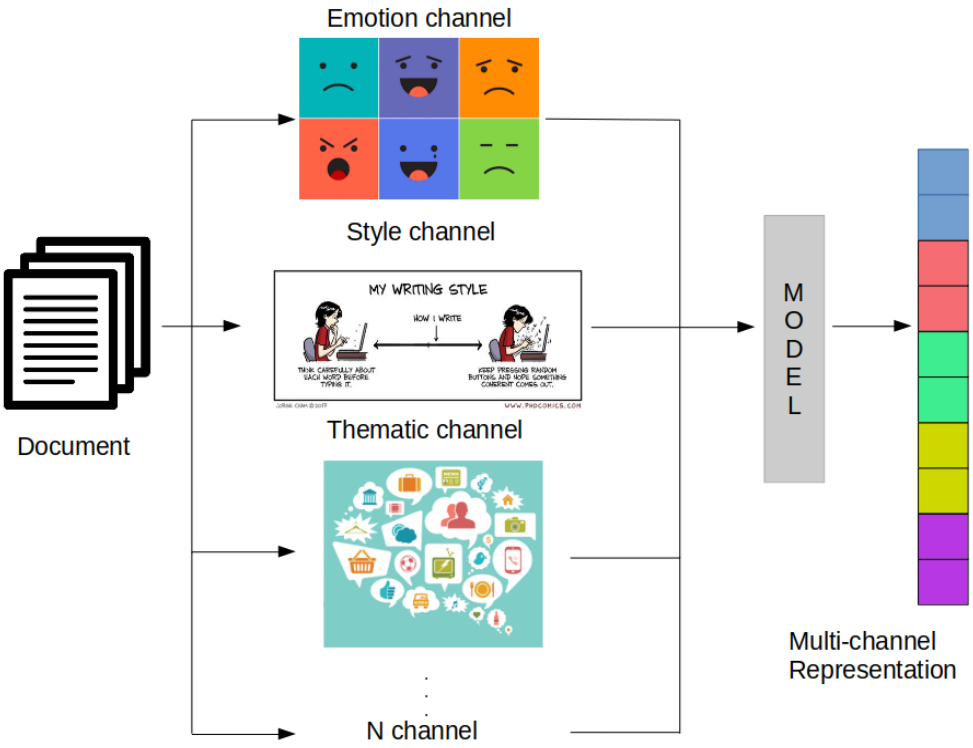

Figure 1.1: Multi-channel Representation, extracting different types of information from the same modality and then a model that automatically learns how to combine it.

1.2 Hypothesis

We stated the hypothesis for this research as follows:

People that present a mental disorder tend to express differently than healthy people. For example, their topics of interest, writing style, relation with others, and even their activity hours had different behavior. The hypothesis is that learning to combine different types of information gives a broader view that could help in the detection of signs of mental disorders and obtains better classification results than using only one type of information.

1.3 Main Objective

Design a method adapting traditional and deep learning NLP techniques to automatically learn a multi-channel representation using the information generated by the users in social media platforms. Then, use this representation for the detection of mental disorders and improve the results obtained by traditional and state-of-the-art approaches.

1.3.1 Specific Objectives

1. Design methods that learn new representations of the different channels in the post history of the users. For example, the context, the style of the author, and the emotions used. With this information, we can improve the representation of the users for the detection of mental disorders.

2. Adapt a model that automatically combines the different information channels and focuses on the critical parts of the data for the detection creating a new representation.

3. Develop a method to incorporate the importance of temporal information presented in the sequences of the posts.

4. Evaluate the utility of our proposed method in different tasks related to mental disorders.

1.4 Contributions Resume

The main contributions of this thesis are the following points:

1. A new representation called Bag of Sub-Emotions (BoSE). Which represents social media documents by a set of fine-grained emotions automatically generated using a lexical resource of emotions and sub-word embeddings.

2. A representation based on sub-emotions that allow capturing the emotional variability of social media users over time.

3. A new representation based on the writing style of the users that allow capturing their writing variability.

4. An approach to combine three different channels: thematic, emotion, and style; which improve the detection.

5. A detailed study of these channels and the importance of combining different types of information. By this characterization, we aim to provide evidence of its robustness.

1.5 Document Outline

In this thesis, our main objective is to focus on the detection of mental disorders in users from social media platforms. The work will focus on the detection of these users, using a multi-channel representation that exploits traditional natural language process methods combined with deep learning methods. For example, extracting from different channels features like semantics, emotions, or writing style. Then, feeding a deep neural network that automatically learns how to combine these features and extract the most relevant information from them. We organized the remaining thesis document as follows:

- In Part II, two chapters present the relevant background elements, to make this thesis as self-contained as possible. This part contains the following chapters:
  - In chapter 2, we describe some of the most relevant concepts in text classification for this research.
  - In chapter 3, we present some of the most relevant works in the literature exploiting NLP approaches to model users with mental disorders.
- Part III organizes the main contributions of this thesis, which are related to the different channels of information.
  - In chapter 4, we present the proposed approaches to capture traces of mental disorders through emotional patterns.
  - In chapter 5, we describe the proposals to capture the thematic and style information.
  - In chapter 6, we explain our fusion strategy and analysis of the combinations.
- Part IV has chapter 7, which outlines the main conclusions and future work of this research.

Part II

Theoretical Background Review

CHAPTER 2

BACKGROUND CONCEPTS

This section presents an overview of the different techniques and core concepts needed for the thesis to be as self-contained as possible. We divided this section as follows: first, a description of the methods for text representations like a bag of words and word embeddings. Then, some of the main ideas for text classification and deep learning, with the description of some relevant neural networks useful for our task.

2.1 Text Representation

Text representation is one of the main problems in text mining, text classification, and information retrieval. One crucial problem to solve is to numerically represent text documents and make them mathematically computable. For example, if we consider a set of unstructured documents, then the objective is to represent each document as a point in a numerical space, where the distance or similarity between each document is well defined.

The performance of any machine learning method is mostly dependent on the choice to represent the data, also known as features. For this reason, researchers applied a lot of effort in the development of preprocessing and data transformation that helps in creating a representation that can support the machine learning methods (Y. Bengio et al. 2014).

Learning representations of the data makes it easier to extract useful information on prediction tasks. In deep learning, representation learning is formed by the combination of multiple non-linear transformations of the data. In the past years, different text representations techniques were proposed, and in the following subsections, we described some of them.

2.1.1 Bag of Words

The bag of words (BoW) is the most simple and well-known technique for text representation and classification, where we described the text by the occurrence of words within a document. Firstly, we create a vocabulary $\mathbf{w}$ from the training data. Then, the measure the presence of the words by their frequency (Goldberg 2017). This representation creates a histogram $\mathbf{d} = [w_1, w_2, \cdots, w_N]$ where $\mathbf{w}$ is the vector that contains w_N words, and ignores the structure of the words, accounting only the occurrence of the words in the document and not the position or order in it. Figure 2.1 presents an example of a a BoW Histogram Vector.

BoW model is a technique of extracting features from the text for a model to use as a representation, like in other machine learning algorithms. This technique is simple and flexible, and we can use it to extract features from documents. The intuition behind this representation is that documents or texts present similar content if they are of the same type.

John likes to watch movies. Mary likes movies too.									
1	2	1	1	2	1	1	0	0	0
John	likes	to	watch	movies	Mary	too	also	football	games

Figure 2.1: Example of a BoW Histogram Vector for the text: "John likes to watch movies. Mary likes movies too"

Weighting schemes

In this subsection, we will briefly examine aspects related to how to weigh them, i.e. how to calculate the w_i values associated with each term, in the vector representing the document $\mathbf{d} = [w_1, w_2, \cdots, w_N]$. The most widely used normalization method is tf-idf, which stands for "term frequency-inverse document frequency" (Salton and Buckley 1988), which, given a document d_i and a term t_j, will calculate its weight $w_{i,j}$ as the product of of two special weights, $tf_{i,j}$ and idf_j as follows:

$$w_{i,j} = tf_{i,j} idf_j \tag{2.1}$$

Where $tf_{i,j}$ is the normalized frequency of term t_j in document d_i and where idf_j is the logarithm of the inverse of the number of documents in which term t_j appears. This method of weighing usually performs consistently well, without depending on the task at hand.

2.1.2 Word Embeddings

In Natural Language Processing a word embedding is a distributed representation of text in an n-dimensional space. Word embeddings is a technique for modeling language, where words presented in the vocabulary are transformed into vectors of continuous real numbers. For example, consider the word "disorder" it would become a vector of size $N \rightarrow$ [0.22,0.15,0.44,..,N]. The main idea is to create a low-dimensional dense vector space where the embedding vector represents the linguistic relationship of the word with the context. Thus, two words that are related have similar vectors.

Word embeddings are a form of word representation that helps a machine to understand the language and the context. Word embeddings represent relationships between words and useful contextual information that benefit when training models on the data. These representations are traditionally used for solving most NLP problems and are sometimes adjusted during the training phase.

There are different techniques to obtain the word embeddings, for example, using neural networks (R. Bengio Y. a. D. et al. 2003; Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013; Mnih and Kavukcuoglu 2013; Morin and Y. Bengio 2005) or matrix factorization (Huang et al. 2012; Pennington and Socher 2014). Nowadays, the most popular embeddings are word2vec, GloVe and FastText, which we describe in the following subsections.

Word2Vec

One of the most used word embeddings is Word2Vec. Where a neural network predicts the target word from the context, for example, word(w) = "playing" and the context = "the musician is w the guitar". The w is the target word that the network model learns. This model is named Continuous Bag-of-Words (CBOW) (Mikolov, Chen, et al. 2013). The other model is named Skip-Gram (SG) (Mikolov, Sutskever, et al. 2013), where the model does the inverse prediction, first learns the word, and then predicts the context of the word. finally, the word is represented by the internal weights of the network. The purpose of the CBOW model is to smooth the big distributional information using the context as an observation. While the SG model uses the context as targets and normally performs better for larger datasets. Figure 2.2 represents these two different methods in the Word2Vec algorithm.

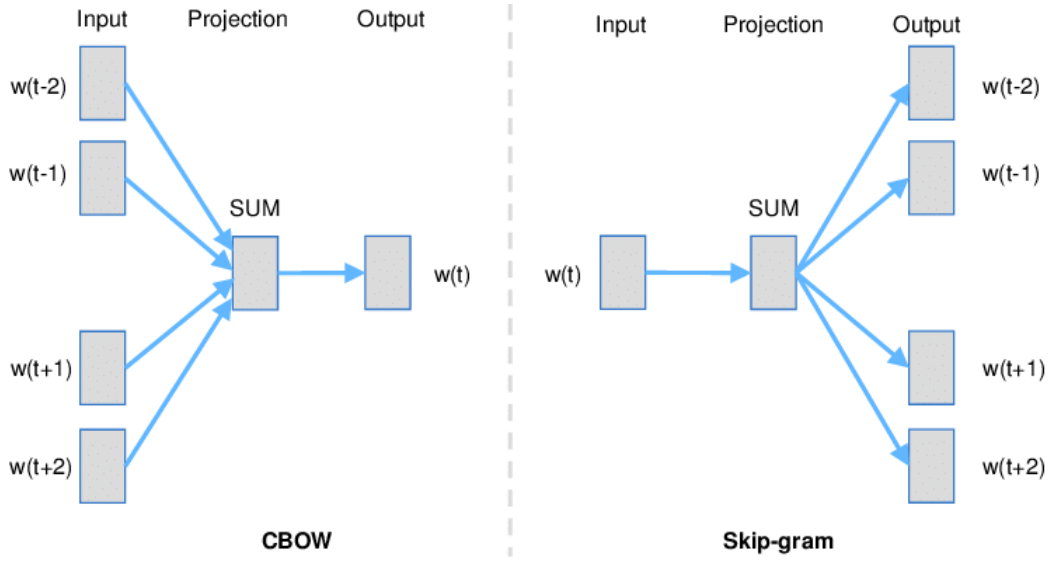

Figure 2.2: Word2Vec CBOW and Skip-gram methods. The first learns the word, and then predicts the context of the word, while SG uses the context as targets.

GloVe

Another traditional approach is Glove (Pennington and Socher 2014), which are embeddings trained using nonzero entries of a global word-to-word co-occurrence matrix. GloVe model is composed of two approaches: a matrix factorization and shallow windows. The main idea of GloVe is to learn word representations of words in a local context. First, it creates a co-occurrence matrix X which contains the co-occurrences of words in the dataset. Thus, a word j appears in the context of a word i if the distance of j is within the window w_s of i. $X \in N^{V \times V}$, where V is the size of the vocabulary and x_{ij} is the number of occurrences of word j in the context of word i. The word vectors are learned according to the co-occurrence matrix X.

This model was pre-trained with high-dimensional corpora from Twitter, Common Crawl and Wikipedia. Unlike Word2Vec, GloVe presents a lower dimensionality in their word vectors, ranging between 50 and 300 dimensions. GloVe pre-trained with information from Twitter performs well in comparison with other models in tasks of classifying texts from social media networks (Pennington and Socher 2014). For this reason, GloVe is a good option due to the nature of the data used in this work.

FastText

FastText embeddings are an extension of Word2Vec proposed by (Bojanowski et al. 2016). The main difference is the way words enter the neural network. In comparison with Word2Vec, FastText splits words into n-character frames named sub-words. They divided the word into these sub-words and represent it as the sum of each char n-gram. After the training, we obtain for the different sub-word a word embedding to depict it and the final vector for each word corresponds to the sum of each sub-word embedding. With this strategy, the embeddings cover a higher number of words and can represent rare words even outside the vocabulary. In addition to these features, the FastText model has pre-trained models in 157 languages.

2.2 Text Classification

Text Classification (TC) is the process of assigning categories or tags to a text or a document according to its content. TC is used to structure and categorize, for example, topics, conversations, and languages. Text Classification has broad applications such as intent detection, information filtering, and sentiment analysis (Aggarwal and Zhai 2012).

Text classification can work in two different ways: i) manual, where a human annotator reviews the text and categorize it accordingly to how interprets the content. ii) automatic, that applies machine learning to classify text faster and with less cost, for example, rule-based systems that organize in groups using a set of linguistic rules (Sasikumar et al. 2007).

Text Classification has become an important part of business as it allows to get insights from the data and automate analysis for different processes. Figure 2.3 described a general process for Text Classification; the model receives an input text and returns a label as an output.

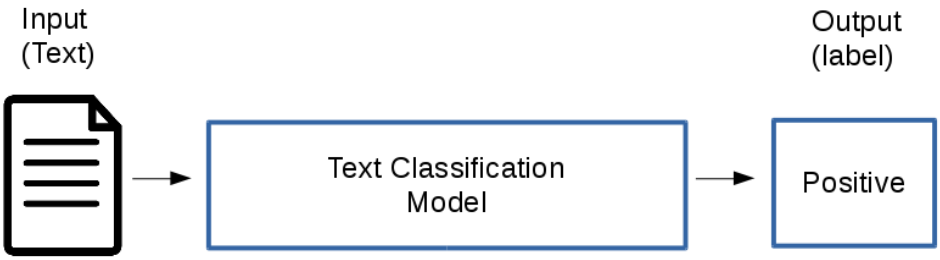

Figure 2.3: Text Classification General Process

There is a wide variety of machine learning models that can be used as classifiers, in the following sub-sections we briefly describe the ones we focused on.

2.2.1 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model, in this sub-section we describe the one used for this thesis.

chi^2 distribution

The chi^2 distribution is a random variable x that can be written as a sum of squared standard normal variables:

$$\chi^2 = \sum Z_i^2 \tag{2.2}$$

where Z_i is a set of standard normal variables. The chi^2 test is used in statistics to obtain the independence of two events. More specifically in our case, we use it to measure if a specific n-gram and the occurrence of a specific class are independent. We can express this formally, given a corpus C, we observed the count and expected count of a term estimating the quantity for each and rank them by their score:

$$\chi^2(C, t, l) = \sum_{e_t \in 0,1} \sum_{e_l \in 0,1} \frac{O_{e_t e_l} - E_{e_t e_l}}{E_{e_t e_l}} \tag{2.3}$$

Where:

- O is the observed count and E the expected count.
- e_t determines if the document contains the n-gram t.
- e_c checks if the document is in the observing class l.

2.2.2 Classifiers

Support Vector Machine

Support vector machine (SVM) is a model that represents the sample points in space, separating the classes into two wider spaces using a hyperplane defined as the vector between the points of the classes (Cortes and Vapnik 1995). Once the separation is done, the samples are arranged according to the spaces they belong to and can be classified more easily. SVM has very efficient training, and it is a robust method for generalization, where its search space has only a global minimum.

For linear classification we have a training set of n points with the form ($\overrightarrow{x_1}$,y_1),...,($\overrightarrow{x_n}$,y_n), where y_i is either 1 or -1, indicating the class to which $\overrightarrow{x_i}$ belongs. Each $\overrightarrow{x_i}$ is a p-dimensional vector. One wants to find the maximum margin hyperplane that divides the set of $\overrightarrow{x_i}$ for each y_i = 1 from the set of y_i = -1. This is defined so that the distance between the hyperplane and the nearest point of the set of *xi* of each set is maximized. Any hyperplane can be written as the set of points *x* satisfying $\overrightarrow{w} \cdot \overrightarrow{x}$ + *b* = *0*, where $\overrightarrow{w}$ is the normal vector of the hyperplane. Figure 2.4 shows this process.

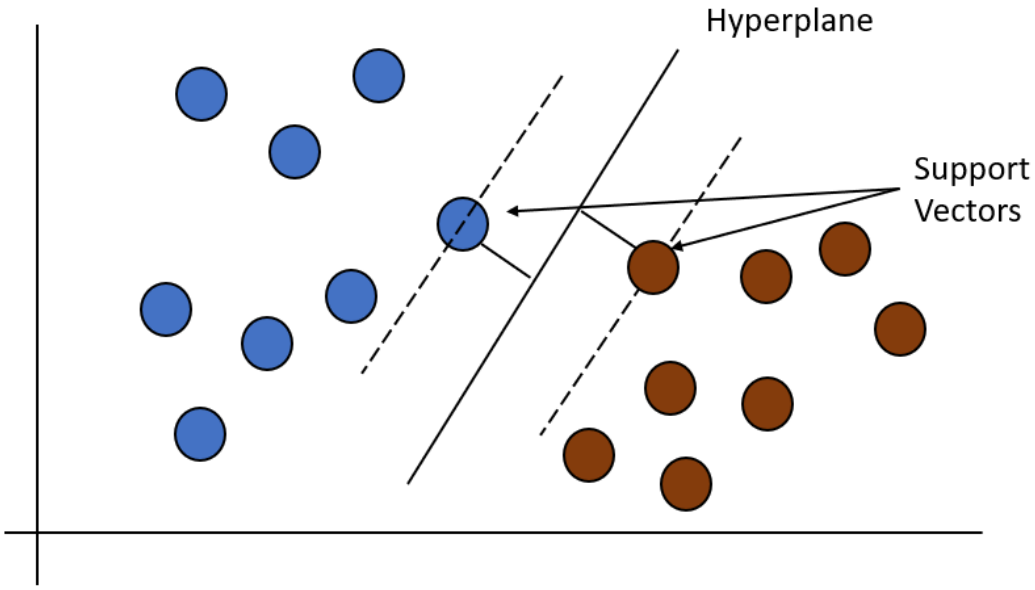

Figure 2.4: A two-dimensional space with one-dimensional hyperplane that divides the data into different classes. The dotted lines represent the margin between the hyperplane and the support vectors.

When we have a nonlinear classification it uses the *"kernel trick"* for the hyperplane. The algorithm is similar, except that each dot product is replaced by a nonlinear kernel function. This allows the algorithm to fit the hyperplane into a transformed feature space. The transformation is nonlinear, and the transformed space is of high dimensionality. Working in feature space with high dimensionality increases the generalization of the support vector machine error, although providing enough examples the algorithm performs well.

Common kernels:
Polinomial (homogeneous):

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} \cdot \overrightarrow{x_j})^d \tag{2.4}$$

Polinomial (non-homogeneous):

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} \cdot \overrightarrow{x_j} + 1)^d \tag{2.5}$$

Gaussian Radial Basis Function:

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = exp(-\gamma \| \overrightarrow{x_i} - \overrightarrow{x_j} \|^2) \tag{2.6}$$

Hyperbolic Tangent:

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = tanh(\overrightarrow{kx_i} \cdot \overrightarrow{x_j} + c) \tag{2.7}$$

Neural Networks

Artificial Neural Networks (ANN) are computational models inspired by the human brain. Artificial neural networks have their origin in the McCulloch-Pitts neuron (McCulloch and Pitts 1943). A simplified model of the human neuron as a computational element described in terms of propositional logic. An ANN is a network of small computational units referred to as "neurons" grouped at different levels named layers (Basheer and Hajmeer 2000; Goodfellow et al. 2016). As illustrated in Figure 2.5, neural networks have three types of layers: input layer, hidden layer, and output layer. The input layer is composed of all the nodes that receive the input values and whose function is to propagate them to the internal nodes. The output layer contains all the nodes that generate the final value of the network. Lastly, we have the hidden layers that perform nonlinear transformations of the inputs entered into the network.

Each node in a neural network takes a vector of input values, $< x0, x1, ..., xm >$, and produces a single output value y. Then, apply a function f to the dot product between the input vector and a vector of weights $< w0, w1, ..., wm >$. Figure 2.6 illustrates this process. We can express y value formally as:

$$y = f(b + \sum_{i=0}^{m} w_i x_i) \tag{2.8}$$

Where f is called "activation function", and its role is to determine whether the result of the dot product is sufficient for that node to "activate" or not. Three of the most commonly used activations are the softmax function, tanh, and ReLu (Ramachandran et al. 2017).

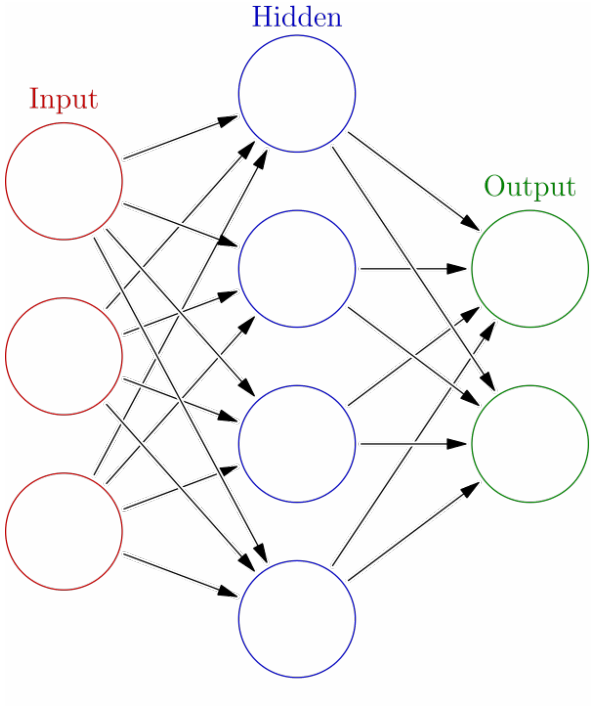

Figure 2.5: Artificial Neural Network architecture. The network is composed of an input layer with input nodes, a hidden layer, and an output layer.

The training of an ANN consists in finding the appropriate values for the weights associated with all the nodes in the network. There are different methods for finding these values, with backpropagation (Chauvin and Rumelhart 1995) being the most widely used algorithm.

Nowadays, with the increase in computational power and the massive data available, neural networks become deeper. With a large number of hidden layers and thousands of nodes. This increase in power leads to a proliferation of different deep network architectures referred to as deep learning architectures (Goodfellow et al. 2016) explained in the following subsections.

2.3 Deep Learning

Deep learning is a group of methods to learn representations that are known as deep architectures (Y. Bengio 2009). These methods consist of multiple layers of nonlinear units that process the data for feature extraction and transformation. The first layers are closer to the input data and learn simple features. The following layers learn sophisticated features extracted from the first layers. These architectures are known as hierarchical representations and can learn without the need of an expert in feature extraction and selection from the original data.

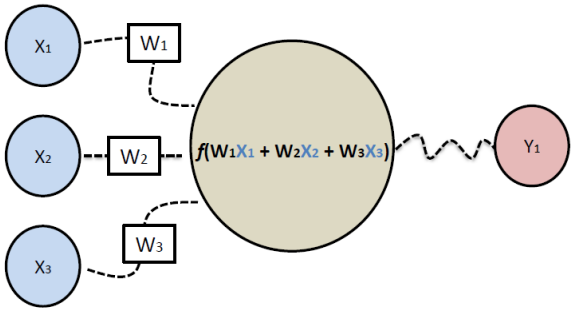

Figure 2.6: Example of a node with 3 inputs, x1, x2 and x3.

Conventional machine learning techniques are limited to processing data in raw form. These techniques required the construction of a pattern recognition system with considerable domain expertise to design a good feature extractor that converts the raw data into a fitting representation for the classification task. Deep learning allows to be fed with the raw data and automatically discover the representation for detection or classification (LeCun et al. 2015). Using the higher layers to amplify relevant aspects of the input data for discrimination between irrelevant information and important variations. The important characteristic of deep learning is that the layers of features are learned from the data using a general learning process, instead of the designed by human experts in the domain.

In the past years, deep learning obtained a state-of-the-art result in many domains; for example, they started in computer vision, speech recognition, and more recently in natural language processing. In the following subsections, we explain the deep learning architectures used for this thesis.

2.3.1 Convolutional Neural Networks

A convolutional neural network (CNN) is a deep learning algorithm that takes a matrix input, for example, an image, and assigns importance to various aspects. We expressed this importance through the weights and biases of the network and make a differentiation between the objects in the image. A CNN can capture the spatial and temporal dependencies in an image using different filters. These filters reduce the number of parameters in the image and perform a better fitting. In previous methods, the filters were hand-engineered, while CNN learned these filters.

How a convolution works?

Think about an input image as a matrix, and each value in the matrix is, for example, a pixel with a value between 0 and 255 that represents the brightness intensity. To understand the convolution operation imagine placing a filter or kernel on the top of the image. Then, multiply the values of the image matrix with the corresponding value in the convolution filter. More formally as:

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k] \tag{2.9}$$

Where f is the input image and h our kernel and the indexes of rows and columns of the result matrix are marked with m and n respectively. Then, add all of the multiplied values together obtaining a single scalar, and is placed in the corresponding position of a result matrix. The kernel moves x pixels to the right, where x is the length of the kernel and is a parameter of the convolutional network. The multiplication is repeated until all the entries in the input image have been covered. This process is called the convoluted feature or input feature map. We can apply multiple convolution kernels at once over an input image and create one output for each kernel.

Pooling

The next step in a convolutional network is the pooling or downsampling layer. This layer consists in applying an operation over regions in the input feature map and extracting representative values for each one. This process is similar to the convolution operation, but instead of transforming local regions via linear transformation (convolution filter), are transformed through a hardcoded operation. The two most common pooling operations are the max-pooling and the average-pooling.

The max-pooling operation selects the maximum of the values in the input feature map of each region. The average-pooling operation obtains the average value of each feature map in the region. The output of this process is a single scalar that results in a significant size reduction in the output size.

The objective of the convolution network is to extract low and high-level features from the input image. To accomplish this, CNN uses multiple Convolutional layers. The first layers capture the low-level features such as edges or color. The following layers adapt this information to high-level features, giving us a better understanding of the images. Figure 2.7 presents an example of different convolutions over an image.

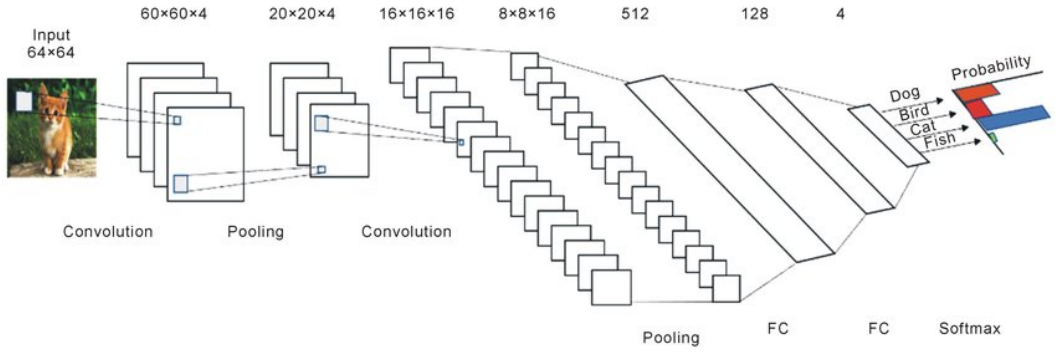

Figure 2.7: Example CNN architecture (Williams and R. Li 2018), with different convolution and pooling layers

2.3.2 Recurrent Neural Network

Recurrent Neural Networks (RNN) are well known by the feedback loop connected to their past decision. RNNs process an input sequence element by element, preserving information about the past elements of the sequence. Due to this process, it is often said that RNNs have memory and captures information in the sequence itself.

RNNs have a purpose to preserve in the hidden state of the network the sequential information and affect the processing of each new example. Then, find correlations between events separated for different moments. RNNs are dynamic systems, but they present a problem maintaining the relation of long sequences because the backpropagated gradient shrink at each time step and after many steps vanish (LeCun et al. 2015).

Just as human memory travels in a sequence way through our brain, affecting the behavior without using the full information. The information that travels in the hidden states of the recurrent nets affects the decisions without revealing all learned. The process of preserving memory in these networks are represented by $h_t = \phi(Wx_t + Uh_{t-1})$, where the hidden state at time step t is h_t. In this function, the input at the same step x_t (the word) is modified by a weight matrix W. Then, added to a hidden state of the previous time step that is represented by h_{t-1} multiplied by the hidden state in the previous time in matrix U. The weights contained in the matrices determine how much importance to grant to the present input and past hidden state. Lastly, the sum of the weights is flattened using a function ϕ, making gradients workable for backpropagation. Figure 2.8 presents a simple example of an RNN unit.

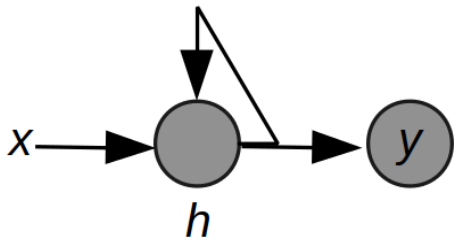

Figure 2.8: A simple example of an RNN unit. Where x is the word, h the hidden state and y the output value.

Gated Recurrent Unit

Gated Recurrent Unit (GRU) is a neural network that also aims to solve the problem of the vanishing gradient that is present in recurrent neural networks. GRU can also be considered as a variation of the LSTM, both have a similar design and produce equal results in some cases (Cho et al. 2014).

GRUs use two gates name update gate and the reset gate. These gates are two vectors that decide what information and the amount of information should be passed to the output. The update gate helps to determine how much of the past information needs to be passed along to the future. Using the information of the actual state multiplied by the weight in the same time step and then added to the multiplication of the previous information and weight. The result is passed to a sigmoid activation function that squashes the result between zero and one. The reset gate is used to decide how much of the previous information to forget. The operation to calculate the gate is the same as the update gate, the difference comes in the weights and the sigma function that is changed for a tanh function.

Step-by-Step GRU

- Update gate: calculate the update gate z_t for time step t. x_t and h_{t-1} are multiplied by its weight W^z and U^z. Both results are added together and a sigmoid function is applied to obtain a result between 0 and 1.

$$z_t = \sigma(W^z x_t + U^z h_{t-1}) \tag{2.10}$$

- Reset gate: decide how much of the past information to forget. Similar to the update gate with the difference in the weights and the gate's usage.

$$r_t = \sigma(W^r x_t + U^r h_{t-1}) \tag{2.11}$$

- Current memory content: multiply x_t with a weight W and h_{t-1} with a weight U. Calculate the product between the reset gate r_t and Uh_{t-1}, sum them up and apply a

nonlinear activation function tanh. This calculation determines what to remove from the previous time steps.

$$\acute{h}_t = tanh(Wx_t + r_t \odot Uh_{t-1}) \tag{2.12}$$

- Memory at current time step: apply element-wise multiplication to the update gate z_t and h_{t-1}. Then, a multiplication to $1 - z_t$ and $\acute{h}_t$. With this process, the network calculates h_t, which holds information for the current unit and passes it down to the network.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \acute{h}_t \tag{2.13}$$

GRUs can save and eliminate information using their gates, helping to eliminate the problem with the vanishing gradient keeping the relevant information that passes to the next step. Figure 2.9 presents a general diagram of the GRU cell unit.

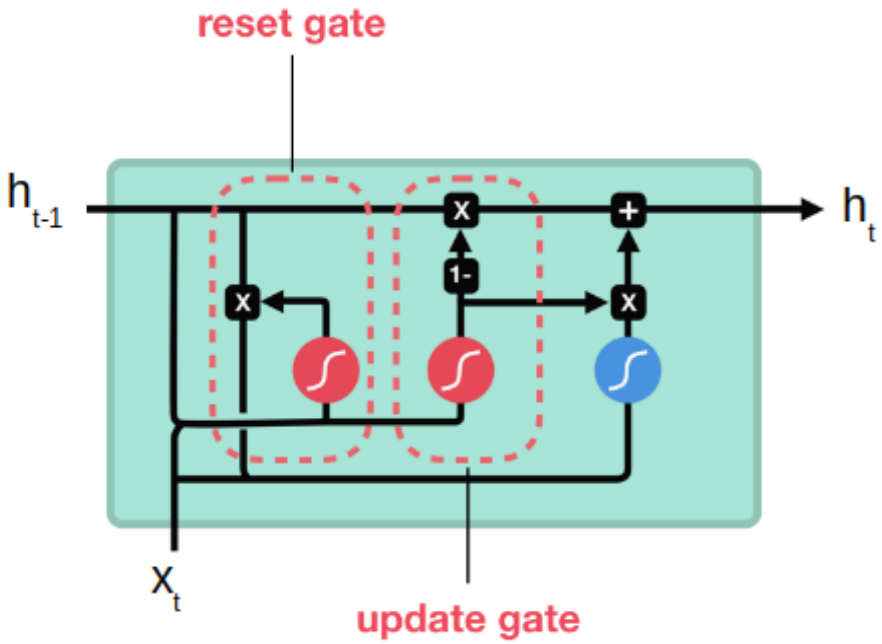

Figure 2.9: General Diagram of the GRU cell unit with the reset and update gates (Phi 2018)

Long Short Term Memory

Recurrent Neural Networks suffer from learning to store information for very long sequences. To solve this problem, researchers proposed the use of explicit memory. A long short-term memory (LSTM) network that uses special hidden units and learns to remember inputs of the sequence for a long time. These hidden units are called memory cells, gated neurons that

leak information through time. Each memory cell has a connection to itself at the next time step, where it copies the value of the current state and accumulates the new values. The cell has a multiplicative gate by another memory cell that learns to decide to clear or keep the content of the memory (LeCun et al. 2015).

The core idea behind LSTMs is to remove or add information to the cell state using gates decided to let information through. These gates are composed out of a sigmoid neural layer and a pointwise multiplication operation. The sigmoid layer outputs a number between zero and one that describes how much of each component should pass. A value closer to one means more information to let pass.

Step-by-Step LSTM

- First step: decide what information we are going to remove from the state of the "forgetting door" cell.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{2.14}$$

- Second step: decide what new information we are going to store. Sigmoid or "gateway layer" decides which values we will update. Tanh creates a vector of new candidate values, which could be added to the state.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2.15}$$

$$\tilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_C) \tag{2.16}$$

- Third step: update the state of the previous cell, C_{t-1}, to the new state of cell C_t.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.17}$$

- Fourth step: Finally, we must decide what output we will produce.

$$O_t = \sigma(W_o[h_{t-1}, x_t]b_0) \tag{2.18}$$

$$h_t = O_t * tanh(C_t) \tag{2.19}$$

LSTM networks have proved to be more effective than conventional RNNs, especially when the sequences are very long and the networks have several layers for each time step. Figure 2.10 presents the general structure of a LSTM.

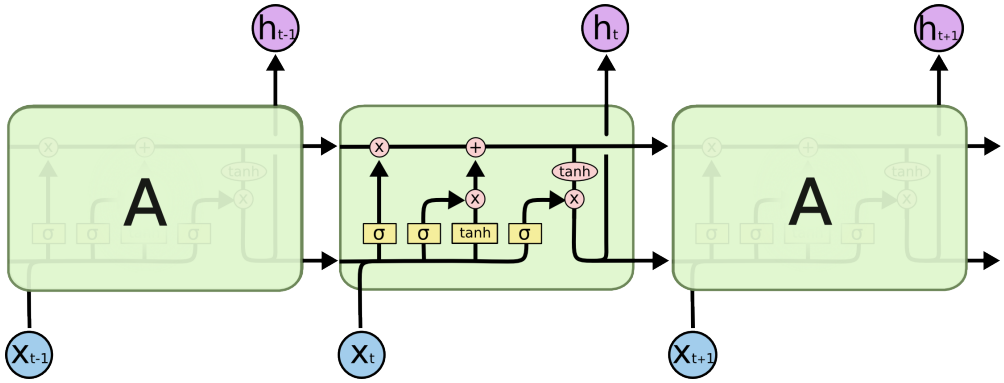

Figure 2.10: General Diagram of the LSTM network with the different gates (Olah 2015).

2.3.3 Attention Models

Attention models are networks similar to short-term memory, but these models can allocate attention over longer periods of time. The attention mechanisms are parts of the networks that learn to access memory that was stored externally instead of learning the sequences of the hidden states like the recurrent neural networks (Bahdanau et al. 2015).

The external data that is stored works like an embedding for the attention mechanism and this data can be altered. Writing the new information learned and reading if a prediction is needed to make. In a recurrent neural network, the hidden states are the sequences of embeddings, while in the memory of the attention model is the accumulation of those embeddings, which is like performing max-pooling on all the hidden states of the network.

To describe the process of attention, first, consider the input sequence as a set of internal states, and then we used the following steps:

1. The first step, is to compute a score of each state h_t. Where the network learns to give high scores to the states with a higher importance in that part of the sequence.

2. After we computed the scores, a softmax function is applied to generate attention weights α. This gives a probabilistic interpretation of the attention weights and the network learns where to put more attention.

3. Once the network computed the attention weights, is to get the context vector C. This vector is calculated by multiplying the attention weights α by their state h and then adding them.

4. The fourth step is to concatenate the context vector with the output word of the previous time step and feed it to the network $[X_{t-1}, h_t]$.

5. The final step is the output of the network that generates the next word in the sequence along with an internal hidden state.

Figure 2.11 presents an example of how we calculate the attention over a sequence.

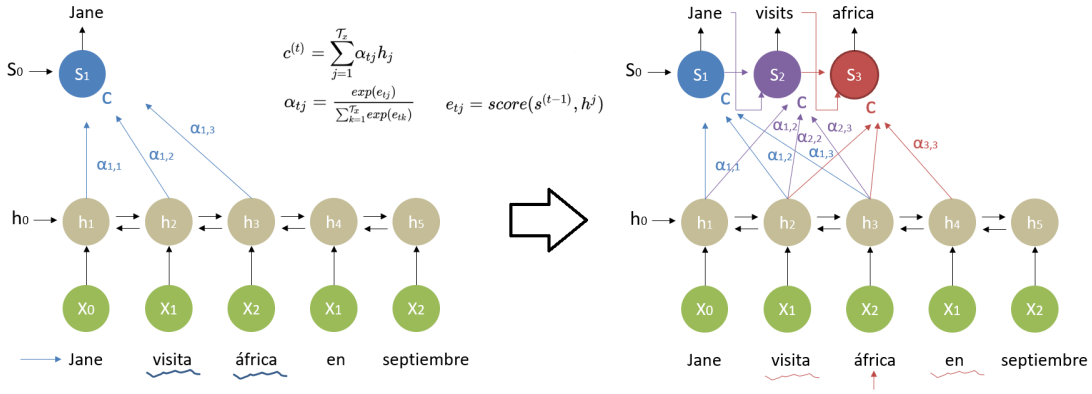

Figure 2.11: Attention over the context in the sequence. In the left part we can appreciate the original text in spanish and in the right part the translated text and the weights learnt through each step.

2.3.4 Transformers

Transformers is a neural network architecture based on a self-attention mechanism, dispensing the usage of recurrence and convolutions (Vaswani et al. 2017). This architecture transforms one sequence into another using an Encoder and Decoder. The Transformer differs from traditional recurrent networks because it does not need the usage of any recurrence like GRU or LSTM.

To capture the timely dependencies present in sequences an LSTM was one of the best ways to do it. However, in recent works (Devlin et al. 2019), using this kind of architecture improves the results in sequence-related tasks. Figure 2.12 shows the general model architecture of the Transformer. The Encoder is on the left, and the Decoder is on the right part. Both of the modules can be stacked on top of each other multiple times as needed (as is referred by *Nx* in the figure). The modules in the architecture mainly consist of Multi-Head Attention and Feed Forward layers. The Multi-Head Attention consists of the dot product of the weight matrices learned during the training. These matrices are defined by how each word in the sequence is affected by the other words of the sequence. For the inputs and outputs, the string sentences need to be represented by their embedding of n-dimensional space.

Using the Positional Encoding part in the architecture, the model could give to every sequence a relative position, and then, the position is added into the embedding. This part is fundamental since the model does not have recurrence to remember how the sequence was feed.

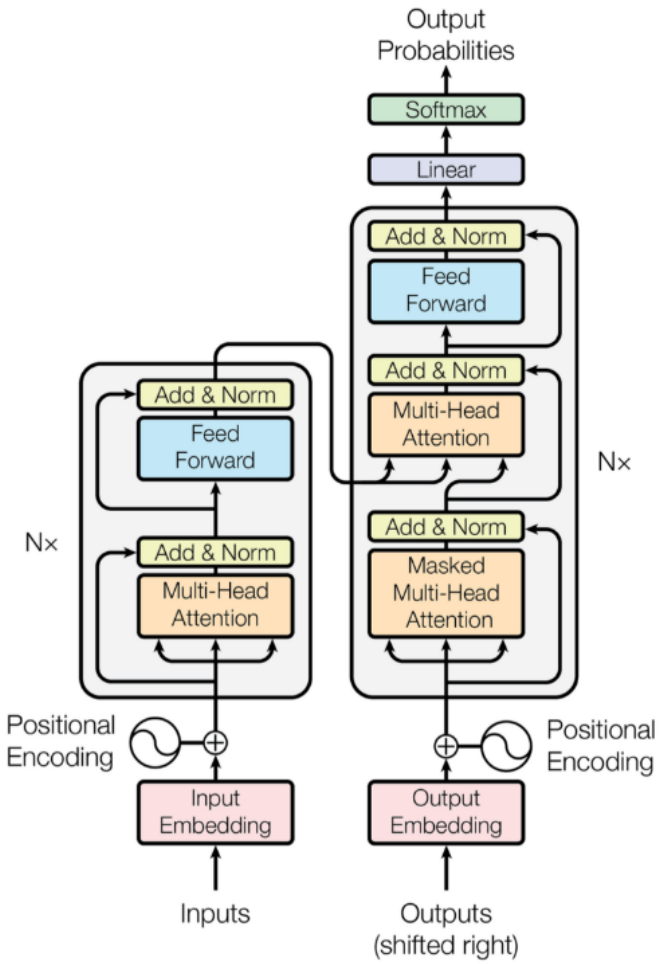

Figure 2.12: Transformer Model Architecture (Vaswani et al. 2017)

BERT

BERT is a recent work published by researchers at Google AI Language (Devlin et al. 2019) and stands for Bidirectional Encoder Representations from Transformers. BERT caused a stir in the NLP community by presenting state-of-the-art results in a variety of tasks. The main innovation from BERT is the bidirectional training of the Transformer's encoder to a language model. This is the main difference from previous works that looked at a text sequence from left-to-right and/or right-to-left during the training. With this technique, the

language model has a deeper understanding of language context in comparison with previous models.

BERT uses an encoder of the Transformer, an attention mechanism that learns contextual relations between words. For the training process, the model receives pairs of sentences as input and learns to predict the second sentence of the pair. During the training, half of the pairs use the original second sentence, while the other half use a random sentence. The intuition behind this process is that the random sentence will be disconnected from the first sentence, and help the model to learn the context.

Figure 2.13 presents the preparation of the input for the training process, where for each pair of sentences the input is processed as follows:

1. First, we need to add a [CLS] token at the beginning of the first sentence, and a [SEP] token at the end for each sentence.
2. Second, we need to add a sentence embedding indicating Sentence A or Sentence B.
3. Lastly, add a positional embedding to each token to indicate the position in the sequence.

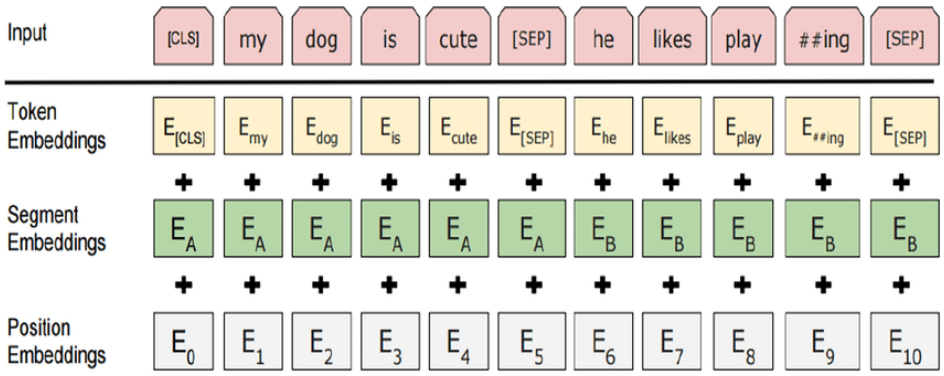

Figure 2.13: BERT input representation (Devlin et al. 2019)

Multimodal BERT

It is a recently proposed supervised multimodal bitransformers model for classifying images and text (Kiela et al. 2019). In this work, the authors proposed a simple yet highly effective multi-modal architecture inspired by the BERT model. The multi-modal BERT (MMBT) transformers use unimodally pre-trained components and outperform a variety of different fusion techniques. This method does not rely on particular feature extraction, and it works

for any sequence of dense vectors. For example, it can compute raw image features and backpropagate through the encoder.

The MMBT model initializes with pre-trained BERT weights. The architecture takes contextual embeddings as input, where these embeddings are obtained as the sum of separate D-dimensional segment, position, and token embeddings. Then the model weights them as $W_n \in R^{P \times D}$ and project each of the image embeddings to D-dimensional token input:

$$I_n = W_n f(img, n) \tag{2.20}$$

where $f(img, n)$ is the n-th output of the image encoder's final pooling operation. For tasks that only have a single text and a single image, the model assigns the text one segment id and image embedding the other one. This architecture can be generalized to any number of modalities. In Figure 2.14, we can appreciate the components of the architecture of the MMBT model.

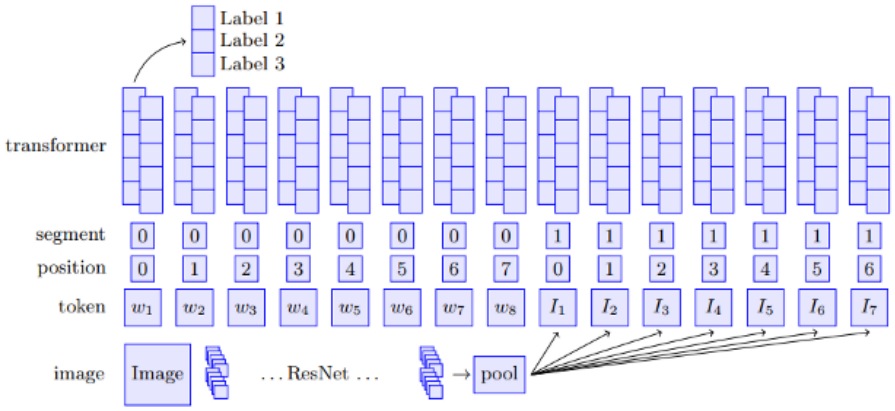

Figure 1: Illustration of the multimodal bitransformer architecture.

Figure 2.14: Multimodal bitransformer architecture (Kiela et al. 2019).

2.4 Gated Multimodal Units

One of the challenges of this work is the problem of fusing information. A simple solution is to concatenate the representations of each channel into one vector or perform an operation like adding or taking the product. However, the use of these operations assume that all channels have the same relevance, which is usually not the case. For our case, depending on the mental disorder, one or a sub-group of channels might have more information.

In recent work (Arevalo et al. 2019), the authors proposed a novel type of hidden unit called Gated Multimodal Unit (GMU). This unit works similarly to the control flow mechanism in gated recurrent units. The gates in the unit let the model regulate the flow of information into the next one. The main idea in a GMU is that the unit learns to weigh the modalities (channels for us) and fuse them according to their relevance. A GMU works similar to a neural network layer and finds an intermediate representation based on the different modalities.

Figure 2.15 presents a general overview of the GMU module. In the figure, we can appreciate that x_i is a feature vector associated with a modality i. For each vector, there will be a weight z_i, which controls the contribution of that modality. At the end of the unit, we obtain a final fused representation as to the weighted sum of each modality. These gates will allow the model to decide how each modality affects the unit's output. One of the advantages of the GMU is its interpretability. After training the model, we can visualize the weights z_i and have a better understanding of which modalities had more contribution to the prediction.

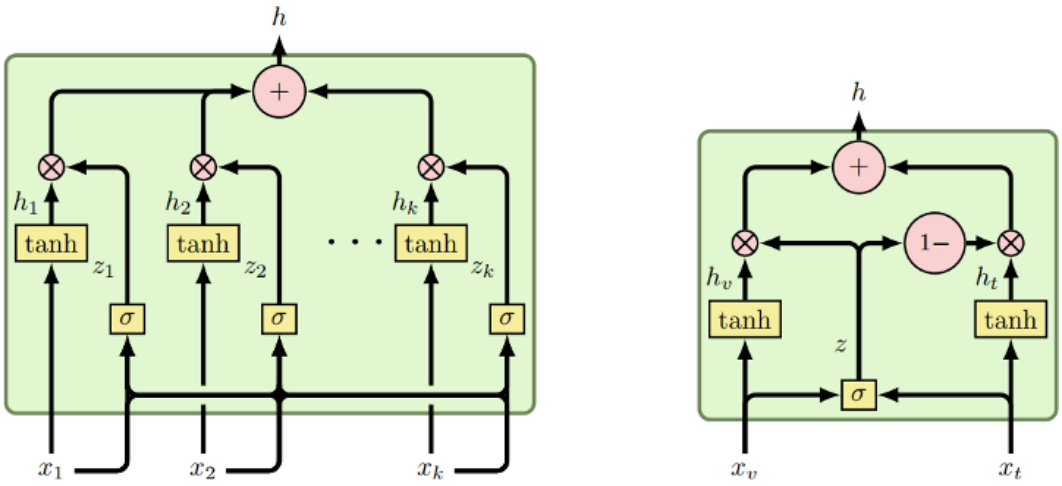

Figure 2.15: Overview of GMU module. Where x_i represents the ith input modality. The final fused representation of all modalities is represented by h at the top.(Arevalo et al. 2019)

2.5 Clustering and Affinity Propagation

Cluster analysis, or clustering, is an unsupervised machine learning task. The objective is to automatically discover natural groups in data (Witten et al. 2016). Contrasting with supervised learning, like prediction or regression learning, clustering algorithms only interpret the input data and find groups in a feature space.

A cluster is an area in the feature space where examples from the raw data are closer to the cluster than others. The clusters are conformed from instances with a strong resemblance

to each other in comparison with instances in different clusters. The cluster may have a center name centroid, which is a sample of the feature space that represents the cluster.

Clustering can perform data analysis to learn more about the problem we are facing. Clustering can also perform feature engineering, where new examples are labeled as belonging to one of the discovered clusters. Figure 2.16 presents some examples of different clustering algorithms, such as k-means, mixture model, hierarchical clustering, and graph based clustering.

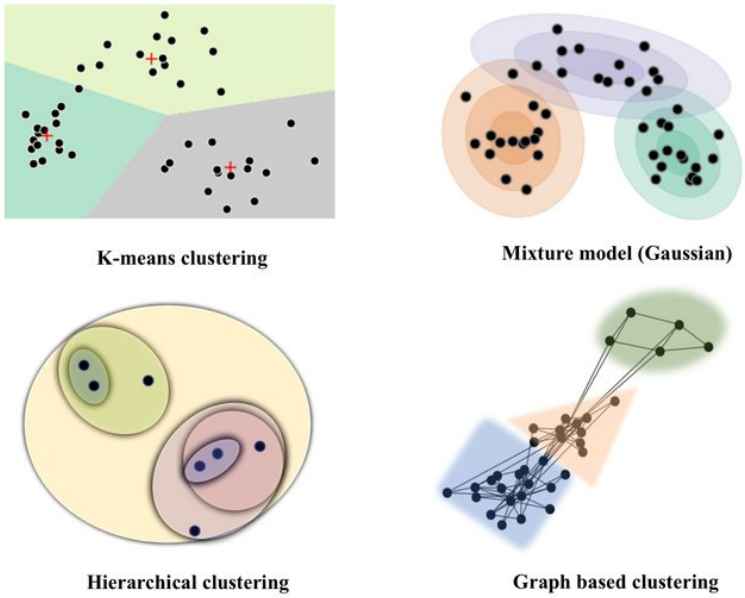

Figure 2.16: Popular clustering algorithms. We can appreciate the k-means clustering, mixture method, hierarchical clustering and graph based clustering (Affinity propagation is based in this algorithm)(Khosla et al. 2019).

Affinity Propagation

Affinity propagation (AP) is a clustering algorithm based on the concept of message passing between data points (Thavikulwat 2008). The main difference between AP and other popular clustering algorithms, like k-means, is AP does not require you to specify the number of clusters. In Affinity Propagation, each data point sends messages to all other points related to the target's relative likeliness. Then, each target responds with the availability to associate with the sender. This process of message-passing continues until they reach a consensus and all senders are associated with one of its targets. This target becomes the point's exemplar (or centroid) and the rest associated with the same examplar are placed in the same cluster.

The algorithm updates two matrices by alternating between two message-passing steps:

1. **R** is the responsibility matrix with values $r(i, k)$ that quantify how well-suited x_k is to

serve as the exemplar for x_i. In comparison with other candidate exemplars for x_i.

2. **A** is the availability matrix that contains the values $a(i,k)$ and represents how appropriate it would be for x_i to pick x_k as its exemplar. Taking into account other points' preference for x_k as an exemplar.

Both matrices are initialized in zeroes, and AP performs the following updates iteratively:

First, responsibility updates are sent as:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \tag{2.21}$$

Then, availability is updated:

$$a(i,k) \leftarrow \min \left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right) \tag{2.22}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)). \tag{2.23}$$

The algorithm continues until either the clusters remain unchanged after a number of iterations or some predetermined number of iterations.

2.6 Evaluation metrics

An evaluation metric quantifies the performance of a predictive model comparing the expected class label to the predicted class label. There are a variety of standard metrics that are widely used to evaluate classification models and rate predictions. In the following sub-sections we describe the ones used in this thesis.

2.6.1 Precision

Precision in pattern recognition and information retrieval is also called positive predictive value. When a program retrieves instances predicted as the ones of interest. The formula is express as:

$$Precision = \frac{TP}{TP + FP} \tag{2.24}$$

where TP are the right predictions and FP are the wrong predictions selected as correct.

2.6.2 Recall

As well as precision, recall is used in pattern recognition and information retrieval to evaluate the fraction of correct instances that have been retrieved over the total correct instances. The formula is express as:

$$Recall = \frac{TP}{TP + FN} \tag{2.25}$$

where TP are the right predictions and FN are the right predictions that were not selected as correct.

2.6.3 F1 score

Is an evaluation measure of the test accuracy, where it considers the precision and the recall to give the score of the evaluation. The F1 measure is considered the harmonic average of the precision and recall, where the score looks for the best value of precision and recall at 1, and also the worst value at 0. The formula is express as:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{2.26}$$

2.7 Mental Disorders

Mental disorders are conditions that affect behavior, thinking, and mood. These conditions may be occasional or chronic (long-lasting) and can affect the ability to relate in society (Medline 2022). There are different types of mental disorders, and some common ones include anxiety disorders, depression, eating disorders, or psychotic disorders. Several factors can cause them, such as life experiences, family history, biological factors, or abuse of alcohol. Mental disorders are not caused by personality flaws like being weak or lazy.

Depression

Mental Disorders affect millions of people around the world. Out of these disorders, depression has been ranked among the most common, even with a high incidence in mortality rates (Mathers and Loncar 2006). Depression is a mental health disorder characterized by persistent loss of interest in activities, which can cause significant difficulties in everyday life, fortunately, it is also treatable (Kessler et al. 2017). Depression can happen at any age, but it often begins in teens and young adults. There are different causes for depression, including genetic, environmental, and psychological factors. Nowadays, depression has effective treatments including antidepressants, or talk therapy.

Anorexia

Anorexia is the most common Eating Disorder (ED) that is related to a mental disorder. It consists of abnormal attitudes towards food and an unusual habit of eating, where generally someone that suffers from anorexia restricts what they eat to maintain low weight or lose more weight (Clinic 2018). Also, they tend to exercise compulsively, purge via vomiting and laxatives, and binge eating[1]. Anorexia, like other eating disorders, can become very difficult to cope with, but with treatment, someone affected can return to healthier eating habits and reverse some of the serious complications. Eating disorders frequently appear during the teen years or young adulthood, and anyone can develop them, but they are more common in women.

Self-harm

Self-harm is defined as the direct and intentional injuring of one's body; in extreme cases with the intent to commit suicide (Laye-Gindhu and Schonert-Reichl 2005). People that commit self-harm commonly use a sharp object to make cuts in their skin. This practice includes other acts such as burning, hitting body parts, ingestion of toxic substances, or scratching. Self-harm affects people of all ages, but similar to depression and anorexia, it usually starts in the teen or early adult years. People who harm themselves are usually at higher risk of attempting suicide if they do not get help. There are different reasons why people hurt themselves. These reasons are often related to having trouble dealing with their feelings or trying to block upsetting memories.

[1] https://www.nationaleatingdisorders.org/learn/by-eating-disorder/anorexia

CHAPTER 3

RELATED WORK

In this chapter, we present the most relevant related work on the detection of mental disorders. This thesis aims to propose novel and effective methods to capture different information from the users. For this, the strategies exploit NLP techniques and neural networks models to create new representations that improve the performance of traditional approaches based only on words or sentiments. In this regard, in Section 3.1 and 3.2 of this chapter, we present the most relevant general approaches and ensemble approaches for the detection of mental disorders. These approaches mainly consider the information among words, sentiment analysis, and neural networks.

3.1 General Approaches

Several recent works have taken advantage of social media platforms to study the manifestation of different mental disorders. Most of them used crowd-sourcing strategies to collect data (De Choudhury, Gamon, et al. 2013). In general, they identify a group of users who expressed in one of their publications having been clinically diagnosed with a mental disorder and then download all or part of their posts (De Choudhury, Counts, et al. 2013; Wang et al. 2017). The authors proposed a method to automatically gather individuals who self-identified as eating disordered in their Twitter profile descriptions. They analyzed their social interactions and found that this kind of user has significant mixed patterns in tweeting preferences, language use, concerns of death, and emotions. Having obtained their data, they apply a variety of methods to find relevant and discriminative patterns from the platform usage behavior, social interactions, or language use.

Regarding the use of language, some works have employed traditional classification al-

gorithms combined with the analysis of words and sequences of words as features (Schwartz et al. 2014; Trotzek et al. 2018; Tsugawa et al. 2015). Through this kind of analysis they aim to compare the sets of most frequent words used by users suffering from a mental disorder and healthy users (Coppersmith, Harman, et al. 2014). The problem with this approach is that the resulting vocabularies from both types of users tend to show a high overlap (Trifan and Oliveira 2019; Van Rijen et al. 2019). For example, in (Funez et al. 2018), the authors implemented a system based on two different models. One that considers the temporal variation of terms, and the other that carries out an incremental classification. The first model uses a semantic representation of documents considering the explicit information available at each chunk, whereas, the second model does an incremental estimation of the association of each user to each class based on the accumulated information at each chunk.

Other works have applied sentiment analysis techniques to study the emotional properties of the users' posts (Preotiuc-Pietro et al. 2015; Ramírez-Cifuentes and Freire 2018). They mainly model the positive, negative, and neutral sentiments that the users express, and explore the relationship between these sentiments and the signs of a mental disorder. One of the aims is to capture the expression of sentiments characterizing the posts along the dimension from positive to negative. They also measure the intensity of the sentiments from low to high and describe different states with these two values. Despite the interesting results of these analyses, they usually fail in detecting users without a mental disorder that tends to express themselves negatively (Coopersmith, Dredze, et al. 2014; Coppersmith, Ngo, et al. 2016). These works also inspired the use of emotions to identify depression (M. 0. Aragon et al. 2019; Xuetong et al. 2018), and a psychological theory that relates the manifestation of feelings and emotions with depression (Association 2013). These results showed that emotion-related expressions capture insights into users' psychological states. The authors also demonstrated that measuring emotional changes in an individual further improves effectiveness.

Previous works have addressed the analysis of emotions of social media users by paying attention to their contrast and tone. They have mainly applied this analysis to predict users' age and gender, as well as a range of sensitive personal attributes including sexual orientation, religion, political orientation (Kosinski, Bachrach, et al. 2014; Kosinski, Stillwell, et al. 2013), income (Preoţiuc-Pietro et al. 2015), and personality traits (Correa et al. 2010; Volkova and Bachrach 2016). According to these studies, the analysis of emotions in social media allows capturing important information related to users. This information presents an opportunity for us to extend the use of emotions in the detection of depression and anorexia in social media.

Other groups of works have used an LIWC-based representation (Tausczik and Pen-

nebaker 2010), which consists of a set of psychological categories that aim to represent users' posts by features of social relationships, thinking styles and individual differences (Coppersmith, Dredze, Harman, and Hollingshead 2015). The authors measure for each user the proportion of word tokens that belong to a category of LIWC and also the proportion of each category. With this measure, they obtained a distribution of the proportion of language applicable to each category. Something interesting from this analysis is that categories show differences across mental disorders. For example, some categories with high value were the ones related to anxiety, auxiliary verbs, function, health, cognitive mechanisms, and death. This suggests that the language used by people with mental disorders can be captured with specific categories. The relation of these categories of LIWC and mental health conditions has already been supported in the mental health literature (Eichstaedt et al. 2018). This strategy allows for a better analysis of the users suffering from a mental disorder, nevertheless, its results are only moderately better than those from word-based approaches (De Choudhury, Gamon, et al. 2013).

3.2 Ensemble Approaches

Recently, some works have considered the use of ensemble approaches, which combine the previously mentioned representations with different deep neural models (Trotzek et al. 2018). The authors submitted results from four machine learning models and an ensemble model that combines the predictions of the previous models. They employed user-level linguistic metadata, a bag of words representation, neural word embeddings from Glove, and a convolutional neural network. In (Masood 2019), the author proposed a solution based on neural networks, multi-task learning, domain adaptation, and Markov models. This work is still in its early stages, one of the issues is how to extend the mental health information resources. In a different work (Liu et al. 2018), they combine the frequencies of words, user-level linguistic metadata, and neural models with word embeddings; it obtained the best-reported result in the eRisk-2018 shared task on depression detection (Losada, F. Crestani, et al. 2018). On the other hand, (Mohammadi et al. 2019) shows a neural network architecture consisting of eight different sub-models, followed by a fusion mechanism that concatenates the features and predicts if a user presents signs of anorexia. In that work, the authors concluded that the combination of different models obtained better performance than using them separately, suggesting that the different types of features enrich the users' representation and provide relevant information for the detection of anorexia.

In a slightly different direction, (Ragheb et al. 2019) models the temporal mood variation using an attention network. Their approach performs the detection of users with mental disorders through two learning phases. The first phase uses an attention-based deep model

to construct a representation for the temporal mood variation. Then, in the second phase, the model uses a Bayesian inference model to obtain the decision. The main idea is to detect clear signs of mental disorders from moods variation and then give a decision. In (Ji et al. 2020), the authors apply an attention model combined with sentiment and topic analysis to detect suicidal ideation. Similar to the previous work, this model consists of two phases. The first phase is the creation of the representation that includes two parts of features extraction of risk-related state indicators and an LSTM text encoder. The second phase is the relation module, where a vanilla relation network connects the state indicators and attention mechanisms to capture the important relation scores of text encoding. These last two studies show the potential of the attention mechanisms in these types of tasks, as their results outperformed those of other deep neural models.

A different and interesting approach is presented in (López-Monroy et al. 2018), where the authors proposed a novel document representation called Multi-Resolution Representation (MulR). A strategy to improve the early detection of risks in social media. This representation generates multiple views or resolutions of the text and captures different semantic meanings for words and documents at different levels of granularity. With this strategy, they model early scenarios with a variable amount of evidence or posts. The experimental results show that using low-resolution information models is better at the early stages of short documents, and for large documents, the higher resolutions capture better information.

3.2.1 Discussion

Despite their good performance, an important limitation of the ensemble approaches is the interpretability of the results, even more if the final objective is to create a tool aimed to support health professionals. In this regard, in (Burdisso et al. 2019; Ríssola et al. 2020) the authors conduct studies to tackle this problem. They characterize users affected by mental disorders and provide methods for visualizing the data in order to provide useful insights to psychologists. Some of the attributes used for the modeling were the activity, vocabulary, psychometric attributes, and emotional indicators in users' posts. The work presents interesting differences that could help predictive systems and health workers to determine if someone has a mental disorder. Based on the good performance shown by the ensemble approaches in the detection of depression and anorexia, and motivated by the design of models that could be easily understood. We decided to implement our multi-channel approach as a new and simple way to combine different views of the information shared by social media users. Furthermore, our emotion and style representations capture important information from the users presenting simplicity and interpretability, then create a more straightforward analysis of the results.

3.2.2 Interactions with healthcare practitioners

As mentioned before, in recent years, there is a growing work from social media to predict the mental health status of individuals (Coopersmith, Leary, et al. 2015; Coppersmith, Dredze, and Harman 2014; De Choudhury, Counts, et al. 2013). Currently, clinicians are exploring how efficient are diagnostic predictions from online data for early diagnosis and providing timely patient interventions (Eichstaedt et al. 2016; Fisher and Appelbaum 2017). One of the main challenges is to obtain clinically valid diagnostic information from sensitive patient populations and researchers build models characterizing online behaviors. However, it is worth mentioning the work in (Ernala et al. 2019), where the authors found that predicting a user with a mental disorder using their social media information although offers strong internal validity, suffers from external validity when tested on mental health patients; demonstrating that there is still work to be done in this area.

Part III

Contributions

CHAPTER 4

DETECTING TRACES OF MENTAL DISORDERS THROUGH EMOTIONAL PATTERNS

4.1 From Sentiments to Emotions

The inspiration behind our work started in the sentiment analysis task for social media corpora. Sentiment analysis is the process of detecting positive and negative sentiments in text. A well-known strategy consists of counting the number of occurrences of positive, negative, and neutral words in texts (Kang et al. 2016), or on measuring how similar are their words to some reference negative and positive words (Htait et al. 2017). Analyzing sentiments has shown interesting results since it has been found that negative comments are more abundant in people with a declared mental health disorder than in comments generated from a control group, which loosely we can call *healthy users* (Coopersmith, Dredze, et al. 2014; Preotiuc-Pietro et al. 2015). Moving a step forward, in a recent work (Xuetong et al. 2018), the authors proposed the use of emotions to identify depression in Twitter users. That study openly exposed the potential of using discrete emotions as features, instead of only using linguistic features, and broad sentiment categories to represent them. The analysis of emotions showed promising results, improving the detection of depressed users in comparison with only using sentiments and word frequency.

Motivated by the results reported in (Xuetong et al. 2018), we decided to apply a similar methodology to analyze the distribution of emotions in the posts of users suffering from a mental disorder. For this analysis, we used a lexical resource of eight emotions and two sentiments (Mohammad and Turney 2013) (which is explained in more detail in the next subsection). We first counted the number of words associated with each emotion and then

computed their average value for all users. Figures 4.1, 4.2 and 4.3 presents the distribution of emotions corresponding to the users from the eRisk dataset. For depression users, we can appreciate that the negative sentiment is prevailing in comparison with the positive and they express a lot of anger in comparison with control users. For anorexia, we can appreciate that sentiments and emotions present similar values for both type of users. Broadly, for self-harm, it shows that control and no-control users tend to express different emotions through their posts. In particular, people suffering from mental disorders use more negative words than control users, whereas the latter express joy and trust more frequently than the former.

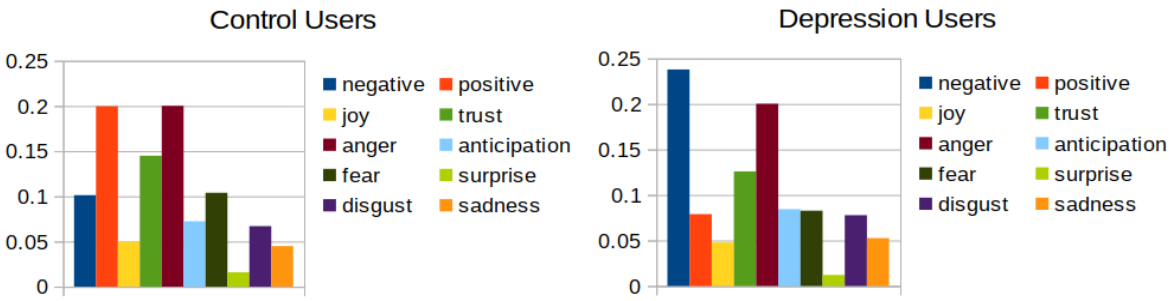

Figure 4.1: Distribution of emotions at the depression task. For this, we counted the number of words with each emotion and average the value for all users.

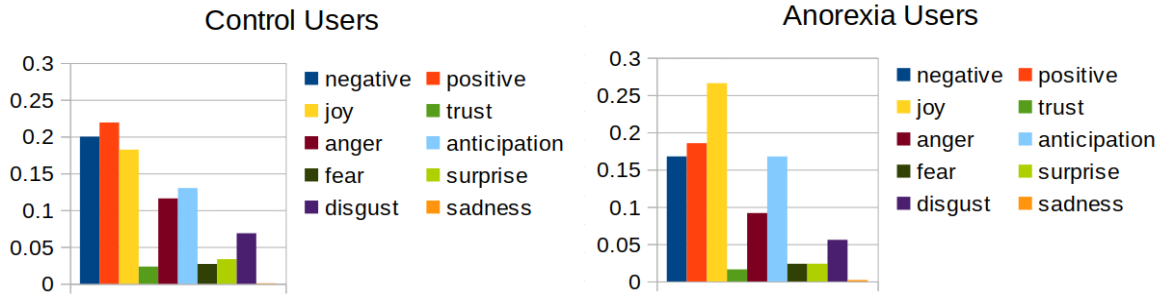

Figure 4.2: Distribution of emotions at the anorexia task. For this, we counted the number of words with each emotion and average the value for all users.

4.2 Can emotions have shades? A sub-emotion based representation

In the previous sub-section, we explained why emotions are important for the detection of users with a mental health disorder. However, we did not mention an important limitation for it: when we consider the words assigned to coarse emotions in lexicons, we can not capture subtle emotional differences. For example, the lexicon associated with the anger

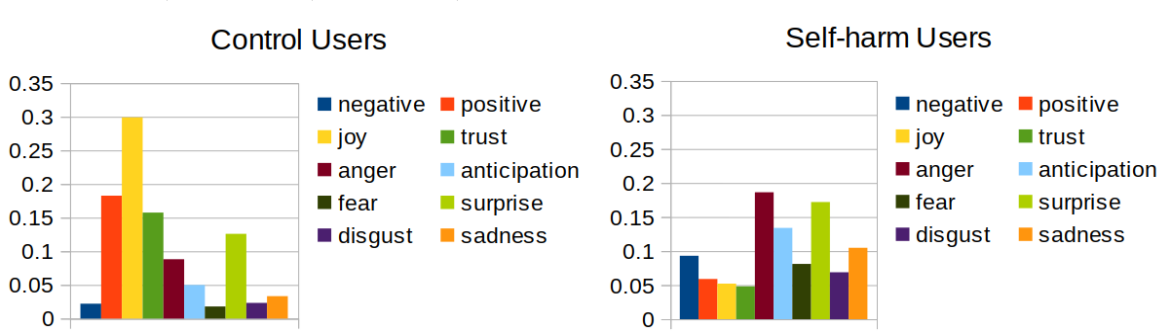

Figure 4.3: Distribution of emotions at the self-harm task. For this, we counted the number of words with each emotion and average the value for all users.

emotion includes words such as *furious*, *angry* and *upset* that represent different degrees of anger, however, they are all tagged with the same emotion, that is, at the same level (in this case, anger). We argue that these differences provide important insights into the mental health condition of users.

Inspired by this intuition, we decided to evaluate emotions, or more precisely, fine-grained emotions as a representation to successfully identify Reddit's users with a mental disorder. We believe that the analysis of the expression of these kinds of emotions is suitable to reveal symptoms or insights of people that have some psychological distress.

Our proposed method for the detection of mental disorders consists of the representation of the user's post history based on their expressed fine-grained emotions. The first step to construct these representations is to generate groups of fine-grained emotions (referred to as *sub-emotions* from here on). For this, we used the EmoLEX lexicon (Mohammad and Turney 2013) and selected each general emotion that belongs to it. EmoLEX is a lexical resource that indicates the relation of words to eight different emotions: Anger, Fear, Anticipation, Trust, Surprise, Sadness, Joy, and Disgust, and two sentiments: Negative and Positive[1]. To construct EmoLEX, its authors manually annotated the words to indicate if they express one or more of the after-mentioned emotions. This lexical resource is available in 40 different languages. Once we have the sub-emotions, we masked the user's posts using the sub-emotions labels instead of the original words. In the following sections, we describe in detail each step of this procedure.

4.2.1 Generating Sub-emotions

To generate the sub-emotions we follow the next process:

[1]In the rest of the chapter we refer to these sentiments as emotions as well.

1. Use a set of words associated with emotions in the EmoLEX lexicon. In the lexicon, there are some words associated with more than one emotion. We formally represent this set of emotions as $E = \{E_1, E_2, ..., E_{10}\}$, where $E_i = \{t_1, .., t_n\}$ is the set of words associated to emotion E_i.

2. Compute an embedding vector for each word that the EmoLEX lexicon resource contains. For this, we used Wikipedia pre-trained sub-word embeddings of size 300 from FastText (Bojanowski et al. 2016). It is important to mention that for this step we could use any other type of word embeddings but we select FT due the flexibility in representing words, specially the ones out of the vocabulary because it works a character level (refer to section 2.1.2). We depict these first two steps in Figure 4.4.

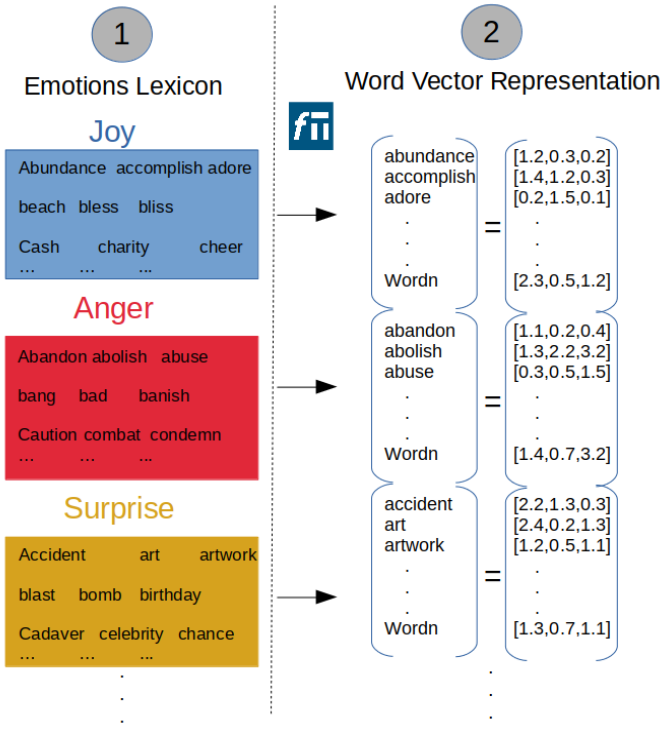

Figure 4.4: Compute word embedding vectors for each word in the EmoLEX lexicon. For simplicity, in the Figure we only use three emotions, but this process was applied to the eight emotions and two sentiments considered.

3. Form groups of sub-emotions. For this, for each emotion, we clustered the words using the *Affinity Propagation (AP)* algorithm (refer to section 2.5). This method is a graph-based clustering algorithm similar to k-means, but that does not require the number of clusters to be specified in advance. An additional advantage of this algorithm is that it finds examples of members of the input set, and uses them as representatives of the

clusters (Thavikulwat 2008). After the clustering, each centroid represents a different sub-emotion.

4. Save for each emotion a set of sub-emotions represented by their vectors, $E_i = \{S_1, ..., S_k\}$, where each S_j is a subset of the words from E_i. This whole process creates a super set S with all computed sub-emotions, where words with similar contexts tend to group together. In Figure 4.5 we present these last two steps.

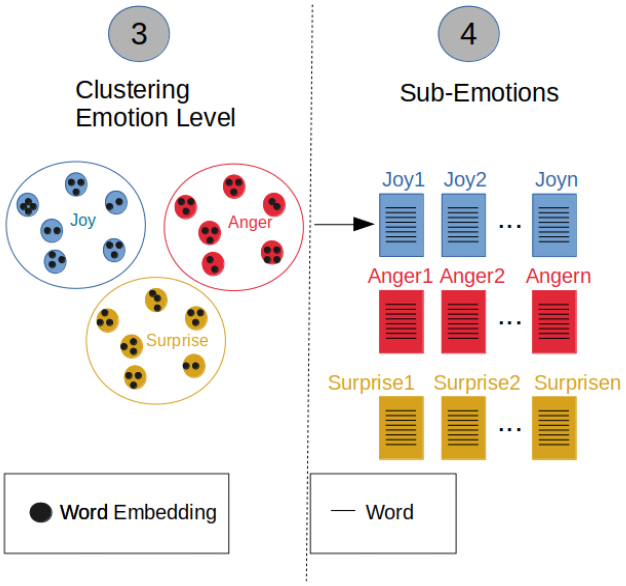

Figure 4.5: Sub-emotions obtained with the affinity propagation algorithm for each emotion. Each sub-emotion represents a group of words with the same label.

4.2.2 Analysis of the novel sub-emotions

For a complete picture of how the vocabulary in the EmoLEX lexicon is distributed among emotions, as well as how it is distributed across the created sub-emotions by the affinity propagation (AP) algorithm, in Table 4.1 we present some statistics of the lexicon. We indicate the number of words and clusters (i.e., sub-emotions) per broad emotion, as well as the average and standard deviation of their internal cohesion (μCoh and σCoh, respectively). The internal cohesion is a metric used to calculate how similar an object is to its cluster; a value closer to 1 indicates high similarity, and a value closer to -1 indicates very low similarity. We obtain this metric by measuring the cosine similarity of each word concerning the others in the same cluster.

Some interesting aspects to note from Table 4.1 are the following. First, the size of the vocabularies associated with each broad emotion is large, as well as diverse, since the num-

ber of clusters obtained per emotion is also large. However, we see that the average number of words per cluster (μW) is similar for all emotions, which indicates that the AP algorithm could find a similar distribution of words in the clusters even for emotions with large vocabularies. Second, the cohesion values indicate that although not all the sub-emotions generated present high cohesion, in all cases there are groups with an important degree of cohesion, which help capture very specific topics. Furthermore, a large part of these sub-emotions were not seen in user histories and only a percentage were used.

Table 4.1: Size of the vocabulary for each emotion presented in the lexical resources, and number of generated clusters (Cls).

Coarse Emotion Stats		**Discovered Sub-Emotions Stats**				
Emotion	Vocabulary	Cls	μ W	σ **W**	μ Coh	σ Coh
anger	6035	444	13.60	16.53	0.2932	0.1588
anticip	5837	393	14.77	20.53	0.2910	0.1549
disgust	5285	367	14.4	21.29	0.2812	0.1601
fear	7178	488	14.70	23.36	0.2983	0.1455
joy	4357	318	13.70	21.25	0.2928	0.1638
sadness	5837	395	14.78	20.48	0.2911	0.1549
surprise	3711	274	13.54	28.68	0.2874	0.1626
trust	5481	383	14.31	21.59	0.2993	0.1609
positive	11021	740	14.89	24.53	0.2967	0.1466
negative	12508	818	15.29	23.75	0.2867	0.1417

It is important to note that some high-cohesive sub-emotions provide an easy understanding and interpretability. For example, the obtained sub-groups of words separate each coarse emotion in different related topics, which in turn capture more specific emotions expressed by users in their posts. In Figure 4.6, we present some examples of the groups of words that represent the sub-emotions. It is important to mention that these groups were automatically obtained using this approach. Note that words with similar contexts tend to group. For example, for the emotion **Anger** one group expresses topics related to fighting and battles, and for another group, we can appreciate topics related to loud noises or growls. In another example, the **Surprise** emotion has one group that is related to accidents and disasters, whereas other groups have words that are related to art and museums, and magic and illusion, respectively.

Anger			Joy		
abomination	growl	battle	accomplish	bounty	charity
fiend	growling	combat	achieve	cash	foundation
inhuman	thundering	fight	gain	money	trust
abominable	snarl	battler	reach	reward	humanitarian
unholy	snort	fists	goal	wealth	charitable
Surprise			Disgust		
accident	art	magician	accusation	criminal	cholera
crash	museum	wizard	suspicion	homicide	epidemic
disaster	artwork	magician	complaint	delinquency	malaria
incident	gallery	illusionist	accuse	crime	aids
collision	visual	sorcerer	slander	enforcement	polio

Figure 4.6: Examples of words grouped in different sub-emotions. Note that words with similar contexts tend to group together.

4.2.3 Converting text to sub-emotions sequences

Our general procedure is to concatenate all the individual posts of each user and create a single document for each one. Then, we mask all users' documents replacing their words with the label of the closest sub-emotion. The following points describe in detail our approach:

1. Compute prototypical sub-emotion vectors by averaging the word embeddings in each cluster. We use the labels of these prototypes to count each word in the text as an occurrence of each sub-emotion. We also compute the word vector for each word in the vocabulary of the users. Fig 4.7 depicts this step.

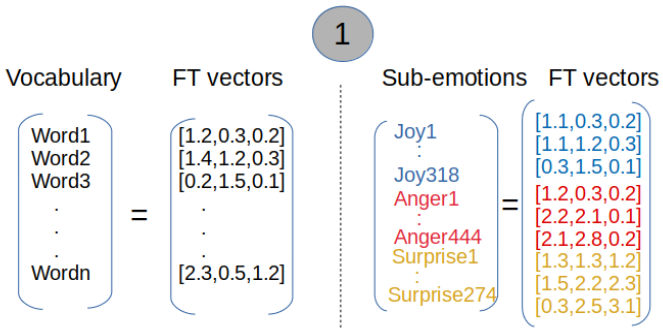

Figure 4.7: Determination of the the vocabulary words and sub-emotions vectors.

2. Measure for each word t in the vocabulary its cosine similarity with all sub-emotions vectors S.

3. Replaced the word t by a label $\tau(t)$ related to its closest sub-emotion[2]. That is, $\tau(t) = S_j : \max_{\forall S_j \in S} sim(\vec{t}, \vec{S_j})$. We present these two steps in Figure 4.8

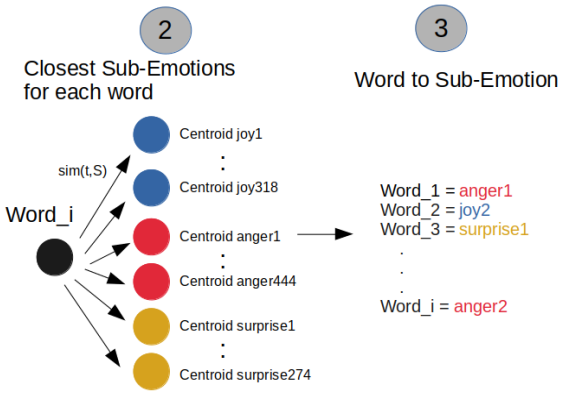

Figure 4.8: Replace each word in the vocabulary with its closest sub-emotion.

4. Mask the words from the post history with their corresponding sub-emotion. After this step, documents are sequences of sub-emotions instead of sequences of words. We present this final step in Figure 4.9

To illustrate this process, consider the following two example sentences:

1. *Two days ago I found out my girlfriend cheated on me and we broke up while I was coming down it was the worst day of my life I tried to buy a gun to kill myself with but the guy never showed up (thankfully).*

2. *I guess my point is comedowns suck especially if your life is collapsing around you.*

These sentences will be masked as:

1. disgust3 trust12 surprise18 joy6 joy4 trust27 surprise18 positive12 disgust6 anticipation22 anticipation16 joy6 surprise18 disgust19 surprise24 disgust9 joy6 joy6 anticipation22 surprise24 anger9 joy6 joy6 disgust19 surprise9 joy6 surprise18 sadness1 joy6 trust27 positive2 joy6 anger8 positive2 surprise7 positive9 anger9 trust15 joy6 anticipation1 joy16 surprise3 surprise24 trust12 joy16 trust23

2. joy6 surprise18 surprise18 anger9 anger9 negative28 disgust23 joy4 surprise18 positive9 sadness1 anger9 disgust19 anger9 positive9

[2]We assigned the labels selecting the name of the emotion followed by the sequential number. For example, for anger, the labels were assigned as anger1, anger2, ...,anger$K$. Where K is the number of clusters in that emotion.

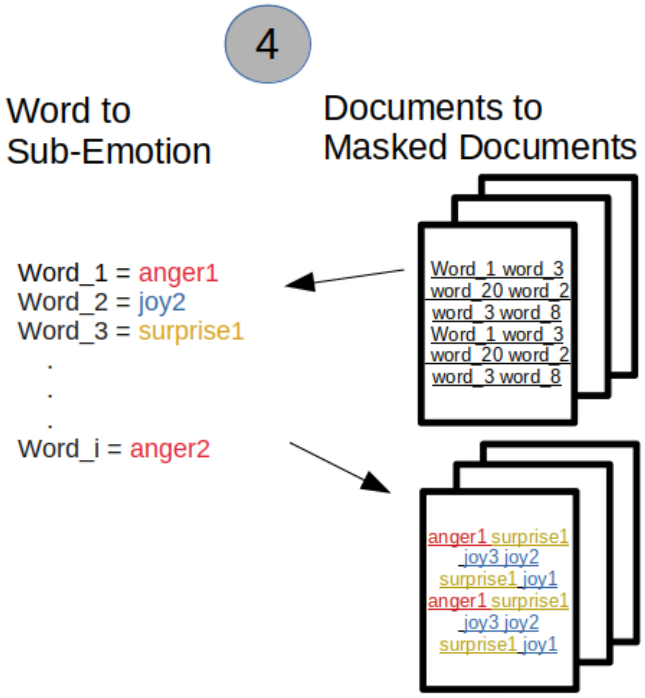

Figure 4.9: Masking the users' documents, replacing word sequences for sequences of sub-emotions.

From these examples, it is possible to appreciate how different contexts are captured by different sub-emotions. It is important to mention that we replaced the whole vocabulary of all users including stopwords, for their closest sub-emotions. Our decision to keep stopwords was completely experimental, we performed experiments removing them before masking the texts, but consistently we got slightly lower performances. We believe some stopwords add important information to the model for interpreting their surrounding words. It is important to remember that not all sub-emotions were present in users' texts and only a percentage was used.

4.3 Using BoSE to identify mental disorders

In this section, we present the sub-emotion-based representation, the so-called Bag of Sub-Emotions (BoSE), which allows us to model the presence of the sub-emotions through the users' posts.

4.3.1 BoSE definition

Once we masked the documents, the next step is to build the BoSE representation by using histograms of sub-emotions. For this, we represent each document d as a vector of weights associated with sub-emotions. Represented in a formal way we have:

$$\vec{d} = [w_1, \cdots, w_m] \tag{4.1}$$

where m is the total number of generated sub-emotions, and $0 \leq w_i \leq 1$ represents the relevance of sub-emotion S_i to the document d. This weight is computed in a *tf-idf* fashion as:

$$w_i = freq(S_i, d) \cdot \log\left(\frac{|\mathcal{D}|}{\#_{\mathcal{D}}(S_i)}\right) \tag{4.2}$$

where $freq(S_i, d)$ represents a function that denotes the frequency of the sub-emotion S_i in the document d, $|\mathcal{D}|$ is the number of documents in the whole collection, and $\#_{\mathcal{D}}(S_i)$ is a function denoting the number of documents containing the sub-emotion S_i. With this information, we determine how relevant that sub-emotion is in the entire training set. Observe that this representation only considers the presence of individual sub-emotions in the documents. Therefore, we call it **BoSE-unigrams**. In the case when we consider the presence of sequences of sub-emotions (continuous word sequences), we named the representation as **BoSE-ngrams**.

4.3.2 Experiments and results

Data sets

To thoroughly evaluate our proposed approach, we use the data sets from the eRisk 2018, 2019, and 2020 evaluation tasks (Losada, 0. F. Crestani, et al. 2019; Losada, F. Crestani, et al. 2020), which are for the detection of depression, anorexia, and self-harm. These data sets contain the post history of several users from the Reddit platform. For each task, we have two categories of users: 1) positive users, those affected by either depression, anorexia, or self-harm; and 2) the control group, composed of people who do not suffer from any mental disorder.

The positive class is composed of people who explicitly mentioned that they were diagnosed by a medical specialist with depression/anorexia or that they had committed self-harm. The creator of these data sets mentioned that they discarded users using vague expressions

Table 4.2: Data sets used for experimentation, where P indicates the positive users and C is used for control users.

Data set	Train		Test	
	P	C	P	C
Depression	135	752	79	741
avg. num. posts	367.1	640.7	514.7	680.9
avg num. words per post	27.4	21.8	27.6	23.7
avg. activity period (days)	586.43	625.0	786.9	702.5
Anorexia	61	411	73	742
avg. num. posts	407.8	556.9	241.4	745.1
avg num. words (post)	37.3	20.9	37.2	21.7
avg. activity period (days)s	800	650	510	930
Self-harm	41	299	104	319
avg. num. posts	169.0	546.8	112.4	285.6
avg num. words (post)	24.8	18.8	21.4	11.9
avg. activity period (days)	495	500	270	426

like "I think I have anorexia" during the gathering of data. The control class for both tasks is composed of random users from the Reddit platform. The Control group also contains users who often interact in the depression, anorexia, or self-harm threads to add more realism to the data and make the detection of positive users more challenging and close to reality. Table 4.2 shows how classes distribute within these data sets as well as some general information regarding the collections.

To offer a glimpse of the data sets, we present some examples of posts from the different classes of users. Our intention is to show that users who suffer from a mental illness, as well as control users, share personal experiences and their feelings about them, which for both can be positive and negative, making their identification a great challenge.

Depression

1. `After coming home from a road trip with a group of friends to celebrate my birthday.`

2. `Sometimes I can't help but think that they will be so much better off without me, and they know that they would be happier without me.`

Anorexia

1. I'm happy to hear that you're okay with realizing you'll be on antidepressan for the rest of your life..
2. My coach looked over at me then muttered; "It's a shame. If she wasn't so BIG I'd consider her for the team.

Self-harm

1. I don't even enjoy video games, I just use them as a distraction since I am not good at anything else I have tried many things but something always stops me from pursuing it further.
2. At some point I realized I wasn't really contributing to the conversation so I moved to the side, my place was immediately replaced and I was forgotten.

Control

1. Nice job; it's not always easy with the clouds. I love the colors of those waters with the glacial moraine. Beautiful image.
2. It was difficult, I do not expect it to be well-received here, but even if one person find it useful i will continue.

In the previous posts we can see how users who have a mental disorders can have posts with content related to their problem or normal day-to-day activities. This type of activity in their publications makes it difficult to detect these users.

Experimental Settings

Preprocessing. The texts were normalized by lowercasing all words and removing special characters like URLs, emoticons, and #; the stopwords were kept. Then, the preprocessed texts were masked using the created sub-emotions.

Classification. The main goal is to classify users into one of the two classes (Depressed / Control, Anorexia / Control, or Self-harm / Control). After building the **BoSE** representation, the most relevant features of the sequences of sub-emotions were selected using the term frequency-inverse document frequency (tf-idf) representation and chi^2 distribution X_k^2 (Walck 2007). With the selected features, we fed a Support Vector Machine (SVM) with a linear kernel. $C = 1$, L2 normalization and weighted for class imbalance. We empirically search for the best number of features for each task. The final prediction is using the whole

Table 4.3: F1 results over the positive class in three eRisk's tasks.

Method	Anor	Dep	SH
Baselines			
BoW-unigrams	0.67	0.58	0.50
BoW-Ngrams	0.66	0.57	0.50
Bag of char 3grams	0.67	0.58	0.53
RNN-word2vec	0.65	0.57	0.55
CNN-word2vec	0.66	0.60	0.56
RNN-glove	0.65	0.58	0.57
CNN-glove	0.67	0.61	0.57
RNN-Attention	0.66	0.50	0.58
BoSE-unigrams	**0.70**	0.61	0.60
BoSE-Ngrams	0.69	**0.63**	**0.62**

post history of the users and we classify the user as positive if the SVM decides the example is closer to this class.

Baselines. The results are compared to the traditional Bag-of-Words representation. Both representations were created using word unigrams and n-grams (size 1, 2 and 3); these are common baseline approaches for text classification. For both approaches, similar to BoSE, we selected the same number of features using tf-idf representation and chi^2 distribution X_k^2. We also add some baselines based on deep learning approaches, using a CNN and a Bi-LSTM. The neural networks used 100 neurons, an adam optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN we use 100 random filters of sizes 1, 2, and 3 (parameters recommended in literature). For all this comparison we consider F_1 score over the positive class, which was suggested as the golden standard by the organizers of eRisk 2018 (Losada, F. Crestani, et al. 2018).

Evaluation of the BoSE Representation. In this study, we exhaustively evaluate the BoSE-based representation, and we contrast them against the BoW schemes (using both unigrams and n-grams) and also against Deep Learning models (using Glove and word2vec). Table 4.3 presents the F_1 score over the positive class for this first evaluation. From this comparison, we appreciate that BoSE outperforms all baseline results, even in some cases with a good margin of difference. Surprisingly, the performance of deep learning models is remarkably low; to some extent, this could be attributable to the small size of the employed data sets. Indeed, most participants of eRisk that employed this kind of model combined them with traditional approaches to leverage their results.

Is there a sub-emotion pattern for mental disorders?

A relevant question that appeared at this point is: *Are there specific sub-emotions related to mental disorders?* In order to answer this question we need to analyze what sub-emotions BoSE captured. In Tables 4.4, 4.5 and 4.6 we show some of the top relevant sub-emotions (according to the chi^2 distribution), as well as some examples of the words that correspond to these sub-emotions.

Most of the sub-emotions that present high relevance for the detection of depression are related to negative topics, for example, the anger sub-emotions are associated with the feeling of abandonment or unsociability, and the disgust sub-emotions are related to delusion, insecurity, and desolation. These sub-emotions captures the way depressed users express their opinions about the world. In contrast, for the anorexia detection task, the sub-emotions that present higher relevance are related to embarrassment, self-harm, and eating topics. For example, the anger sub-emotions are related to victimization, bleeding, or bruises. The disgust sub-emotions are associated with mental states of defeat and internal organs related to eating. The latter clearly showed that these sub-emotions capture the essence of the problems of a person that suffers from anorexia. For self-harm, some topics are related to negative aspects, like hate, criticism, or refusal. An interesting sub-emotion captured automatically by our model is the negative sub-emotion related to young people, since people that commit self-harm usually are teenagers or closer to that age.

Table 4.4: Examples of relevant sub-emotions for depression detection

Sub-emotion	Associated words
anger1	abandoned, deserted, unattended
anger11	unsociable, crowd, mischievous
anticip10	disappointed, inequality, infidelity
anticip99	desolate, seclude, inhospitable
disgust16	unsatisfactory, delusion, influence
disgust11	insecurity, desolation, incursion
fear17	hysterical, immaturity, injury
fear20	suffer, unhealed, sickness
joy1	abundant, abundance, plentiful
joy14	life, time, moment
negative10	accident, incident, fatal
negative11	accuse, complaint, wrongdoing
positive10	approval, authorize, approved
positive100	attention, notoriety, publicity

surprise8	anxious, desire, wanted
surprise20	distress, embarrassment, shame
trust9	completion, continuation, progress
trust20	ambition, desire, wanted

Table 4.5: Examples of relevant sub-emotions for anorexia detection

Sub-emotion	Associated words
anger4	bruising, contusion, bleeding, fracture
anger15	delinquent, victimized, victimization
anticip10	hurting, refused, anxious, afraid
anticip12	ashamed, embarrass, upset, disgust
disgust32	breakdown, fight, crushed, abandoned
disgust21	stomach, intestinal, bile, esophagus
fear19	food, eating, eat, consume
fear101	illness, sickness, suffer
joy10	gain, attain, surpass
joy13	age, young, youngster
negative65	bathroom, toilet, washroom
negative105	bread, cake, tart
positive10	feeding, nutriment, nutrient
positive19	helpful, information, relevant
surprise18	oddness, quirkiness, strangeness
surprise28	betrayal, deceiver, desolation
trust23	admiration, fondness, esteem
trust25	hunger, thirst, solace

Table 4.6: Examples of relevant sub-emotions for self-harm detection

Sub-emotion	Associated words
anger11	unsociable, crowd, mischievous
anger108	traumatic, trauma, traumatize
anticip104	attack, offensive , harass
anticip10	hurting, refused, anxious, afraid
disgust17	condemn, criticise, refuse, repudiate
disgust32	breakdown, fight, crushed, abandoned

fear5	dreadful, hate, bad, nasty
fear114	aggression, hostile, discord
joy14	life, time, moment
joy104	effort, chance, opportune
negative18	adolescence, teen, juvenile
negative115	conceal, hide, concealment
positive110	protect, safeguard, protection
positive19	helpful, information, relevant
surprise20	distress, embarrassment, shame
surprise108	oddness, quirkiness, strangeness
trust22	impatient, desire, anxious
trust20	ambition, desire, wanted

4.4 Learning Sequential Information from Sub-Emotions (Δ-BoSE)

Sequential information is valuable to discover discriminative patterns in natural language processing tasks. We can interpret this information in terms of emotions and sub-emotions as finding emotional variability and regularities among the posts of users. Unfortunately, this information is not naturally captured by approaches based on a bag of terms or concepts, such as the previously described BoE/BoSE strategies, which lose the order of the occurrence of emotions. In this section, we describe different sequential strategies adapted to take advantage of our proposed ideas to model emotions and sub-emotions. In Section 4.4 we introduce **Δ-BoSE** (Aragon et al. 2021), where we extract new features based on statistical values to capture the emotional variability through the sequence at chunk level. This strategy allows us to exploit our proposed ideas for emotions and sub-emotions by a wide range of traditional classifiers such as SVMs. Then, in Section 4.5, we go a step further by using deep learning techniques to model the emotional variability adapting convolutional and recurrent neural networks with attention to find patterns at local and global levels of the masked text.

Δ-BoSE

The main idea of **Δ-BoSE** is that users express their emotions differently and that this could reveal a valuable pattern to improve their identification. In other words, users that present signs of self-harm tend to expose greater emotional variability than a healthy person. Following this hypothesis, we propose a new representation to capture this variability in the emotions by extracting some statistical features from the original BoSE representation.

We named this representation **Δ-BoSE**. The following steps describe the construction of Δ-BoSE; Figure 4.10 graphically shows this process.

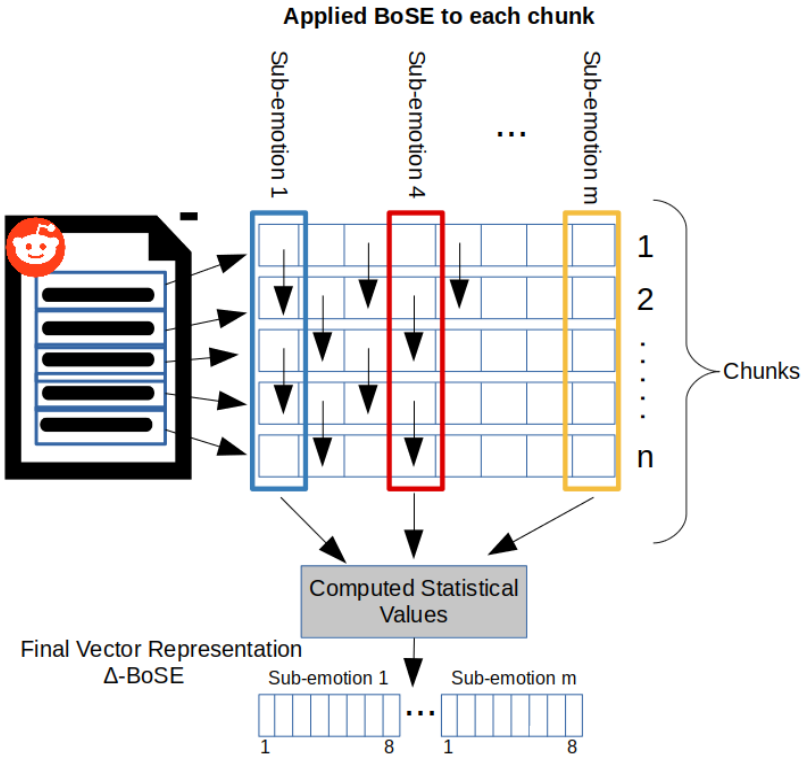

Figure 4.10: Construction of the Δ-BoSE representation. First, BoSE is obtained for each part of the document; then, statistical values are computed for each sub-emotion creating a new vector representation.

1. Process the post history of each user using a chunk by chunk framework similar to the e-Risk competition. Where the post history is divided into n equal parts, each of which is called a chunk. Therefore, consider from now on that we have n chunks of text for each user.

2. Compute for each chunk the BoSE representation as described in Subsection 4.3.1, then consider the sequence of chunks as a sequence of documents represented with BoSE.

3. For each chunk represented under BoSE, expand the representation of each m dimension (sub-emotion) by a vector of n values $\vec{S_i} = \langle w_{i,1}, \cdots, w_{i,n} \rangle$, where $w_{i,j}$ indicates the weight of sub-emotion S_i in the chunk j as determined by Formula 4.2.

4. Transform each sub-emotion vector $\vec{S_i}$ into a vector $\Delta\vec{S_i}$ that models the temporal variability of the sub-emotions. For this, we represent each sub-emotion by a $\Delta-$vector of

Table 4.7: F_1**,** ***Precision*** **and** ***Recall*** **results over the positive class.**

Task	Method	F1	Precision	Recall
anorexia	BoSE	**0.70**	**0.91**	0.57
anorexia	Δ-BoSE	0.67	0.86	**0.60**
depression	BoSE	**0.63**	**0.65**	**0.61**
depression	Δ-BoSE	0.53	0.54	0.52
self-harm	BoSE	**0.62**	**0.78**	**0.51**
self-harm	Δ-BoSE	0.58	0.72	0.48

the following eight statistical values that capture its changes through the n-chunks sequence: mean(μ), sum($\sum$), max-value(max), min-value(min), standard deviation(σ), variance(σ^2), average($\bar{x}$), and median($\tilde{x}$). With these statistical values we create a new vector $\Delta\vec{S_i} = \langle \mu, \sum, max, min, \sigma, \sigma^2, \bar{x}, \tilde{x} \rangle$ that represents the changes of the sub-emotion S_i in the post history of the user.

5. Concatenate the Δ-vectors from all sub-emotions in one single vector of size $8 \times m$, where m is the number of sub-emotions (see Figure 4.10).

One of the main benefits of Δ-BoSE representation is that it makes it possible to observe the emotional change in users. The latter significantly improves the interpretability of the model, which becomes highly relevant in medical domains. To appreciate this, in Figure 4.11 we compare the occurrences of certain sub-emotions over time (i.e., through the 10 chunks) in the control group (colored in orange) and the mental disorder group (colored in blue). We selected some of the top sub-emotions based on their chi-squared value for each task. These signals indicate the average occurrence of the sub-emotions in all users from each group. We observe that the control group presents fewer changes or peaks through time than the mental disorder group. Thus suggesting that emotional variability may exist and that it could be further exploited through emotion-based representations and Machine Learning methods to identify certain types of users.

Finally, our experimental evaluation in Table 4.7 shows that, although Δ-BoSE is slightly below the performance of BoSE, it offers better interpretability and is still better than some baselines, especially in anorexia and self.harm we can observe results similar to the best baseline.

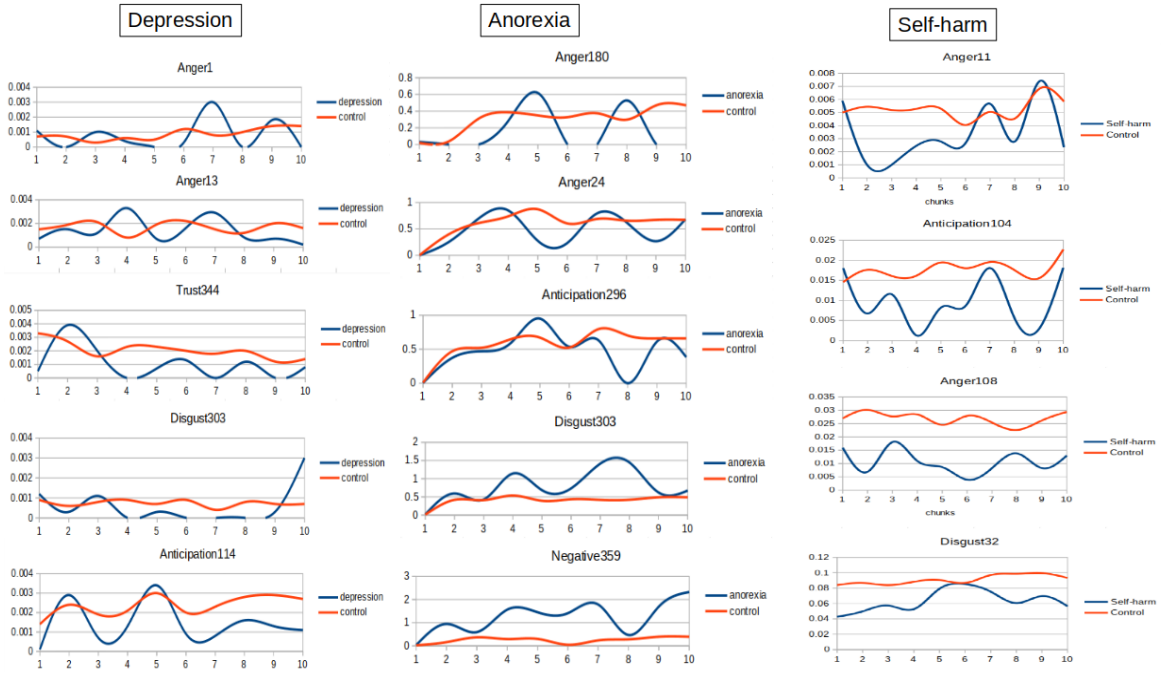

Figure 4.11: Comparison of the emotional signals between control and mental-disorder groups. X-axis represents the chunks (time span) and Y-axis represents the average value of the sub-emotion at each chunk.

Table 4.8: Depression data set used for experimentation, where P indicates the positive users and C is used for control users.

Data set	Train		Test	
	P	C	P	C
Depression'20	214	1493	40	49

4.5 Deep Learning for emotion patterns

To this point, we have only used our proposed sub-emotions (Section 4.2.1) to build representations using the framework of *bag of features*, which consist in building emotion vectors that are fed into standard classifiers. However, some sequential patterns will be hardly captured by this *bag of emotions*. This is because the order of occurrence of the emotions is lost once they are codified into the representation. In spite that some sequential information can be captured by extracting sub-emotion n-grams or by using Δ-BoSE, there are better approaches to capture long dependencies and complex patterns among sequences of user emotions (M. Aragon et al. 2020).

In this section, we show how to adapt deep neural networks to take advantage of their

capacity to automatically learn sequential patterns. The key idea is to see the documents as sequences of sub-emotion vectors associated with each word. This step is similar to what we described in Section 4.2.3 and we shown in Figure 4.9, where we treated the document as a sequence of sub-emotion labels. The representation of documents as a sequence of sub-emotion vectors allows us to feed them into Convolutional and Recurrent Neural Networks with Attention just as if they were processing standard sequences of word embeddings. The advantage is that we can capture emotional patterns at local and global levels at the same time to better model users with a mental disorder. For the following experiments, we changed the depression data set for a new one created by the eRisk organizers. For this data set, each user filled a standard BDI questionnaire (Beck et al. 1961), which contains 21 questions that allow to assess the level of severity of the depression. In the original task, the organizers asked the participants to predict, for each user, the possible answers to each input of the questionnaire. In contrast, for this study, we exclusively consider a binary prediction task, i.e., to distinguish between positive and control users. In particular, the positive class is composed of users that obtained 21 points or more in the final result of the questionnaire (presence of moderate or severe depression), whereas the control class is formed by the rest of the users, having 20 points or less in their final result. The training set for this collection is the previous depression data set from eRisk 2018. Table 4.8 shows how classes distribute within this data set.

4.5.1 Adapting Convolutional Neural Network and BoSE

Now that we have our embeddings for the sub-emotions, we want to take advantage of the contextual information presented. For this purpose, our next experiments consist of the usage of neural networks. More specifically a Convolutional Neural Network (CNN, refer to section 2.3.1). A CNN is an algorithm designed to recognize patterns. These patterns are numbers contained in a vector, in our case, sub-emotion embeddings. The CNN is a neural network that applies convolutional layers to local features. These convolutional layers, also named kernels or filters, can transform the large input data into local features. We use these filters, multiply its values element-wise with the original matrix, then sum them up. A full convolution consists of sliding the filter over the whole matrix and to the convolutional operation for each element.

For our problem, we explain this process in Figure 4.12 and through the following steps:

1. Represent each sub-emotion with an embedding vector. This embedding vector corresponds to the centroid obtained when we cluster the sub-emotions.

2. Use a convolutional neural network architecture for feature extraction of the sub-

emotions inspired in (Kim 2014). We represent this step in Figure 4.12 with the filters in blue and red.

3. Obtain different feature maps for each region and concatenate them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns.

4. Apply a sigmoid to classify the vector.

The intuition about this network is to see the post history as a sequence of sub-emotion embeddings. We look for the convolution taking one or two sub-emotions at once since our filter sizes are 1 and 2 (we empirically search for the best number of filters). In this process, we choose the maximum value of the result from each filter vector. We can think of filter sizes as unigrams and bigrams since they are finding patterns at this level of the neighborhood.

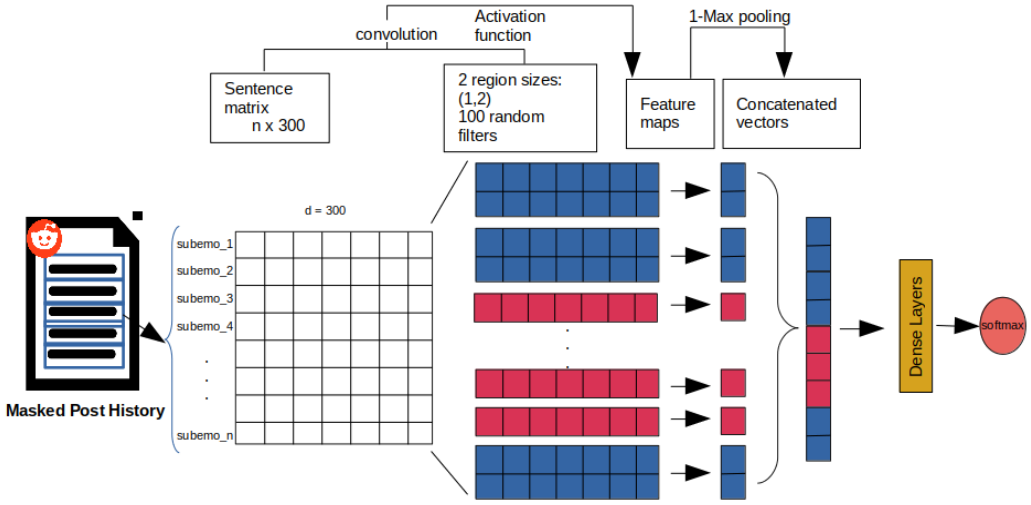

Figure 4.12: Diagram of the Convolutional Neural Network Model. We used 100 random filters of size 1 and 2 to extract the local features.

The purpose of the following experiment is to know if using the sub-emotion embeddings allows obtaining better results than the regular word embeddings and our previously BoSE representation. Table 4.9 presents the results of this experiment. We observe that using BoSE with a CNN obtains better performance in two of three task in comparison with traditional embeddings in terms of $F1$ score and recall, this could be due to the fact that with the filters, sequences are found throughout the text that confirm the presence of one of these disorders. Whereas using only BoSE obtains a better precision, which confirms the relevance of feature extraction supported with a CNN.

Table 4.9: CNN F_1, *Precision* and *Recall* results over the positive class in the different mental disorders data sets.

Task	Method	F1	Precision	Recall
anorexia	CNN-glove	0.67	0.93	0.52
anorexia	CNN-word2vec	0.66	**0.95**	0.51
anorexia	BoSE	0.70	0.91	0.57
anorexia	CNN-BoSE	**0.76**	0.77	**0.75**
depression	CNN-glove	**0.61**	0.56	**0.68**
depression	CNN-word2vec	0.58	0.55	0.60
depression	BoSE	**0.61**	**0.57**	0.64
depression	CNN-BoSE	0.59	0.56	0.62
self-harm	CNN-glove	0.57	0.54	**0.59**
self-harm	CNN-word2vec	0.56	0.54	0.58
self-harm	BoSE	0.62	**0.78**	0.51
self-harm	CNN-BoSE	**0.65**	0.73	**0.59**

4.5.2 Adapting Recurrent Neural Network and BoSE

Recurrent Neural Networks (RNN) are neural networks that are ideal for sequential information such as text. In our scenario, the idea is to capture sequential patterns of sub-emotions expressed by users, which can be achieved because the RNN has a feedback loop that allows inspection of longer dependencies among sub-emotions. The proposed strategy based on RNNs process an input sequence sub-emotion by sub-emotion instead of word by word, which helps to preserve the information about past sub-emotions on the sequence. This information flows through the hidden states and affects the decisions without revealing all learned. RNNs are dynamic systems, but they present a problem maintaining the relation of long sequences. Because the backpropagated gradient shrink at each time step and after many steps vanish (LeCun et al. 2015). To alleviate this problem, we use a special type of RNN, with explicit memory named Gated Recurrent Unit (GRU, refer to section 2.3.2). GRUs use two gates named update gate and reset gate. These gates are two vectors that help to decide the amount of information that should be passed to the output.

For this experiment, we used the sub-emotions embeddings and used as input for a BiDirectional GRU. GRU helps us to remember previous information learning the sequential structure of the sub-emotions. Furthermore, the BiDirectional GRU keeps the contextual information in both directions. We present this process in Figure 4.13. Similar to the CNN

experiment, we wanted to know if our sub-emotion embeddings obtain better results than regular word embeddings and BoSE representation. Table 4.10 presents the results of this experiment. We observe that using BoSE with an RNN again obtains better performance than traditional embeddings in terms of $F1$ score and recall, and better performance in two of three tasks in comparison with BoSE. Again using only BoSE obtains a better precision. This result shows that this strategy has high confidence when saying that a user has traces of mental disorder, but struggles in finding all the users with these symptoms. The CNN network obtains in general a better performance than using a RNN.

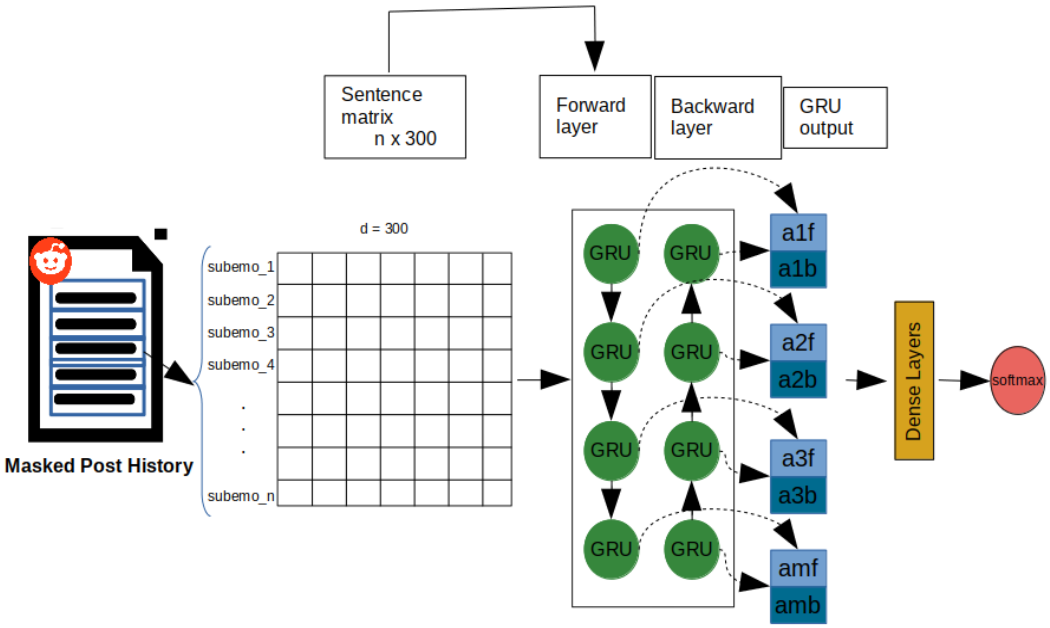

Figure 4.13: Diagram of the Recurrent Neural Network Model. We used a bi-GRU network and combining the outputs before the decision making.

4.5.3 Adapting Attention to the Sub-Emotions

Although GRUs can capture the long-range dependencies in sequences of words, the information of the whole sequence is forced to be encoded into the fixed-size hidden vector. Determining the size of the hidden vector is not trivial, and sometimes when an input sequence exceeds the average length it could not be well codified. Another problem is that we can not give more importance to some of the input sub-emotions depending on the context. Then, the following question arises: *Is there a way we can keep all the relevant information of the input sentence?*. First, we need to remember the intuition behind previous representations. In BoSE, we used a conventional method like tf-idf with a count vectorizer. This method finds features from the users' posts by doing a keyword extraction. Some words

Table 4.10: RNN F_1, *Precision* and *Recall* results over the positive class in the different mental disorders data sets.

Task	Method	F1	Precision	Recall
anorexia	RNN-glove	0.65	0.93	0.51
anorexia	RNN-word2vec	0.65	**0.95**	0.49
anorexia	BoSE	**0.70**	0.91	**0.57**
anorexia	RNN-BoSE	0.69	0.91	0.55
depression	RNN-glove	0.58	0.59	0.57
depression	RNN-word2vec	0.57	0.62	0.53
depression	BoSE	0.61	0.57	**0.64**
depression	RNN-BoSE	**0.62**	**0.63**	0.60
self-harm	RNN-glove	0.57	0.54	**0.59**
self-harm	RNN-word2vec	0.57	0.62	0.53
self-harm	BoSE	0.62	**0.78**	0.51
self-harm	RNN-BoSE	**0.64**	0.70	**0.59**

(sub-emotions) are more helpful in determining the class of the posts than others. However, when we use this method, we lost the sequential structure of the text. With the GRU network, while we can learn from the sequence structure, we lose the ability to give different weights to the more important words. Then, *could we combine both of them?* The answer is affirmative, the authors in (Bahdanau et al. 2015) came up with a simple but elegant idea. They suggested considering all the input words in the context but also adding relative importance to each word. This idea is called "Attention".

The attention mechanisms are parts of the networks that separately weigh the hidden states each time the network process an input item (sub-emotion vector), instead of only using the final RNN hidden-state to encode the whole sequence (Bahdanau et al. 2015). The advantage of having a mechanism (attention) to see how the recurrent hidden states are evolving after each input sub-emotion, is that it makes it easier to find higher-level patterns among input sub-emotions and highlight only those hidden states that contribute more to the representation. In this experiment, we added the ability to give higher weight to more important sub-emotions using an attention mechanism. This mechanism extracts sub-emotions that are important in the sentence and add the information of those sub-emotions. Then, we multiplied each sub-emotion score with their GRU output obtaining a weight according to their importance. In Figure 4.14 we present this whole process. We also experimented with a CNN before the GRU and the attention layer to capture local features before the sequence

learning.

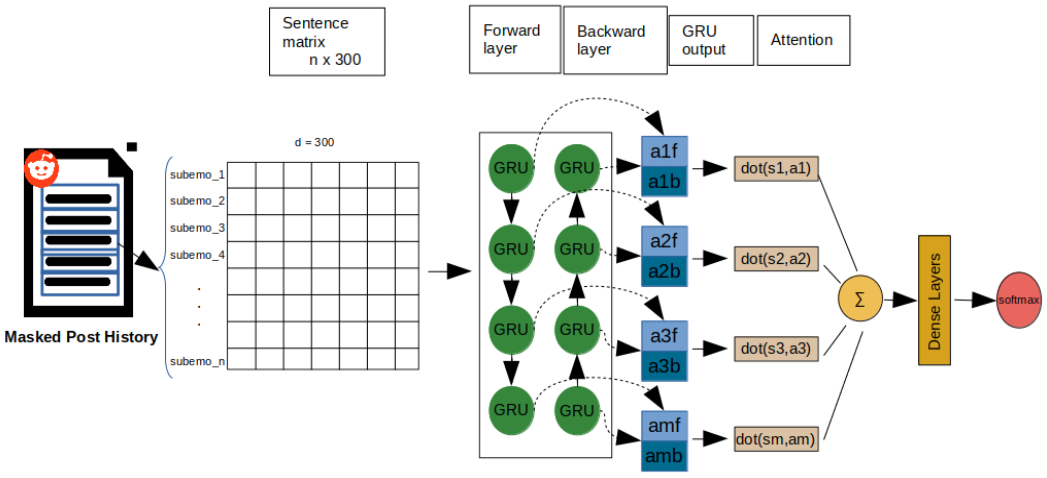

Figure 4.14: Diagram of our bi-GRU model with Attention. We add the attention after the concatenation of both the forward and backward output.

For this experiment our goal was to corroborate whether adding attention improves the results of our aforementioned approaches. Table 4.11 presents the results of this experiment. We observe that using BoSE with attention mechanism obtains a better performance that normal FT embeddings. Again using only BoSE obtains a better precision.

Analysis of the results. For the analysis, we want to know what the attention model captures. For this, we extracted the weights given to each sub-emotion on different sequences. We selected examples of sequences with a high probability of being positive cases and extracted the weights assigned to them. Table 4.12 presents the sub-emotion with the highest mean and variance value in their weights. In Figure 4.15, we also present some examples of these sequences. The shading represents the weight given to the sub-emotion, a darker shade means a higher weight. We can appreciate in these examples that the weight of each sub-emotion depends on their surrounding context. For example, if we analyze the sub-emotion "anticipation16" that is related to life experiences and events, in the first sequence, the weight assigned is high because it is close to a context of worries, mistakes, and incidents. But if we analyze the second sequence, the weight is lower because is closer to a context of gain, growth, and home. It is also interesting to notice that this sub-emotion has low variance in their value. If we analyze the sub-emotion "positive43", we can appreciate that it has high weight value in general but also has high variance depending on the context. Using an attention model allows capturing the importance of the sub-emotions in the post history using the context around it. These weights in the sub-emotions help the model to learn patterns related to people with a mental disorder.

Table 4.11: Attention F_1, *Precision* and *Recall* results over the positive class in the different mental disorders data sets.

Task	Method	F1	Precision	Recall
anorexia	RNN-Att	0.66	**0.95**	0.51
anorexia	BoSE	0.70	0.91	0.57
anorexia	RNN-Att-BoSE	**0.72**	0.93	0.59
anorexia	CNNRNN-Att-BoSE	0.71	0.72	**0.70**
depression	RNN-Att	0.50	**0.67**	0.40
depression	BoSE	0.61	0.57	**0.64**
depression	RNN-Att-BoSE	**0.63**	0.64	0.61
depression	CNNRNN-Att-BoSE	0.54	0.52	0.56
self-harm	RNN-Att	0.58	0.56	0.60
self-harm	BoSE	0.62	**0.78**	0.51
self-harm	RNN-Att-BoSE	**0.66**	0.72	0.62
self-harm	CNNRNN-Att-BoSE	**0.66**	0.64	**0.69**

Table 4.12: Mean and variance value for the attention weights in the sub-emotions.

Sub-emotion	Mean	Sub-emotion	Variance
positive43	0.1375	trust108	0.0082
joy16	0.1335	positive58	0.0051
trust108	0.1260	positive23	0.0045
negative54	0.1258	surprise65	0.0034
fear38	0.1250	fear470	0.0033
positive23	0.1198	negative62	0.0032
anticipation16	0.1144	positive43	0.0030
positive209	0.1133	joy27	0.0029
negative310	0.1086	positive240	0.0029
surprise65	0.1046	anticipation344	0.0028

Table 4.13: Significance of pairwise differences in output between submissions (using F1-score asthe reference metric).

Task		BoSE	CNN-BoSE	RNN-BoSE	Att-BoSE	Baseline
anorexia	BoSE	-	***	=	**	*
anorexia	CNN-BoSE		-	***	**	***
anorexia	RNN-BoSE			-	**	**
anorexia	Att-BoSE				-	**
depression	BoSE	-	*	=	*	**
depression	CNN-BoSE		-	*	**	*
depression	RNN-BoSE			-	*	*
depression	Att-BoSE				-	***
self-harm	BoSE	-	*	*	**	***
self-harm	CNN-BoSE		-	=	*	***
self-harm	RNN-BoSE			-	*	*
self-harm	Att-BoSE				-	***

For an extra analysis, we perform a pairwise significance comparison of the F1-scores (according to approximate randomization test (W. Noreen 1989)). We compare our different BoSE approaches between them and the best baseline of each type (CNN, RNN, Att). Table 4.13 shown this comparison, for the notation we use the following thresholds: '=' (not significantly different: $p > 0.5$), '*' (significantly different: $p < 0.05$), '**' (very significantly different: $p < 0.01$), '***' (highly significantly different: $p < 0.001$). The results suggest that the approaches have a significant difference with their baselines, and depending on the task some BoSE representations are better than another.

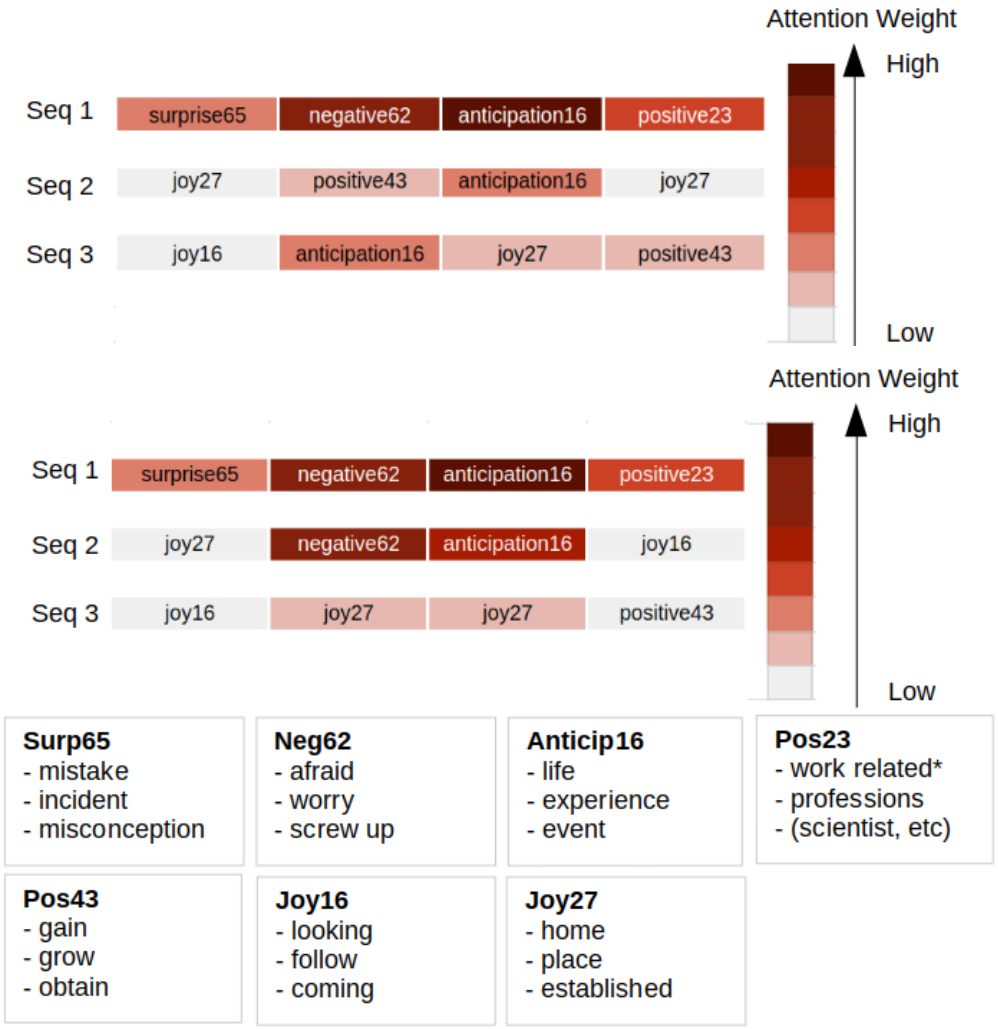

Figure 4.15: Examples of weighted sequences of sub-emotions with different contexts. Each sequence corresponds to the label of the sub-emotion assigned to each word in a sequence of words. The lower part shows the topics related to each sub-emotion.

CHAPTER 5

WHAT THEY SAY AND HOW THEY SAY IT?: THEMATIC AND STYLE REPRESENTATIONS

In order to analyze the information shared by social media users from various views, we propose to represent each word from their posts by using *three different embedding vectors*, aiming to emphasize their thematic, emotional, and writing style contexts, respectively. In chapter 4 we emphasis on the emotional channel, while in this chapter we will focus on presenting the other two channels considered and the following subsections describe each of these representations; thematic and style channels.

5.1 Thematic Embeddings

For this channel, since our aim is to capture the thematic information that is related to each word, we use vanilla GloVe embeddings (Pennington and Socher 2014). These embeddings can capture the topic or semantic similarity among words. One important limitation is that these embeddings did not take the context of the words during the test phase. For example, take the word *"bank"*. This word has different meanings in different contexts. However, Glove will give us the same vector for different contexts during the training and stay the same after. These embeddings are also called static embeddings.

To overcome these limitations, we also use for the thematic channel the embeddings from BERT (Devlin et al. 2019) (refer to section 2.3.4). A pre-train deep bidirectional representation inspired from the Transformers architectures. With BERT embeddings we can capture the thematic context from each word. For example, for our experiments, we use both separately and evaluate which one contributes the most in the multi-channel representation.

5.2 Proposed Style Embeddings

We propose a new representation based on the writing style of the users that allow capturing their writing variability. This representation complements the emotion-based representation described in the previous chapter and the thematic representation from previous sub-section.

In this channel, the representation of the words aims to capture some aspects of the writing style of social media users. The intuition behind capturing the style is that users with a mental disorder tend to talk more often and differently about the events in the past or the uncertainties in the future than healthy users. To capture this kind of information we devise a new word representation inspired in the FastText (refer to section 2.1.2) vectors (Bojanowski et al. 2016), where the idea is to weigh the contribution of each char n-gram according to its discriminative value as measured by the chi^2 distribution. Historically, char level n-grams help capturing writing style and FT use the characters for their embeddings. Figure 5.1 depicts the whole process, which consists of two main modules.

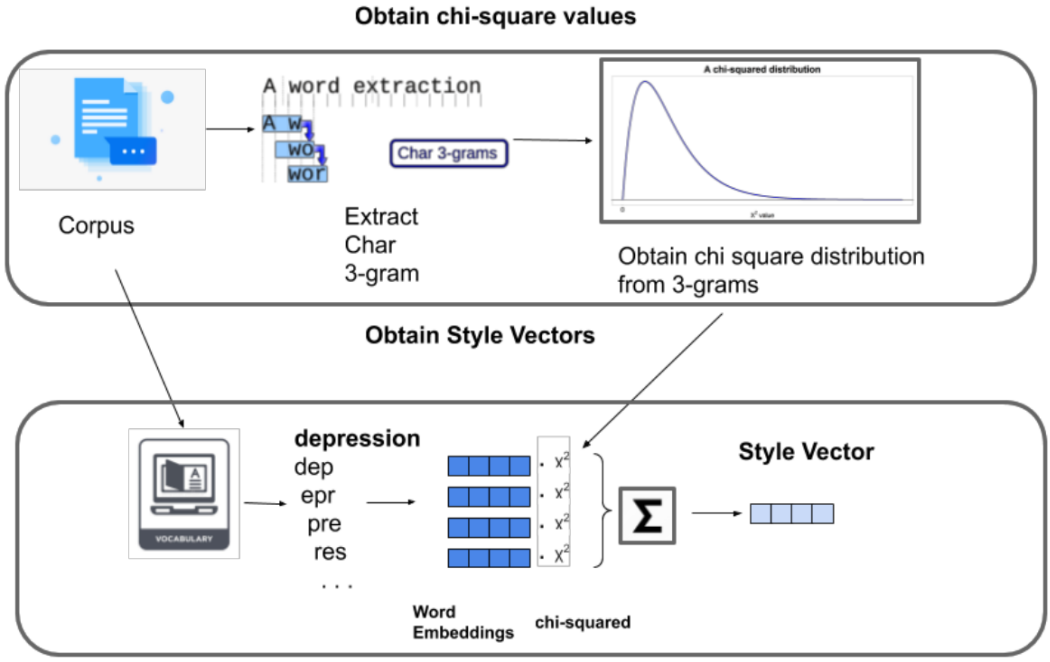

Figure 5.1: Diagram of the generation of style embeddings. The first step is to obtain the chi^2 values, then weight the vectors.

The first module uses the corpus of the task at hand to compute the relevance of each sub-word. To do that, it first divides the users' posts in all their char 3-grams, and then it computes their chi^2 distribution according to the two given classes (positive and healthy users).

For each n-gram (term), we obtain a corresponding chi^2 score that indicates if the document class has influence over the n-gram's frequency. With this approach, we want to capture the most important n-grams and use them to weight the words.

On the other hand, the second module builds the word embeddings combining the previously extracted char n-grams. It selects each word from the users' posts and divides it into char 3-grams. Then, for each 3-gram, it computes its embedding $\mathbf{c_i}$ using FastText. Finally, it obtains the embedding vector of the word by applying a weighted sum of the vectors of its char 3-grams, considering as weights their chi^2 values. We can express this formally as:

$$\mathbf{S_w} = \sum_{i=1}^{n} \mathbf{c_i} \cdot \chi_i^2 \tag{5.1}$$

Where:

- $\mathbf{S_w}$ is the final style vector for each word w.
- $\mathbf{c_i}$ represents the vector of each n-gram.
- χ_i^2 is the chi^2 value of each n-gram.

Take for example the word "depression", its style-based embedding is obtained by the weighted sum of the vectors corresponding to its character 3-grams "dep", "epr", ..., "ion".

It is important to notice that the style embeddings are similar for words that have similar spelling rather than meaning. For example, words in superlative resemble each other, or regular verbs in past tense, or words with the same root. Take for instance the word "mental" some of their closest words are "dental", "mentality" and "decremental". For a more detailed discussion of how the style embeddings differentiate from the original semantically-oriented embeddings, in Figure 5.2 we show the similarity of eight different pairs of words related to mental health. It is clearly appreciated that style vectors find high cosine similarity between word sharing affixes.

5.3 What does each individual channel capture?

A reasonable question would ponder how different are the word embeddings for the different channels. To offer a glimpse of this, first, we selected some of the words with the highest information gain. Then, we computed their word embeddings for the three channels. Finally, we obtained their closest words using the cosine similarity:

$$Cos\Theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|} \tag{5.2}$$

Figure 5.2: Similarities of several word pairs using the style and the original embeddings.

where, $\mathbf{a} \cdot \mathbf{b} = \sum_i^n a_i b_i = a_1 b_1 + a_2 b_2 + ... + a_n b_n$ is the dot product of the two vectors. Table 5.1 presents the obtained results. For each query word we can observe that the three channels offer very different information, that could later be used to improve the detection of users suffering from a mental disorder. For example, for the word *"depressed"*, the thematic channel captures topics related to insecurities or concerns, whereas the style channel retrieves some word variations such as depressing and depressants, and the emotion channel includes some negative adjectives as unhappy and demotivated.

5.4 Experimental Settings

Similar to the emotion representation, for the evaluation of the style and thematic channel we used the data sets of Anorexia '19, Depression '20, and Self-harm '20.

5.4.1 Pre-processing

Users' posts tend to contain a lot of noisy text and irrelevant information for the detection of a mental disorder. Thus, the application of pre-processing techniques is required to allow the classifier to focus on the key information and to be able to obtain reliable results. The first pre-processing step was to normalize the text by lowercasing all words and removing special characters like URLs, emoticons, and #; the stopwords were kept.

Table 5.1: Examples of the closest words for five query words according to the three considered channels.

Query Word	Thematic	Style	Emotion
mental	psychiatric	dental	autistics
	psychological	mentality	psychosocial
	retardation	mentalism	psychopathy
	illness	mentality	behavioral
	disorders	decremental	sociological
depressed	weak	depressedly	unhappy
	distressed	depressants	demotivated
	insecure	anti-depressive	uninterested
	worried	depressing	fatigued
	disturbed	depressive	troubled
therapy	treatments	aromatherapy	therapeutic
	therapies	hydrotherapy	medicine
	psychotherapy	radiotherapy	hydrotherapy
	chemotherapy	inmuno-therapy	clinical
	medication	physical-therapy	pharmacy
medications	medicines	dedications	anesthesia
	prescription	adjudications	ibuprofen
	drugs	meditations	paracetamol
	antidepressants	dedication	analgesia
	pills	edication	promethazine
compulsions	obsessions	envisions	obsessive
	phobias	fusions	compulsive-disorders
	compulsion	inclusions	obsessive
	preoccupations	visions	paranoid
	fascinations	emisions	personality-disorders

5.4.2 Classification & Predictions

We separate each post history into N parts. We select N empirically testing recommended sizes of sequences in the literature, i.e., $N = \{25, 35, 50, 100\}$. For training, we process each part of the post history as an individual input and train the model. For the test, each part receives a label of 1 or 0; then, if the majority of the parts are positive, the user is classified as showing a mental disorder. The main idea is to consistently detect the presence

of major signs of depression, anorexia, or self-harm through all the user posts.

5.4.3 Baselines

As main baseline we employed a traditional Bag-of-Words representation, considering words as well as word n-grams (size 1, 2, and 3). We also consider a bag of character trigrams, a common approach for style analysis. For these two approaches, we selected the features using a tf-idf weighting and the chi^2 distribution X_k^2. In addition, we consider some baselines based on deep learning approaches, using a CNN, a Bi-LSTM, and a state-of-the-art approach for text classification based on a Bi-LSTM with an attention layer. All these neural networks used 100 neurons, an ADAM optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN, we used 100 random filters of sizes 1, 2, and 3. We also add a BERT model with a fine-tuning over the training data set (similar to previous experiments we explore recommended parameters for the networks). Additionally, the obtained results are compared against the top-three participants of the eRisk evaluation tasks. For all these comparisons, we considered the F_1 score over the positive class, which was suggested as the golden standard by the organizers of eRisk (Losada, F. Crestani, et al. 2018).

5.5 Evaluation

In this study, we evaluate our new representations and compare them with different baselines. Table 5.2 presents the results in terms of F_1 score over the positive class to detect Anorexia (eRisk'19), Depression (eRisk'20) and Self-harm (eRisk'20). We organize the results in baseline methods, and our proposals of single channel.

From this evaluation we can observe that most of our proposals outperform the baseline results. First of all, some single-channel representations obtain a considerable improvement in comparison with baselines, in particular those based on style and emotion information. Surprisingly, the performance of deep learning models applied over word-based representations is somehow poor and closer to traditional approaches like BoW; we presume this could be attributable to the small size of the data sets in conjunction with their large thematic diversity. Something interesting to notice is that CNN networks obtains a better performance than RNN networks. The latter could be due to the fact that CNN networks search for the presence of specific local information important for the detection of these disorders. We can also notice that using a GMU module improves the results in comparison with a simple concatenation on anorexia and depression, and competitive results for self-harm.

Table 5.2: F1 results over the positive class in three eRisk's tasks.

Method	Anor	Dep	SH
Baselines			
BoW-unigrams	0.67	0.58	0.50
BoW-Ngrams	0.66	0.57	0.50
Bag of char 3grams	0.67	0.58	0.53
RNN-word2vec	0.65	0.57	0.55
CNN-word2vec	0.66	0.60	0.56
RNN-Attention	0.66	0.50	0.58
Our methods: Single-channel			
RNN Thematic$_{Glove}$	0.65	0.58	0.57
CNN Thematic$_{Glove}$	0.67	0.61	0.57
CNN Thematic$_{BERT}$	**0.77**	**0.64**	0.60
RNN Style	0.73	0.63	0.64
CNN Style	0.70	**0.64**	**0.65**
RNN Emotion	0.69	0.62	0.64
CNN Emotion	0.76	0.59	**0.65**

From these round of experiments, we highlight the following:

1. Most single-channel representations outperformed the baselines for the detection of mental disorders in online environments.

2. Depending on the task different information channel obtains better performance. For example, for anorexia the thematic channel obtains the best performance. For depression, the style and thematic information obtain equal performance. In self-harm, the style and emotion channel are the best results.

3. These results opens an opportunity to combine the types of information and improve the detection.

5.5.1 Completeness and diversity analysis

For closing up this analysis, we investigate how diverse but also complementary, in terms of the information they capture, are these channels.

To measure their complementarity, we used the Maximum Possible F1 (MPF) metric. This measure is a variation of the Maximum Possible Accuracy (MPA), which is defined as

the quotient of the correctly classified instances over the total number of test instances. For this analysis, we considered an instance as correctly classified if at least one of the channels classified it correctly.

To measure the diversity of predictions, we use the Coincident Failure Diversity (CFD) metric (Tang et al. 2006). This measure focuses on obtaining the error diversity among the predictions of the three channels. The minimum value of this measure is 0, obtained when all the channels misclassified the same subset of users; on the contrary, the maximum value is 1, which is obtained when the classification errors are different for each channel.

Table 5.3 presents the MPF and CFD scores for each task, measured over the positive class. For the MPF values, we can appreciate an improvement in comparison with our best reported results (last row of table). These results indicate that the channels are complementary to each other. However, the CFD results indicate a low error diversity for the three channels, suggesting that for many users there are one correct and two incorrect decisions. Combining the obtained insights from both results, it is clear that there is still room for improvement, but to achieve it, it will be necessary to explore more channels as well as other fusion strategies.

Table 5.3: MPF and CFD results in the three tasks, measured over the positive class

Metric	Anorexia	Depression	Self-harm
CFD	0.3150	0.175	0.2788
MPF	0.8872	0.7568	0.8442
Best Emotion	0.76	0.62	0.65
Best Thematic	0.77	0.64	0.60
Best Style	0.73	0.64	0.65

CHAPTER 6

LEARNING A MULTI-CHANNEL REPRESENTATION

As we previously mentioned, depending on the task different types of information obtain better performance. These results open an opportunity to learn to combine these channels and improve the detection of mental disorders. For this problem, we propose the multi-channel learning paradigm that aims to create a new data representation by combining two or more channels of information. In this work, to accomplish this, we select a well established feature extractor strategy, namely, a Convolutional Neural Network (CNN)[1].

6.1 A dynamic channel fusion approach

A CNN is an algorithm designed to recognized local patterns and, thus, it aligns very well with our hypothesis that in order to identify a user suffering from a mental disorder, it is enough to detect a set of thematic, stylistic or emotional evidences distributed throughout all of his/her posts. One of the challenges of this work is the problem of fusing information. A simple solution is to concatenate the representations of each channel into one vector or perform an operation like adding or taking the product. However, using these operations assume that all channels have the same relevance, which is usually not the case. For our case, depending on the mental disorder, one or more channels might have differently and complementary information. Figure 6.1 shows our multi-channel architecture, which is summarized as follows:

[1] We also evaluated a Recurrent Neural Network (RNN) with an attention mechanism to learn the relation between the channels, but experiments show a better performance for the CNN alone. We discuss more of this in the analysis of the results section.

1. Represent each word in the user' posts with an embedding vector. This embedding vector corresponds to the channel that is being analyzed. The main idea is to process in parallel the three channels described in previous chapters.

2. Use a CNN for feature extraction, as described in (Kim 2014). To extract relevant unigram and bigrams, we employ filters of size equal to 1 and 2, which are represented in Figure 6.1 in red and blue respectively.

3. Obtain for each channel different feature maps for each region and concatenate them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns.

4. Use a GMU module and linear layers to learn the relations between each channel feature vector. Then, apply a sigmoid activation function to classify the final vector with the information of the three channels.

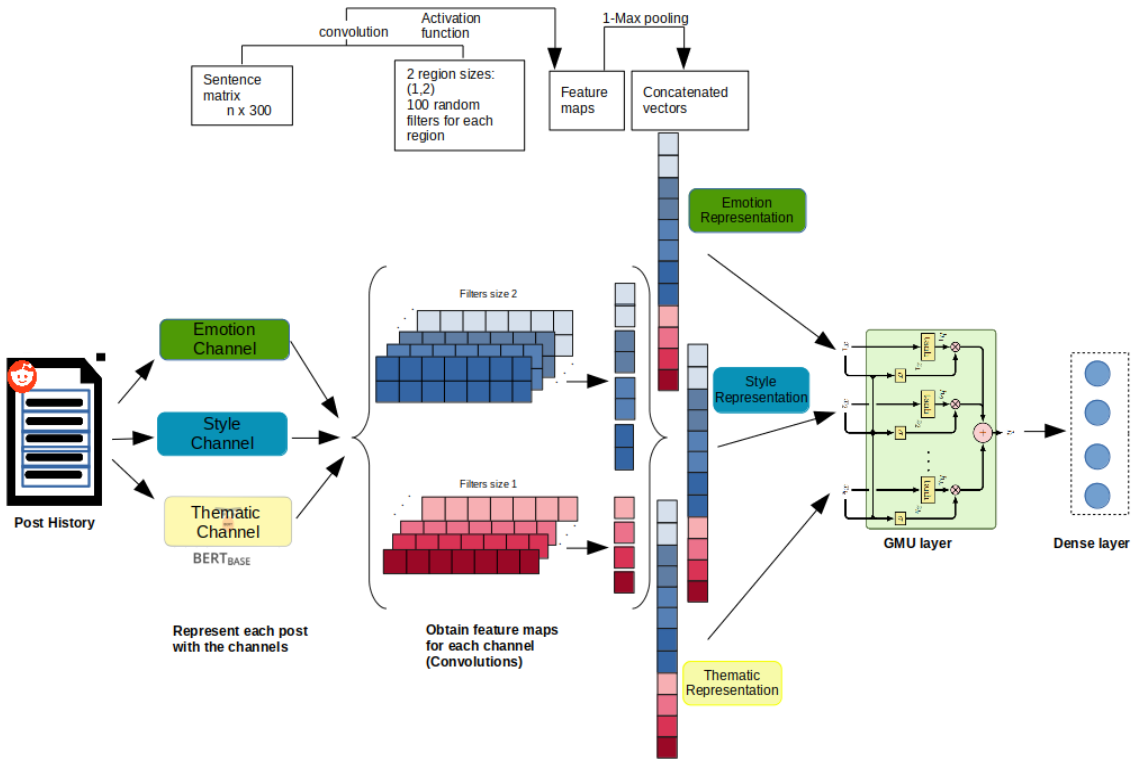

Figure 6.1: Diagram of the Convolutional Neural Network Model for the creation of the multi-channel representation. We used 100 random filters of size 1 and 2 to extract the local features.

6.2 Evaluation

For experimental settings, we used the same pre-processing and classification configuration mentioned in previous experiments (chapter 5). Table 6.1 presents the results in terms of F_1 score over the positive class to detect Anorexia (eRisk'19), Depression (eRisk'20) and Self-harm (eRisk'20). We organize the results in three groups: baseline methods, our proposal but limited to only one channel, and our original proposal using all information channels.

From this evaluation we can observe that the full-channel representation is clearly the performant approach in this comparison, then suggesting the pertinence of combining different types of information. From these experiments, we highlight the following observations:

1. The use of a multi-channel representation improves the results than only using one type of information. This result confirms our intuition that learning the fusion is very relevant to capture signs of mental disorders in users.

2. Using a GMU improves the results of anorexia and depression detection in comparison with a simple vector concatenation. Automatically weighting the information helps to create a better representation of the posts and the users.

6.3 Comparison against the eRisk participants

To set a context regarding this competition consider that a total of 54 models were submitted to the anorexia detection task and 57 to the self-harm detection task in eRisk-19 and 20 editions (Losada, 0. F. Crestani, et al. 2019; Losada, F. Crestani, et al. 2020). It is important to mention that the participants focused on obtaining early and accurate predictions of the users, while our approach focuses exclusively on determining accurate classifications.

Table 6.2 shows how our best approach (i.e., the CNN model with 2 and 3 channels) compares against the top places at the eRisk 2019 and 2020 evaluation tasks. We observe that our approach achieves competitive results in both tasks, first place for Anorexia and tied in first place for Self-harm. For the depression task, organizers changed the evaluation task, and thus we cannot directly compare these results against the participants[2].

[2]To clarify this point, our approach focuses on a binary classification task, i.e., to discriminate between users suffering from depression and control users, while, on the other hand, the eRisk task considered the assessment of the level of depression severity for each user.

Table 6.1: F1 results over the positive class in three eRisk's tasks. St = style channel, Em = emotional channel and Th = Thematic channel. For the RNN notation: Entry = combination of channels at the input layer; BA = combination of the channels before the attention layer, and AA = combination of the channels after the attention layer.

Method	Anor	Dep	SH
Baselines			
BoW-unigrams	0.67	0.58	0.50
BoW-Ngrams	0.66	0.57	0.50
Bag of char 3grams	0.67	0.58	0.53
RNN-word2vec	0.65	0.57	0.55
CNN-word2vec	0.66	0.60	0.56
RNN-Attention	0.66	0.50	0.58
BERT	0.77	0.62	0.67
Martinez-Castaño et.al	0.77	0.64	**0.75**
Our methods: Single-channel			
RNN Thematic$_{Glove}$	0.65	0.58	0.57
CNN Thematic$_{Glove}$	0.67	0.61	0.57
CNN Thematic$_{BERT}$	0.77	0.64	0.60
RNN Style	0.73	0.63	0.64
CNN Style	0.70	0.64	0.65
RNN Emotion	0.69	0.62	0.64
CNN Emotion	0.76	0.59	0.65
Our methods: Multi-channel (simple concatenation)			
RNN St+Em (Entry)	0.75	0.65	0.66
RNN St+Em (BA)	0.76	0.48	0.56
RNN St+Em (AA)	0.76	0.55	0.56
CNN St+Em	0.78	0.65	0.72
CNN St+Em+Th$_G$	0.78	0.67	0.71
CNN+RNN St+Em	0.71	0.64	0.71
CNN+RNN St+Em+Th$_G$	0.72	0.66	0.70
CNN St+Em+Th$_B$	0.78	0.66	**0.75**
Our methods: Multi-channel (GMU combination)			
CNN+GMU St+Em+Th$_G$	0.79	0.67	0.73
CNN+GMU St+Em+Th$_B$	**0.82**	**0.70**	0.73

Table 6.2: F_1**,** $Precision$ **and** $Recall$ **results over the positive class**

Task	Anorexia 2019			Self-harm 2020		
Metric	F1	P	R	F1	P	R
1st place	0.71	0.64	0.79	**0.75**	0.82	0.69
2nd place	0.68	0.77	0.60	0.62	0.62	0.62
3rd place	0.68	0.67	0.68	0.62	0.59	0.65
Our best	**0.82**	**0.88**	0.76	**0.75**	0.69	0.82

For a further analysis of these results, Figure 6.2 presents a boxplot of the F_1, precision, and recall scores of all participants from both tasks. The green **X** represents our best result of the combination of channels. In the Figure, we appreciate that our results are in the highest quartile for both tasks. These results indicate that our multi-channel representation obtains competitive results in comparison with the participants of the anorexia and self-harm detection tasks.

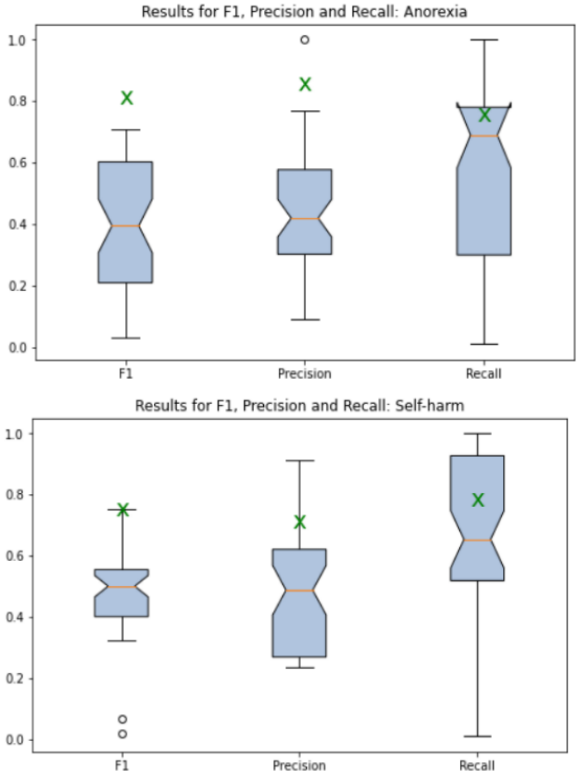

Figure 6.2: Boxplot of the F1 scores for anorexia (upper part) and self-harm (bottom part), where the green X represents our best approach.

6.4 Analysis of the Method

6.4.1 Contribution of each information channel

We want to understand how the GMU units are weighting the relevance of each channel. For this experiment, we analyze the gates z_i of the GMU module selecting the posts in the test set and average the gate activations per channel. This z_i value represents the contribution of the feature calculated from x_i to the overall output of the unit. Figure 6.3 presents the results for the three channels, where each row already takes into account the average of all posts per mental disorder. In general, it can be notice how the activations for each channel are different depending on the mental disorder. For example, for depression and anorexia BERT (thematic) channel has the highest value, and for self-harm, the style channel has the highest value. Something interesting is that the thematic channel presents the highest variation. With the highest value in anorexia and the lowest value in self-harm. This variation indicates that the posts of users who suffer from anorexia are probably more homogeneous than those who suffer self-harm, whose content is presumed to be much more heterogeneous.

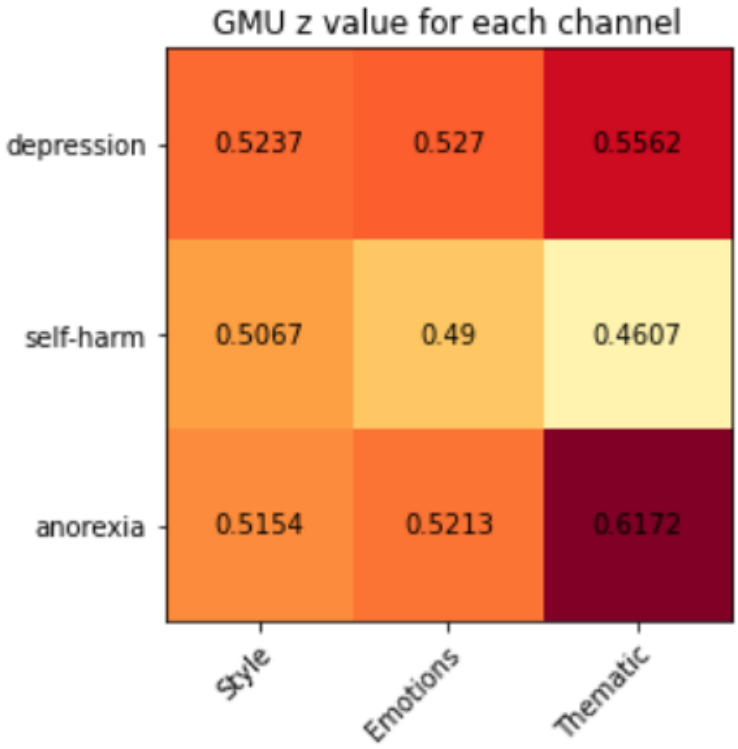

Figure 6.3: Average proportion of GMU unit activations for the channels over the test set. The Figure presents the avarage z_i value for each channel and mental disorder

.

For an extra analysis of the GMU activations, Table 6.6 presents the posts with the highest z_i value for each mental disorder and each channel. We can appreciate posts related to personal opinions and concerns even when the topics are not directly related to mental disorders. For example, for depression the posts are related to users playing video games, while self-harm focus on personal concerns in different topics.

Table 6.3: Posts with highest z_i value for each mental disorder and channel. The posts are related to personal opinions and concerns.

Channel	Post
Anorexia	
style	*"...produced in a manner to sell more rather than staying true to what the lore is..."*
emotion	*"... actually made to portray stories that are in touch with the issues that the current world is going through..."*
thematic	*"...they fit in with they are trying to achieve because of this. Is trying to push social justice..."*
Depression	
style	*"I use to try and humanize myself. In their eyes is fake nightmares and say that thinking of talking to them stopped the nightmare..."*
emotion	*"I think I know who they're playing, but who knows. Sent to losers by eliminated."*
thematic	*"... civil war losers, top gg. its all gaming after this, losers..."*
Self-harm	
style	*"... thank you I appreciate it very much and I hope to be able to serve to the best of my ability no matter what others think..."*
emotion	*"...me Im usually fine, Im a christian and love that show supernatural too also lucifer. What the hell guys you re giving me a bad name..."*
thematic	*"...she yelled out loud, it scared me so bad that the knife slipped and I ended up slicing into my thumb."*

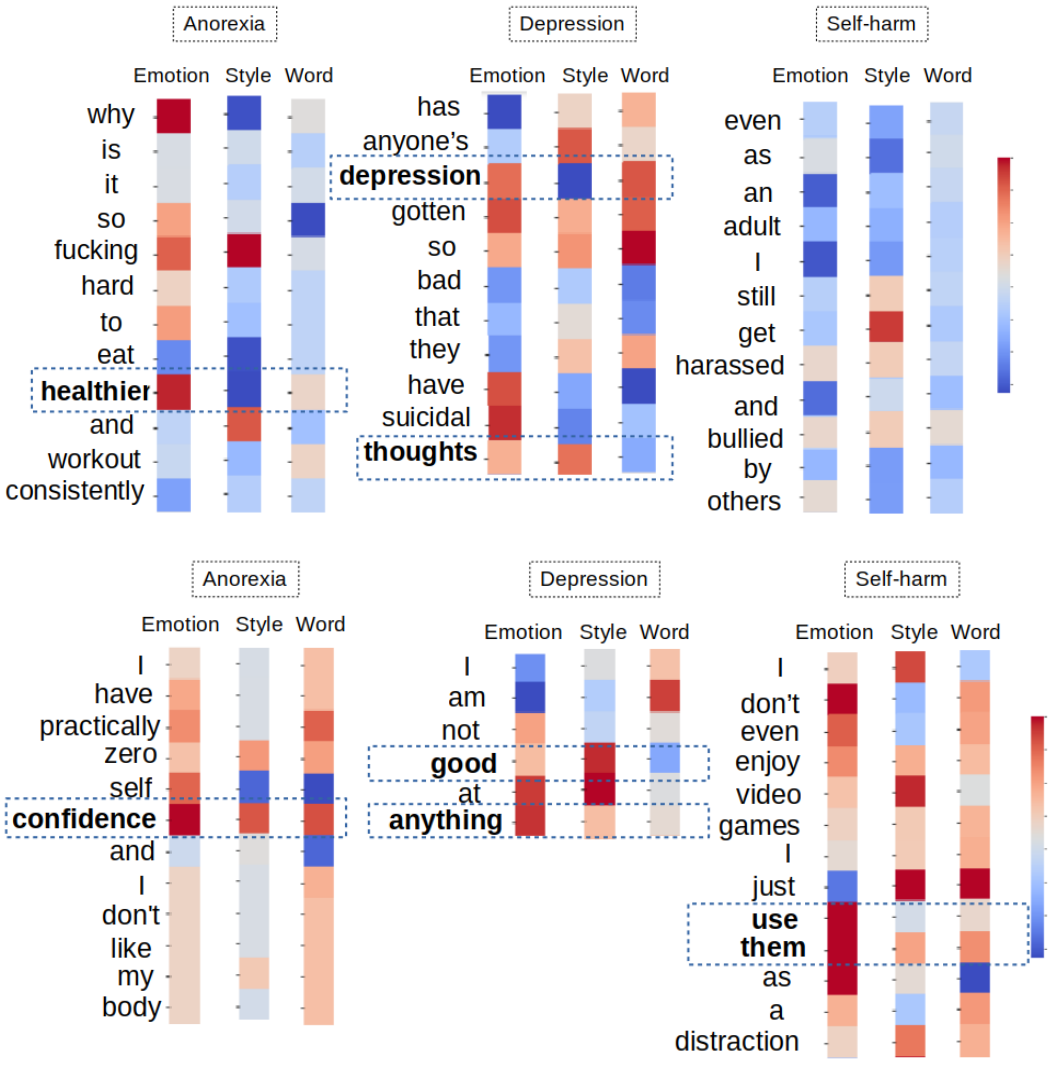

Figure 6.4: Saliency obtained with the different type of channels for the positive class. The red color indicates a high value and the bluer color represents a lower value.

6.4.2 Qualitative analysis of each channel

This analysis aims to investigate to what extent each information channel captures different information. For visualizing this, we used a strategy inspired by the back-propagation in vision. This strategy named as saliency, measures how much each input contributes to the final decision, which is obtained using its first derivative. More formally, we analyze the output of our model, and computed the saliency $\mathcal{S}_j = \sum_{x_l \in \mathbf{x}_j} \left|\left| \frac{\partial \tilde{y}_i}{\partial x_l} \right|\right|$ of the three channels with different sample texts that were extracted from users with a mental disorder. We define the saliency $\mathcal{S}_j$ of a specific word as the average of the magnitude of the gradient of each component in the embedded representation (J. Li et al. 2016). We present these saliency

maps in Figure 6.4, where the red color indicates a high value and the bluer color represents a lower value. We notice that depending on the channel and the context, the saliency is higher for different words. For example, the word "healthier" has high saliency for the emotion channel, but it is low for the style channel. For the word "thoughts", the saliency is high for the emotion and style channel but is low for the thematic channel. See how the word "confidence" gets high saliency for the three channels (at different level of importance), but the rest of the words in the context have different saliency.

For a further analysis of the saliency, we compute for each task the average saliency for each word. Then, we select the words with the highest value, avoiding words with less than ten occurrences in the documents (we want to avoid words that appear few times but obtain high saliency and do not generalize the task). In Table 6.4 we show these words. Note that for each channel, the words are different, but they have a close relationship with each task. For example, for self-harm, the style words are disorders, addicted or tension, while for the emotion channel the highest words are killed, bother, or cutting. We can conclude that the channels contribute with individual and contrasting information between each other, and this information helps to improve the detection of mental disorders in online environments.

Table 6.4: Words with the highest saliency for each task and each channel. The words are different, but they have a close relationship with each task.

Task	Thematic	Style	Emotion
Anorexia	mouth	stereotype	frusterating
	dicyclomine	stigmatized	likeness
	young	promise	mean
	stomach	lunch	jealous
	meal	anonymous	far
Depression	pills	mysteries	frusterating
	need	borderline	likeness
	together	professional	pathological
	suicide	ditsy	mean
	depressive	stigmatized	life-call
Self-harm	disinterest	addicted	killed
	shyness	tensions	bother
	accused	disorders	codependent
	dies	homework	cutting
	friendzone	crime	boyfriend

6.4.3 On the predicted posts' probabilities

As we previously mentioned, decisions about users are generated by combining the predictions made for each post. To better understand this process, Figure 6.5 presents the distributions of the posts' prediction values for the three tasks considered. In this case, the prediction values are nothing other than the probabilities of the posts to belong to the positive class in accordance to the classifier used.

Figure 6.5 shows some interesting information. For the self-harm and depression tasks it is possible to observe that most of the prediction values for the positive users' posts are higher than 0.6, thus suggesting the suitability of our approach to detect evidence of the presence of those mental disorders. Nevertheless, control users' posts also show some high probabilities, which may indicate that their topics, emotions and style overlap to some degree with those from the positive users. On the other hand, for the anorexia task it can be observed that control users are clearly distinguishable from positive users, since most of their posts have little or no probability of belonging to the positive class. However, in this case not all posts from positive users show high probabilities, perhaps due to their greater thematic and style diversity. In summary, the figure shows that for the detection of depression and self-harm the false positives are the main concern, while for the detection of anorexia the false negatives are the key issue.

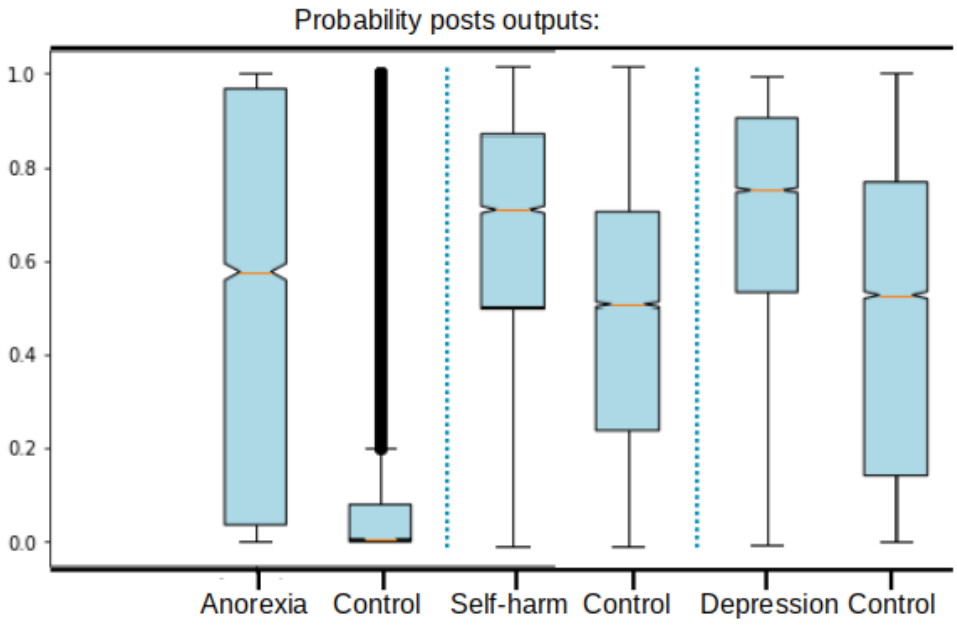

Figure 6.5: Distribution of the posts' predictions. The prediction values are the probabilities of the posts to belong to the positive class.

To shed some light on what is being detected by the approach proposed, Figure 6.6 shows the distribution of the decisions in the post history of some users. Taking as a reference the answers to the BDI questionnaire used for depression detection, we selected the users with the three highest scores, the three lowest, and three users with borderline scores to be depressive, which correspond to severe depression, normal status, and borderline clinical depression, respectively. Then, we represent each post of their history with a dark blue color if the post obtains a high probability of belonging to a depressive user, and with white color if the probability was closer to a control user. We can appreciate how the post history of the users with a high score of depression is darker than the users with a low score. It is interesting to notice that the borderline users present different distribution on their post history. This difference could be due to the diversity in the topics they write or express in their posts.

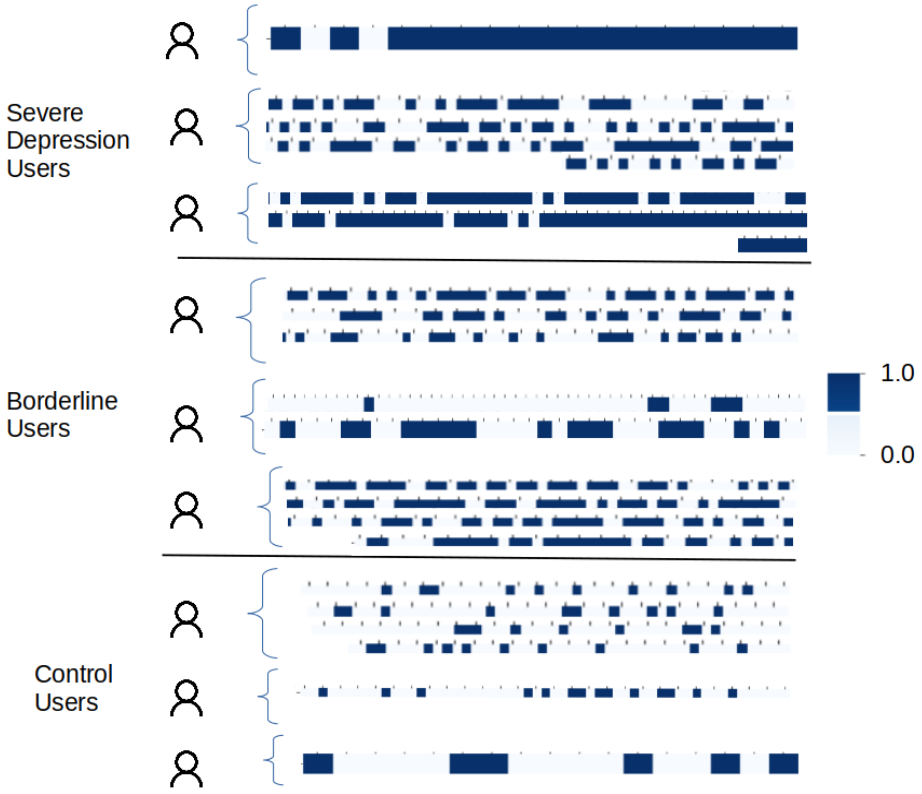

Figure 6.6: Output of the different posts of being a positive class. We can appreciate how the post history of the users with a high score of depression is darker than the users with a low score.

6.5 Multi-Modal BERT (MMBT) Experiments

It is a recently proposed supervised multimodal bitransformer model for classifying images and text Kiela et al. 2019. The MMBT model starts with pre-trained BERT weights, and takes their contextual embeddings as input. These contextual embeddings are obtained as the sum of the segment, position, and token embeddings of each word. Then, the model weights them and project each of the embeddings to a token input. Although proposed for only two modalities, this architecture can be generalized to any number of modalities, assigning a different segment id to each of them.

In the original work, the authors took contextual embeddings as input, learned the weights, and projected each image's embeddings to a dimensional token input. Instead of image embeddings, we used the sequence of the words for the fine-tuning with our different channels as embeddings. Once we fine-tunned our model with the channels, we took advantage of the contextual information learned for classifying the users' posts. For this purpose, we used the first output of the final MMBT layer as input to a Convolutional Neural Network (CNN) for feature extraction, and then add a dense layer for achieving the classification. Figure 6.7 presents the general architecture of this approach.

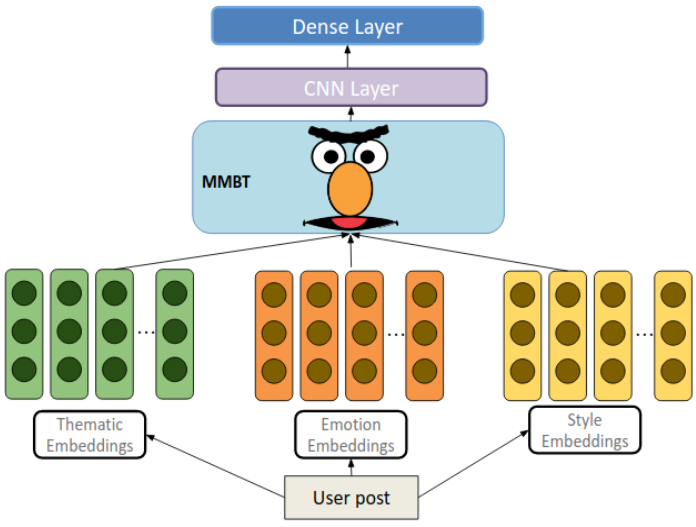

Figure 6.7: General diagram of the Model 1: Multimodal BERT with vectors of three channels, then a CNN layer, and a classification layer.

6.5.1 Combining information using multiple BERTs

The authors of MMBT Kiela et al. 2019 noted that their method is compatible with scenarios where not every modality is present and can be generalized to an arbitrary number

of modalities. Then, for our second model, we proposed to train different BERTs and fine-tunning them with our channels separately. After the training, similar to our first approach, we used the first output of the final layer of each BERT and concatenate them as input for a CNN layer, we will refer to this model as BERT-CNN. In Figure 6.8, we present the general diagram for this process.

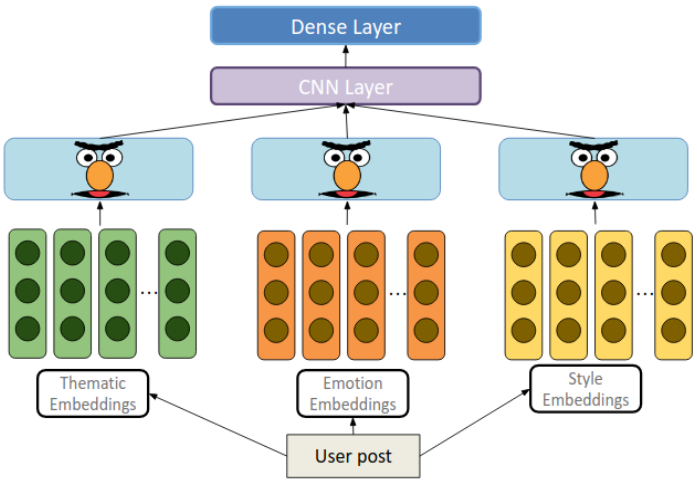

Figure 6.8: General diagram of the BERT-CNN model: Each channel separately enters to a BERT model, then join vectors feed a single CNN layer, and a classification layer.

For our third model, our approach is similar to the previous one, however, instead of concatenating the vectors and using a single CNN layer, we separate them into different convolutional layers and used the output for a dense layer. With this approach, the model obtains for each channel different feature maps of each region and concatenates them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns, and after that combine the information. The hypothesis that the local information per channel is important and should be extracted before it is combined, we call this model BERT-3CNN. Figure 6.9 presents the general diagram for this model.

Motivated by the results of the GMU module in different multimodal tasks, our fourth model takes advantage of it. That is, after the feature extraction, we implement a Gated Multimodal Unit (GMU) module to learn the relations between each channel feature vector. Then, apply a dense layer to classify the final vector with the information of the three channels, we call this model BERT-GMU. Figure 6.10 describes the process for this model.

Table 6.5 present the results in terms of F_1 score over the positive class. We compared the new results with our previous results obtained for every single channel and multi-channel. From this evaluation, we observe that most of our proposals outperform the baseline results. For our representations based on the fusion of information, their performance is higher in

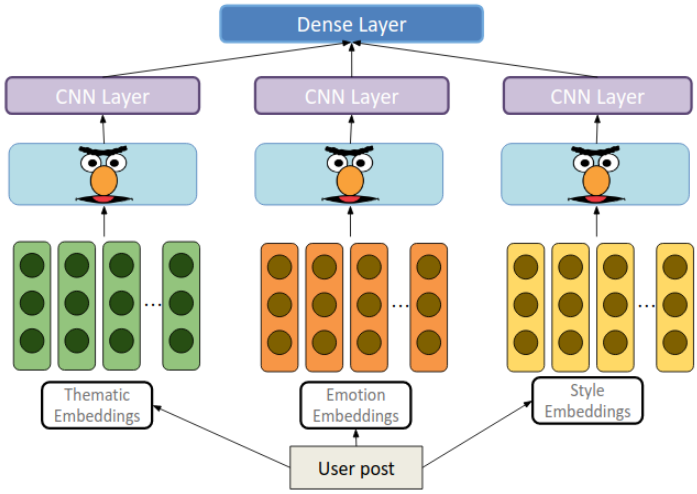

Figure 6.9: General diagram of the BERT-3CNN model: Each channel separately enters a BERT model and a CNN layer, then their outputs are concatenated and fed to a single classifier.

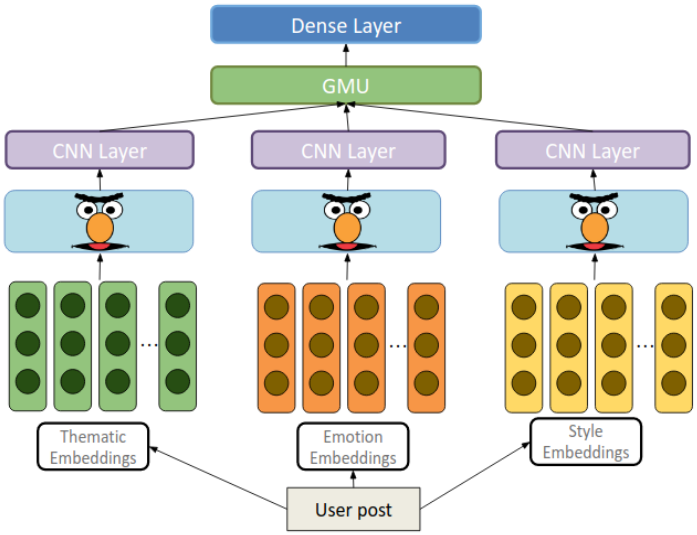

Figure 6.10: General diagram of the BERT-GMU model: Each channel separately enters a BERT model and a CNN layer, then their outputs are combined by a GMU module, and fed to a single classifier.

comparison with the other models, suggesting the relevance of combining the information from different channels. Something interesting to notice is that all models using multiple BERTs outperformed the multimodal BERT model in F1, indicating that, for this particular task, and with this way of representing the channels, it is better to represent each channel independently and combine them later. In general, the model that use the GMU module showed the best average performance, which suggest that weighting the information helps to

create a better representation of the posts and the users.

From these experiments, we highlight the following observations:

1. The use of a multi-channel representation improves the results than only using one type of information. This result confirms our intuition that learning to combine different types of information is very relevant to capture signs of mental disorders in users.

2. The architectures using multiple BERTs obtained better performance than the multi-modal BERT.

Table 6.5: F1 results over the positive class in three eRisk's tasks.

Method	Anor	Dep	SH
Single-channel			
Thematic$_G$	0.67	0.61	0.57
Thematic$_B$	0.77	0.64	0.60
Style	0.70	0.64	0.65
Emotion	0.76	0.59	0.65
Multi-channel (simple concatenation)			
St+Em+Th$_G$	0.78	0.67	0.71
St+Em+Th$_B$	0.78	0.66	**0.75**
Multi-channel (GMU combination)			
St+Em+Th$_G$	0.79	0.67	0.73
St+Em+Th$_B$	**0.82**	**0.70**	0.73
Multi-channel (MMBT+GMU)			
MMBT	0.76	0.65	0.65
BERT-CNN	**0.82**	**0.70**	0.70
BERT-3CNN	0.80	0.68	0.70
BERT-GMU	0.81	**0.70**	0.73

6.5.2 Contribution of each information channel

For this analysis, we obtained the gates' z_i values of the GMU module corresponding to the test set posts. Figure 6.11 presents the results for the three tasks, where each value already takes into account the average of all posts. For anorexia, we can appreciate that the thematic information contributes the most to the final decision, followed by emotion and style information. For depression, we can observe that the thematic information is also the

most important and the value for the style information is higher than the emotional value. Finally, for the self-harm task, the thematic information obtains the lowest value and the style information the highest. In general, it can be noticed how the activation for each channel are different depending on the mental disorder. Something interesting is that the thematic channel presents the highest variation, with the lowest value in self-harm and the highest value in anorexia. We think that this variation indicates that the posts of users who suffer from anorexia are probably more homogeneous than those who suffer self-harm.

For an extra analysis of the GMU, Table 6.6 presents the posts with the highest z_i value for each channel. We can notice that the posts are related to personal opinions, different topics, and in general express negative emotions even when they are not directly related to mental disorders. Take for example the emotion information, where the post is related to regrets in life and feelings. The thematic post where the model captures the contexts of the user expressing having a good day and then ruined it for the new conversation topic.

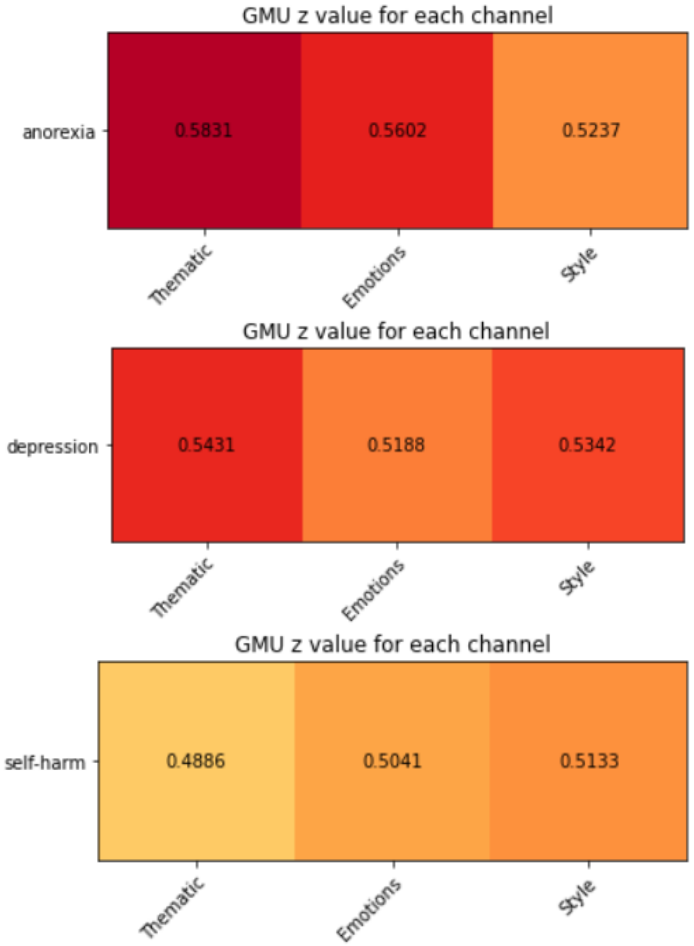

Figure 6.11: Average z_i value for the three mental disorders over the test set instances.

Table 6.6: Posts with highest z_i value for each channel over the depression task.

Channel	Post
thematic	*"...I have no idea what either of them were trying to communicate tbh i was* ***having a really good day*** *and then you had to bring up hughes..."*
emotion	*"...take the chance and have no* ***regrets*** *in life, its always better to know if the other person* ***feels*** *something so that you are not wasting your time..."*
style	*"...these days parents don't know a whole lot about* ***mental illness*** *they were told it was nothing so that's all..."*

Part IV

General Conclusions

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this study, we explored the detection of Anorexia, Self-harm, and Depression in users of social media environments utilizing a novel multi-channel representation. Each information channel focuses on extracting information that corresponds to the users' writing style, emotions, and thematic interests. Our proposal can automatically learn how to combine these features and extract the most relevant information from each channel. Results suggest that combining different types of information helps in the detection of users with mental disorders, outperforming traditional and state-of-the-art baselines, and being strongly competitive with the performance of top eRisk participants. Our analysis of the method yields that the complementarity in the types of information is important to get a better picture and understanding of the posts written by the users. Below we detail the main conclusions of this research work:

- BoSE outperformed the BOW representation in three tasks, indicating that considering emotional information is more relevant for the detection of depression, self-harm, and anorexia in online communications than only considering the use of words. The use of sub-emotions as features helps to improve the results from a representation that only considers coarse emotions. This result confirms our hypothesis that such an approach is more effective to capture subtle changes of emotions in users with mental disorders.
- The inclusion of a dynamic analysis over the sub-emotions, called Δ-BoSE, improved the detection of users that presents signs of a mental disorder, showing the usefulness of considering the changes of sub-emotions over time. It is worth mentioning the simplicity and interpretability of both representations, then creating a more straightforward analysis of the results.

- Most single-channel representations outperformed the baselines, especially noting that style and emotional information are more relevant for the detection of mental disorders in online environments than the thematic aspect without the contextual information.
- The use of a multi-channel representation improves the results than only using one type of information. This result confirms our intuition that learning the fusion is very relevant to capture signs of mental disorders in users. Using a GMU improves the results of anorexia and depression detection in comparison with a simple vector concatenation.

Scope and Limitations:

- This study aims to detect Anorexia, Self-harm, and Depression in users of social media environments using a multi-channel representation.
- This study also presents some limitations, mainly because these data sets are observational studies and we do not have access to the personal and medical information that is often considered in risk assessment studies. There are also some limitations given by the nature of the data, as they might differ from users at risk who do not have an online account or decided to not make their profiles public. In addition, in the data sets of anorexia and self-harm, it is not guaranteed that the users annotated as positive are actually at risk because the annotation was performed just from reading a few posts.

Future work and further comments:

- For future work, we want to explore more sophisticated combination techniques that could improve the results and understanding of mental disorders detection. We note that most of the analysis of mental disorders has been made for the English language, then, one of our interests lies in the expansion of this study to the Spanish language.
- The capability to model the behavior of users using their social media data presents an opportunity for future wellness facilitating technologies. This kind of technology can serve as warning systems that provide wide-area analysis and information related to a mental disorder respecting user privacy. This information could include the presence of mental disorders in certain areas, and the authorities could decide to create professional assistance or emotional support, that the users will decide whether to take or not.
- We believe that it is important to mention when we analyze social media content, we may have concerns regarding individual privacy or certain ethical considerations.

These concerns appear due to the usage of information that could be sensitive, given the personal behavior and emotional health of the users. The experiments and usage of this data are for research and analysis only, and the misuse or mishandling of the information is prohibited.

7.1 Academic Production

The following list shows the research papers derived from this thesis, or those where ideas from this work have been used:

- Journal papers:
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Luis C González, Manuel Montes-y-Gómez. Information fusion for mental disorders detection: multimodal BERT against fusioning multiple BERTs. Sociedad Española del Procesamiento del Lenguaje Natural. SEPNL (2022) (Under review)
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Luis C González, Manuel Montes-y-Gómez. Approaching what and how people with mental disorders communicate in social media – Introducing a multi-channel representation. Neural Computing and Applications. (Under review)
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Luis C González, Manuel Montes-y-Gómez. Detecting Mental Disorders in Social Media Through Emotional Patterns - The case of Anorexia and Depression. IEEE Transactions on Affective Computing. Q1. IEEE-T-AFFC doi: 10.1109/TAFFC.2021.3075638 (2021)
- Book chapters:
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Luis C González, Manuel Montes-y-Gómez. Detecting traces of self-harm on Reddit through emotional patterns. Studies in Computational Intelligence: The best of eRisk in the last 5 years. Springer-SCI (2021) (Accepted)
- Conference papers:
  - Juan S Lara, **M. Ezra Aragón**, Fabio A Gonzalez, Manuel Montes-y-Gómez. Deep Bag-of-Sub-Emotions for Depression Detection in Social Media. International Conference on Text, Speech, and Dialogue. TSD (2021)
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Luis C González, Manuel Montes-y-Gómez. Attention to Emotions: Detecting Mental Disorders in Social Media. International Conference on Text, Speech, and Dialogue. Lecture Notes in Computer Science, vol 12284. Springer, Cham. TSD (2020)
  - **M. Ezra Aragon**, A. Pastor Lopez-Monroy, Luis C. González-Gurrola, M Montes-y-Gomez. Detecting Depression in Social Media using Fine-Grained Emotions.

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT (2019)

- Shared tasks:
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Manuel Montes-y-Gómez. INAOE-CIMAT at eRisk 2020: Detecting Signs of Self-Harm using Sub-Emotions and Words. CLEF 2020, eRisk - Early Risk Prediction on the Internet. CLEF (2020)
  - **M. Ezra Aragón**, A. Pastor López-Monroy, Manuel Montes-y-Gómez. INAOE-CIMAT at eRisk 2019: Detecting Signs of Anorexia using Fine-Grained Emotions. CLEF 2019, eRisk - Early Risk Prediction on the Internet. CLEF (2019)
- Extra collaborations:
  - Marco Casavantes, **M. Ezra Aragón**, Luis C González, Manuel Montes-y-Gómez. The message is not all you need – Paying attention to posts and authors' metadata to spot abusive comments on Twitter. Knowledge-Based Systems. (2022) (Under review)
  - Proceedings of the Iberian Languages Evaluation Forum, IberLEF 2021. Editors: Manuel Montes, Paolo Rosso, Julio Gonzalo, **Mario Ezra Aragón**, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez Mellado, Jorge Carrillo-de-Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima-López, Flor Miriam Plaza-de-Arco, Mariona Taulé. CEUR Workshop Proceedings, Vol. 2943, ISSN 1613-0073.
  - **M. Ezra Aragón**, Horacio Jarquín, Manuel Montes-y-Gómez, Hugo J. Escalante, Luis Villaseñor-Pineda, Helena Gómez-Adorno, Gemma Bel-Enguix, Juan P. Posadas-Durán. Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain. SEPLN-IberLEF (2020)
  - **M. Ezra Aragón**, Miguel Á Álvarez-Carmona, Manuel Montes-y-Gómez, Hugo J. Escalante, Luis Villasenor-Pineda, Daniela Moctezuma. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain. SEPLN-IberLEF (2019)

BIBLIOGRAPHY

Aragon, M.E., AP. Lopez-Monroy, LC. Gonzalez-Gurrola, and M. Montes-y-Gomez (2021). "Detecting Mental Disorders in Social Media Through Emotional Patterns - The case of Anorexia and Depression". In: *IEEE Transactions on Affective Computing*.

Aggarwal, C.C. and C. Zhai (2012). "A survey of text classification algorithms." In: *Mining text data. Springer.*, pp. 163–222.

Aragon, M.E., AP. Lopez-Monroy, LC. Gonzalez-Gurrola, and M. Montes-y-Gomez (2020). "Attention to Emotions: Detecting Mental Disorders in Social Media". In: *International Conference on Text, Speech, and Dialogue*.

Aragon, M.E. , AP. Lopez-Monroy, LC. Gonzalez-Gurrola, and M. Montes-y-Gomez (2019). "Detecting Depression in Social Media using Fine-Grained Emotions." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1481–1486.

Arevalo, J., T. Solorio, M. Montes-y Gómez, and FA. González (2019). "Gated multimodal networks". In: *Neural Computing and Applications*, pp. 10209–10228.

Association, American Psychiatric (2013). "Diagnostic and statistical manual of mental disorders (5th ed.)" In: *American Psychiatric Association*.

Bahdanau, D., K. Cho, and Y. Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *The International Conference on Learning Representations (ICLR)*.

Basheer, I. and M. Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application". In: *Journal of microbiological methods*, pp. 3–31.

Beck, A.T., C.H. Ward, M. Mendelson, J. Mock, and J. Erbaugh (1961). "An Inventory for Measuring Depression". In: *JAMA Psychiatry 4(6)*, pp. 561–571.

Bengio Y. and Ducharme, R., P. Vincent, and C. Jauvin (2003). "A neural probabilistic language model." In: *Journal of Machine Learning Research.*, pp. 1137–1155.

Bengio, Y. (2009). "Learning deep architectures for AI." In: *Foundations and trends in Machine Learning.*, pp. 1–127.

Bengio, Y., A. Courville, and P. Vincent (2014). "Representation Learning: A Review and New Perspectives." In: *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1798–1828.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics*, pp. 135–146.

Burdisso, S., M. Errecalde, and M. Montes-y-Gómez (2019). "A Text Classification Framework for Simple and Effective Early Depression Detection Over Social Media Streams". In: *Expert Systems With Applications, Vol.133*, pp. 182–197.

Chauvin, Y. and D. Rumelhart (1995). "Backpropagation: theory, architectures, and applications". In: *Psychology press*.

Chikersal, Prerna, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme (2020). "Understanding client support strategies to improve clinical outcomes in an online mental health intervention". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16.

Cho, K., D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." In: *Conference on Empirical Methods in Natural Language Processing.*, pp. 1724–1734.

Clinic, Mayo (2018). "Anorexia Nerviosa". In: *https://www.mayoclinic.org/es-es/diseases-conditions/anorexia-nervosa/symptoms-causes/syc-20353591*.

Coopersmith, G., M. Dredze, and C. Harman (2014). "Quantifying mental health signals in Twitter". In: *Workshop on Computational Linguistics and Clinical Psychology*, pp. 51–60.

Coopersmith, G., R Leary, E. Whyne, and T. Wood (2015). "Quantifying suicidal ideation via language usage on social media". In: *Joint Statistics Meetings Proceedings, Statistical Computing Section*.

Coppersmith, G., M. Dredze, and C. Harman (2014). "Quantifying mental health signals in Twitter". In: *Proceedings of the Workshop on Computational. Proceedings of the Workshop on Computational. Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60.

Coppersmith, G., M. Dredze, C. Harman, and K. Hollingshead (2015). "From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses". In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pp. 1–10.

Coppersmith, G., C. Harman, and M. Dredze (2014). “Measuring post traumatic stress disorder in Twitter”. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 579–582.

Coppersmith, G., K. Ngo, R. Leary, and A. Wood (2016). “Exploratory analysis of social media prior to a suicide attempt”. In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pp. 106–117.

Correa, Teresa, Amber Willard Hinsley, and Homero Gil De Zuniga (2010). “Who interacts on the Web?: The intersection of users’ personality and social media use.” In: *Computers in human behavior 26.2*, pp. 247–253.

Cortes, C. and V. Vapnik (1995). “Support-vector networks”. In: *Machine learning*.

De Choudhury, M., S. Counts, and E. Horvitz (2013). “Social media as a measurement tool of depression in populations.” In: *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 47–56.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz (2013). “Predicting depression via social media”. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pp. 128–137.

Devlin, J., MW. Chang, K. Lee, and K. Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *NAACL-HTL.*, pp. 4171–4186.

Duong, C., R. Lebret, and K. Aberer (2017). “Multimodal Classification for Analysing Social Media.” In: *arXiv:1708.02099*.

Eichstaedt, J., R. Smith, R. Merchant, L. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. Asch, and H. Schwartzm (2016). “A decade into Facebook: where is psychiatry in the digital age?” In: *The Lancet Psychiatry 3, 11*, pp. 1087–1090.

– (2018). “Facebook language predicts depression in medical records”. In: *Proceedings of the National Academy of Sciences*, pp. 1535–1559.

Ernala, Sindhu K., Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury (2019). “Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals”. In: *In Proceedings of the 2019 chi conference on human factors in computing systems*.

Fisher, C. and P. Appelbaum (2017). “Beyond Googling: The Ethics of Using Patients’ Electronic Footprints in Psychiatric Practice”. In: *Harvard Review of Psychiatry*.

Funez, DG., MJ. Garciarena-Ucelay, MP. Villegas, SG. Burdisso, LC. Cagnina, M. Montes-Y-Gómez, and ML. Errecalde (2018). “UNSL’s participation at eRisk 2018 Lab”. In: *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*.

Goldberg, Y. (2017). "Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies)." In: *Graeme Hirst*.

Goodfellow, I., Y. Bengio, and A. Courville (2016). "Deep learning". In: *MIT press Cambridge*.

Group, Canadian Reference (2016). "Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Study-Adolescent Supplement (NCS-A)". In: *American College Health Association-National College Health Assessment II*.

Guntuku, SC., D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt (2017). "Detecting depression and mental illness on social media: an integrative review". In: *Current Opinion in Behavioral Sciences*, pp. 43–49.

Htait, A., S. Fournier, and P. Bellot (2017). "LSIS at SemEval-2017 Task 4: Using Adapted Sentiment Similarity Seed Words For English and Arabic Tweet Polarity Classification". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 718–722.

Huang, E.H., R. Socher, C.D. Manning, and A.Y. Ng (2012). "Improving word representations via global context and multiple word prototypes." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*, pp. 873–882.

Ji, S., X. Li, Z. Huang, and E. Cambria (2020). "Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks". In: *Neural Computing and Applications*.

Kang, K., C. Yoon, and E.Y. Kim (2016). "Identifying depressive users in Twitter using multimodal analysis". In: *In Big Data and Smart Computing (BigComp), 2016 International Conference on. IEEE*, pp. 231–238.

Kessler, R., E. Bromet, P. Jonge, V. Shahly, and Marsha. (2017). "The Burden of Depressive Illness". In: *Public Health Perspectives on Depressive Disorders*, pp. 40–66.

Khosla, M., K. Jamison, GH. Ngo, A. Kuceyeski, and Sabuncu MR. (2019). "Machine learning in resting-state fMRI analysis." In: *Magn Reson Imaging.*, pp. 101–121.

Kiela, D., S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine (2019). "Supervised Multimodal Bitransformers for Classifying Images and Text". In: *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop*.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.

Kosinski, Michal, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel (2014). "Manifestations of user personality in website choice and behaviour on online social networks". In: *Machine learning*, pp. 357–380.

Kosinski, Michal, David Stillwell, and Thore Graepel (2013). “Private traits and attributes are predictable from digital records of human behavior”. In: *Proceedings of the national academy of sciences*, pp. 29–48.

Laye-Gindhu, A. and Kimberly A. Schonert-Reichl (2005). “Self-Harm Among Community Adolescents: Understanding the "Whats" and "Whys" of Self-Harm”. In: *Journal of Youth and Adolescence*, pp. 447–457.

LeCun, Y., Y. Bengio, and G. Hinton (2015). “Deep learning”. In: *Nature 521, no. 7553*, pp. 436–444.

Li, J., X. Chen, E.H. Hovy, and D. Jurafsky (2016). “Visualizing and Understanding Neural Models in NLP”. In: *HLT-NAACL*.

Liu, N., Z. Zhou, K. Xin, and F. Ren (2018). “TUA1 at eRisk 2018”. In: *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*.

López-Monroy, Adrian Pastor, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio (June 2018). “Early Text Classification Using Multi-Resolution Concept Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)"*, pp. 1216–1225.

Losada, DE., F. Crestani, and J. Parapar (2019). “Overview of eRisk 2019: Early Risk Prediction on the Internet”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland*, pp. 1–18.

Losada, DE., F. Crestani, and J. Parapar (2018). “Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview)”. In: *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*.

– (2020). “Overview of eRisk 2020: Early Risk Prediction on the Internet.” In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*, pp. 1–18.

Masood, Razan (2019). “Adapting Models for the Case of Early Risk Prediction on the Internet”. In: *Advances in Information Retrieval. ECIR 2019. Lecture Notes in Computer Science, vol 11438.*, pp. 353–358.

Mathers, C. and D. Loncar (2006). “Projections of global mortality and burden of disease from 2002 to 2030”. In: *PLOS Medicine, Public Library of Science*.

McCulloch, W. and W. Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics*, pp. 115–133.

Medline, Plus (2022). “Mental Disorders”. In: *https://medlineplus.gov/mentaldisorders.html*.

Merikangas, KR., J. He, M. Burstein, SA. Swanson, S. Avenevoli, L. Cui, C. Benjet, K. Georgiades, and J. Swendsen (2010). “Lifetime prevalence of mental disorders in U.S.

adolescents: Results from the National Comorbidity Study-Adolescent Supplement (NCS-A)". In: *Journal of the American Academy of Child and Adolescent Psychiatry*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). "Efficient estimation of word representations in vector space." In: *Proceedings of International Conference on Learning Representations.*

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems.*

Mnih, A. and K. Kavukcuoglu (2013). "Learning word embeddings efficiently with noise-contrastive estimation." In: *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems.*

Mohammad, S.M. and P.D. Turney (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence*, pp. 436–465.

Mohammadi, E., H. Amini, and L. Kosseim (2019). "Quick and (maybe not so) Easy Detection of Anorexia in Social Media Posts". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Morin, F. and Y. Bengio (2005). "Hierarchical probabilistic neural network language model." In: *In Proceedings of the International Workshop on Artificial Intelligence and Statistics.*, pp. 246–252.

Olah, C. (2015). "Understanding lstm networks". In: *https://colah.github.io/posts/2015-08-Understanding-LSTMs/.*

Pennington, J. and C.D. Socher R. and Manning (2014). "GloVe: global vectors for word representation." In: *In Proceedings of the Conference on Empirical Methods on Natural Language Processing.*, pp. 1532–1543.

Pestian, JP., H. Nasrallah, P. Matykiewicz, A. Bennett, and AA. Leenaars (2010). "Suicide Note Classification Using Natural Language Processing: A Content Analysis in Heidelberg". In: *Biomed Inform Insights*, pp. 19–28.

Phi, M. (2018). "Illustrated Guide to LSTM's and GRU's: A step by step explanation". In: *https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step- explanation*.

Preotiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, HA. Schwartz, and L. Ungar (2015). "The role of personality, age and gender in tweeting about mental illnesses". In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pp. 21–30.

Preoţiuc-Pietro, Daniel, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras (2015). "Studying user income through language, behaviour and affect in social media." In: *PloS one 10.9*.

Qianli, M.A., S. Lifeng, C. Enhuan, T. Shuai, W. Jiabing, and C. Garrison (2017). "WALKING WALKing walking: Action Recognition from Action Echoes". In: *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2457–2463.

Ragheb, W., J. Aze, S. Bringay, and M. Servajean (2019). "Attentive Multi-stage Learning for Early Risk Detection of Signs of Anorexia and Self-harm on Social Media". In: *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland*.

Ramachandran, P., B. Zoph, and Q. Le (2017). "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941*.

Ramírez-Cifuentes, D. and A. Freire (2018). *UPF's Participation at the CLEF eRisk 2018: Early Risk Prediction on the Internet*.

Renteria-Rodriguez, M. (2018). "Salud mental en Mexico". In: *NOTA-INCyTU NÚMERO 007*.

Ríssola, E.A., M. Aliannejadi, and F. Crestani (2020). "Beyond Modelling: Understanding Mental Disorders in Online Social Media". In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal*, pp. 296–310.

Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval". In: *Information processing management*, pp. 513–523.

Sasikumar, M., S. Ramani, S. Muthu-Raman, KSR. Anjaneyulu, and R. Chandrasekar (2007). "A Practical Introduction to Rule Based Expert Systems." In: *Narosa Publishing House, New Delhi.*

Schwartz, HA., J. Eichstaedt, M. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar (2014). "Towards assessing changes in degree of depression through facebook". In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 118–125.

Tang, E.K., P.N. Suganthan, and X. Yao (2006). "Visualizing and Understanding Neural Models in NLP". In: *Mach. Learn.*

Tausczik, YR. and JW. Pennebaker (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". In: *Journal of Language and Social Psychology*, pp. 24–54.

Thavikulwat, P. (2008). "Affinity Propagation: A clustering algorithm for computer-assisted business simulation and experimental exercises". In: *Developments in Business Simulation and Experiential Learning*.

Trifan, A. and JL. Oliveira (2019). "BioInfo@UAVR at eRisk 2019: delving into social media texts for the early detection of mental and food disorders". In: *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland*.

Trotzek, M., S. Koitka, and CM. Friedrich (2018). "Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia". In: *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*.

Tsugawa, S., Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki (2015). "Recognizing depression from twitter activity". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3187–3196.

Van Rijen, P., D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, and P. Ruch (2019). "A Data-Driven Approach for Measuring the Severity of the Signs of Depression using Reddit Posts". In: *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland*.

Vaswani, A., N. Shazeer, Uszkoreit Parmar N., J., L. Jones, AN. Gomez, L. Kaiser, and I. Polosukhin (2017). "Attention Is All You Need." In: *1st Conference on Neural Information Processing Systems*.

Volkova, Svitlana and Yoram Bachrach (2016). "Inferring perceived demographics from user emotional tone and user-environment emotional contrast". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1567–1578.

W. Noreen, E. (1989). "Computer-Intensive Methods for Testing Hypotheses: An Introduction". In: *A Wiley-Interscience publication*.

Walck, C. (2007). "Hand-book on Statistical Distributions for experimentalists". In: *University of Stockholm, Internal Report SUF–PFY/96–01*.

Wang, Tao, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis (2017). "Detecting and characterizing eating-disorder communities on social media". In: *Proceedings of the Tenth ACM International conference on web search and data mining*, pp. 91–100.

Williams, T. and R. Li (2018). "An Ensemble of Convolutional Neural Networks Using Wavelets for Image Classification". In: *Journal of Software Engineering and Applications*, pp. 69–88.

Witten, I., E. Frank, M. Hall, and C. Pal (2016). "Data Mining: Practical Machine Learning Tools and Techniques." In: *Morgan Kaufmann Series in Data Management Systems*.

World Health Organization, WHO (2019). "Mental health: Fact sheet". In: *https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health*.

Xuetong, C., D.S. Martin, W.J. Thomas, and E. Suzanne (2018). "What about Mood Swings? Identifying Depression on Twitter with Temporal Measures of Emotions". In: *Companion*

Proceedings of the The Web Conference 2018, International World Wide Web Conferences Steering Committee.