Detecting Mental Disorders in Social Media using a Multichannel Representation

Mario Ezra Aragón Saenzpardo,
Adrián Pastor López Monroy,
Manuel Montes y Gómez

Reporte Técnico No. CCC-20-006
22 de octubre de 2020

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Detecting Mental Disorders in Social Media using a Multichannel Representation

Mario Ezra Aragón Saenzpardo⋆,
Adrián Pastor López Monroy† and **Manuel Montes y Gómez**⋆

Computer Science Department
⋆ Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro 1, Santa María Tonantzintla, Puebla, 72840, México

Computer Science Department
† Centro de Investigación en Matemáticas (CIMAT)
Callejón Jalisco, Valenciana, 36023 Guanajuato, GTO, México
E-mail: mearagon@inaoep.mx, pastor.lopez@cimat.mx, mmontesg@inaoep.mx

---

Abstract

Currently, millions of people around the world are affected by different mental disorders that interfere in their thinking and behavior, damaging their daily life. Timely detection of mental disorders is important to help people before the illness gets worse, minimizing disabilities and returning them to their normal life. The stigma related to mental disorders creates barriers to improve the resources that help the detection of these problems.

The most popular way for people to share information is using social media platforms, and people tend to share topics related to work issues and personal matters. People with mental disorders tend to share more about their concerns looking for some advice, support or just because they want to relieve suffering. This creates an excellent opportunity to automatically detect users that have a mental disorder and refer them as soon as possible to seek professional help.

In this work to detect mental disorders in social media, we propose: 1) different representations from the information shared by the users. For example, semantic or topic information, phonetic or writing style, and emotion information. 2) A model that automatically creates a representation combining the previous representations. With these, the model can learn to represent social media documents (a.k.a. posts) by using the combination of these different types of information. The generated representations (individual and combined) will be evaluated in different tasks related to mental disorders, for example, depression detection, anorexia detection and post-traumatic stress

disorder (PTSD). Learning to automatically combine these different types of information, creating a new representation of the social media documents, could improve the results for detecting mental disorders in comparison with state of the art approaches.

As preliminary results; we design a new representation considering emotions as information called Bag of Sub-Emotion(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. We evaluated this first representation in depression and anorexia detection. The results are encouraging; the usage of fine-grained emotions improved the results from traditional representations and a representation based on the core emotions and obtained competitive results in comparison to state of the art approaches. We also present results from a representation inspired by the emotional changes of a user, this representation combined with BoSE obtain better results than using them separately.

Contents

1 Introduction

Common mental disorders such as depression, anorexia, dementia, post-traumatic stress disorder (PTSD) or schizophrenia affect millions of people around the world [20, 21]. Most people believe that mental disorders are not usual or happen to other people, that have specific personal damage. When in fact, mental disorders are prevalent and familiar. 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 cause emotionally and physically damaged that could make people feel fear for the idea of being vulnerable to criticism, judgment or wrong opinions.

A mental disorder is a disease that causes different disturbances in the thinking and behavior in the affected person. The disturbances could vary from mild to severe, where it could result in an inability to live ordinary demands or routines in daily life. The mental problem may be related to a particular event that generated excessive stress on the person or a series of different stressful events. One or a combination of different factors like environmental stress, genetic factors, different hard life situations, could be the cause that affects people.

The National Institute of Mental Health made a study where they found that young people are more affected by one mental disorder [7]. This study found that one of every five young people are affected by at least one mental disorder. The researchers of this study also found that the percentage of someone that is suffering from a mental disorder is higher than other frequent primary physical conditions, such as diabetes or asthma. In another 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 [8]. The study also found that one of five students have depression and feel anxious, or are dealing with other mental disorder. The students also claim their health was bad or sick, and the 13% had considered suicide at least once. This presents an alarming rise in mental disorders, and the numbers of suicide are increasing. It is imperative then, to create useful approaches that are capable of detecting these mental disorders before they cause irreparable damage to many people that suffer these problems and the people that surround them.

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 [69]. Nowadays, of the people that are affected only one in five get treatment. Mental disorders increase in countries that have gone through phenomena of generalized violence or natural disasters, such as Mexico with the war against drug dealers. There are thousands of people that are direct or indirect victims, whose mental health requires appropriate and effective attention. Of the health budget in Mexico, only about 2% is destinated to mental health, when the World Health Organization, recommends 5 to 10%. Besides, 80% of the spending on mental health is used to maintain psychiatric hospitals instead of detection, prevention, and rehabilitation.

In a developed world, for many people the majority of their social life does not take place in their surroundings or immediate environment, but in reality, it takes place in a virtual world created by social media platforms like Facebook, Twitter, Reddit, or another similar platform. Social media has become a vital link for many people that live far from their loved ones like family and friends. However, psychologists express concern and have started research that suggests that the usage of social media has increased in fact that people feel lonelier, insecure and isolated than before using it, rather than increasing the connection with the people they love or care. Independently of the pros or cons that social media have, it is improbable that they are going to disappear any time soon. This presents an opportunity, to understand different mental disorders through the analysis of their social media documents and increases the chances to detect people that present signs of mental disorders and help them to guide or provide them to professional help as soon as possible [10, 9].

In posts that are shared by the users, different properties or channels from their texts could be analyzed, providing useful information to detect if some of them present signs of a mental disorder. For example, the emotions expressed in the posts, the style of writing, or the kind of topics discussed presented in the posts. This different information in texts could help to represent the information that users shared.

For this work, we proposed three main contributions: 1) new approaches to model different channels, using the information shared by users in their social media platforms, for example, new representations of emotions, style or topics. 2) The creation of a multichannel representation combining the previous channels. A model learns to automatically represent social media documents, using different channels. 3) The incorporation of sequential information to represent the documents. Multimodal representations inspire these contributions, where it is very important to discover the relationship between different modalities (text, images, voice, etc.). The proposed contributions will be evaluated in depression detection, anorexia detection, and post-traumatic stress disorder (PTSD) detection, three of the most common mental disorders with access to databases.

The remainder document organized the content as follows: Section 2 introduces a brief discussion about the related work to mental disorders. Section 3 presents the research proposal that includes the problem statement, hypothesis, research questions, objectives, and expected contributions. Section 4 contains the Methodology to accomplish the objectives and contributions that Section 3 proposes. Section 5 adds a Gantt Diagram with the work schedule to illustrate the different steps to finish the dissertation. Section 6 includes the preliminary work to support this dissertation proposal. Section 7 and 8 present the conclusions and published papers from this work. Finally, Section 9 presents and describes in detail the background with the core concepts and techniques needed for this dissertation.

2 Related Work

This section presents an analysis of the previous related work, different approaches, and techniques for the detection of different mental disorders using social media. This section is focused on works related to the areas of depression detection, anorexia detection and Post Traumatic Stress Disorder (PTSD). There are related works of the features and predictors they implemented.

2.1 Depression detection in social media

Depression is one type of mental disorder that have an increasing number of studies that focus on automatically detect high scores of depression in a user. To accomplish this detection, automated analysis of social media is made using predictive models that use features or variables that are extracted from the data post from the users in their social media accounts.

For example, one of the most commonly used features are the frequencies of each word that are encoded to create a users' language [33, 34, 35, 36, 37, 38, 39, 40]. In this approach each word or pairs of words frequencies are used as features, the main idea is to considered sequences of words to build a rule-based approach, but it was found that is harder to distinguish between people with depression vs people without depression, suggesting an overlap in the language associated.

Other works focus on the usage of a Linguistic Inquiry and Word Count [32], a program to extract basic counts/ratios. It contains different dictionaries for languages such as English, Spanish, German, Italian and Dutch. With this program it could extract the different word in psychologically meaningful categories like social relationships, thinking styles or individual differences [33, 34, 35, 36, 37, 38, 41, 42, 43], their used this dictionaries to characterize differences between mental disorders conditions and perform some success in the detection. Authors also proposed other dictionaries or lexicons related to depression, for example, in [19] the authors proposed a method to exploit a micro-blog platform for detecting psychological pressures from teenagers. They construct a stress-related lexicon and provide two methods to aggregate tweets in time series to get an overview of teenager's stress fluctuation and variation over time.

Other type of traditional feature is the extraction of a sentiment analysis in the post [34, 36, 37, 39, 41, 42, 43], a features that determines if a post has a positive, negative or neutral emotional charge, with this features their model the general sentiment that a user express in their post, getting some interesting results when a user tend to express a lot of negativity but did not perform well when users without depression tend to express also in a negative way. For example in [17], the authors worked in a model to predict depression of different users from a Chine social media. Their proposed a method that combines: 1) a sentiment analysis to calculate the polarity of the tweets considering the structure of the sentences, and 2) 10 features derived from psychological research like the usage of first-person pronouns, user interaction with others, user behaviors in the microblog, etc. Then they combine the features and used 3 different classifiers(nb, treeJ48 and rules decision table).

Another common feature extractor is the analysis of topics used in the post [33, 34, 37], where the idea is to understand the themes or subjects that users with depression tend to share in their social media platforms. The extraction of meta-data like the average of the length of the vocabulary, the number of words by post, the total number of words, are another kind of common features that are extracted for the analysis of the users [34, 36, 39, 41, 42, 43]. Other kinds of features such as the user activity in the social media are common of extract [34, 36, 37, 41, 42, 43]: such as the post by an hour in a day, the hour they post, mention of other users, friend, followers. This kind of features helps to enrich the information of the users and helps to improve the detection of depression. In [18] the authors proposed a suicidal detection over Sina Weibo, a Chinese social media. They used linguistic features from HowNet a lexicon used for sentiment analysis. They analyze the polarity of the words or phrases posted by the users to use it as features. They find the usage of temporal features could be useful. They made an analysis of preferences of people with high-risk suicide like time of the posts, originality in the posts and self mention.

2.2 Anorexia detection in social media

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. Most of the previous studies focus on identified anorexia using user-generated content from their social media platforms to generate features. Some of the most common are: the analysis of syntactic and semantic content in the posts [44, 45, 46, 47, 48], this approaches divided a sentence analyzing the structure and meaning a linguistic level.

Other traditional feature, is the usage of sentiment analysis to analyze the emotional characteristics for every person [47, 49], similar to depression, this approach search for a relation in the sentiments that are posted by users that presents signs of anorexia.

Another common feature is the extracting using words or dictionaries that are related to the topic of anorexia [47]. Recently some works had explored the usage of Deep Learning techniques, and getting competitive results [44, 48, 50]. The combination of these different approaches performs better than used them separately, each kind of feature enriches the representation giving important information for the detection of anorexia. For example, the combination of models that employ user-level linguistic metadata, frequencies of words, neural word embeddings and a convolutional neural network, gets the best result for the detection of anorexia in [2].

2.3 Post-traumatic stress disorder detection in social media

Post-traumatic stress disorder (PTSD) is a mental disorder that is caused when a person experiences a terrifying event, either experiencing it or witnessing something. People who suffer traumatic events tend to have difficulties in adjusting in society, but with time they can get better. PTSD is not as popular to study as depression or eating disorders. Some works focus more on the semantic and syntax analysis [14, 15, 16]. For example, in [14] the author examined a range of supervised topic models to find groups of words with differentiate between each class, and then calculate topics over the posts. In [15], they examined inferring topics automatically, combined with unigram words. Other works focus on the usage of LIWC to extract basic counts and ratios [13].

The results suggest an open room for future improvement and work, the task is not solved yet. The techniques that were employed provide insights from the PTSD problem and the opportunity for a new direction for mental health research.

2.4 Evaluation Forums for Mental Disorders

CLEF eRisk: Early risk prediction on the internet[1]. eRisk is 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. Their main goal is to pioneer in a new interdisciplinary research area that focuses on early alerts that could be sent when, for example,

[1] https://early.irlab.org/

people with suicidal inclination or people susceptible to depression or other mental disorders start to interact in social networks, forums or blogs. Early detection technologies have the potential to be applicable to a wide variety of areas, especially those related to health and safety [1].

CLPSYCH: Computational Linguistics and Clinical Psychology Workshop[2]. CLPsych is a workshop that introduces a union between clinical psychology and natural language processing for mental health. Their goal is to bring together scientists and clinicians interested in improving mental health through language understanding. CLPsych focus on an interdisciplinary audience where they share their findings and methods to improve assessment of mental health care.

3 Research Proposal

This section presents in detail the research proposal. In the first part we present the problem statement, then a description of Multichannel Learning, the next part present the research questions, then the objectives, the hypothesis and in the final part the expected contributions of the research.

3.1 Problem Statement

Previous studies that focus on the detection of mental disorders like depression, anorexia or PTSD suggest that these symptoms are detectable on online environments. Most of the works focus on the usage of dictionaries related to the topics, sentiment analysis looking for the polarity of the post or counting the frequency of the words and then combine the information using generally a simple concatenation. The performance is still modest, suggesting the challenging of the problem. This presents an opportunity for exploration and analysis of new techniques to extract types of information from the user's posts and create a model that learns to automatically combine these channels of information that could better represent the posts and improve the detection of signs and symptoms of different mental disorders. On the other hand, the nature of the social media platforms is dynamic, where the information is constantly increasing in sequence order, a study and analysis of the sequentiality presented could also help to improve the results of detection.

3.2 MultiChannel Learning

In the real world, the information usually is presented in different modalities that help to learn a new combined representation [63]. For example, images that are associated with text that describes it or videos that contains audio, images and text (subtitles). Sometimes available datasets only contain one of these modalities and a multichannel learning inspired in the multimodal learning is possible.

For this work, a *channel* is defined as a different property or view from the same modality, for example, in [56] they divided the 3D skeleton sequences in different channels and then learn to combine the information of the channels. For this work we use the text modality and some examples of channel could be the semantic aspects that are contained, the phonetics that are used,

[2] http://clpsych.org/

the emotions presented or the style of the author for writing or expressing. Multichannel Learning creates a representation that combines two or more of these channels, discovering the relationship between different channels. This learning is a good representation of the joint of different channels.

Figure 1 shows the process of extracting the different types of information (channels) from the documents, and then, a model that learns how to automatically combine the channels in a single representation.

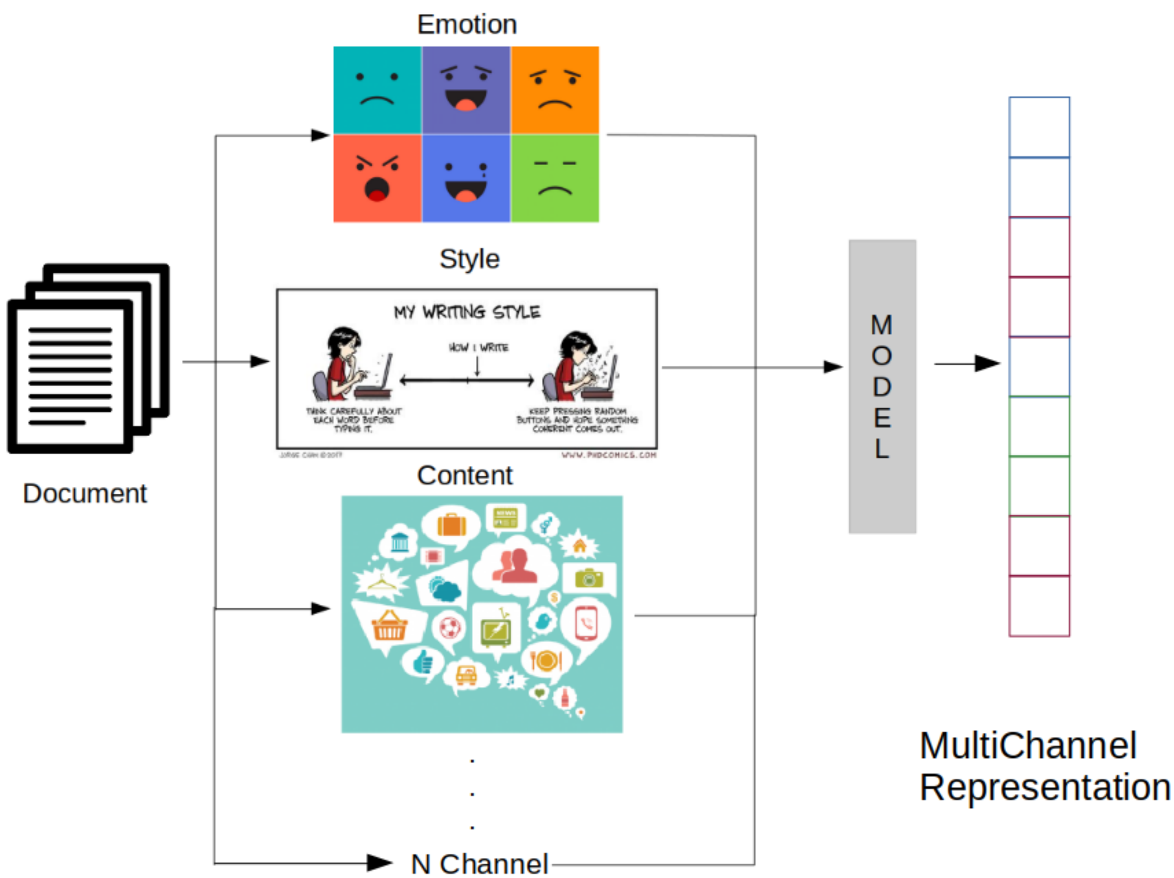

Figure 1. Multichannel Representation, extracting different types of information from the same modality and then a model that automatically learns how to combine it.

3.3 Hypothesis

People that present some mental disorder tend to express differently than healthy people; their topics of interests, writing style, relation with others and even their activity hours had different behavior. The hypothesis is that learning to combine different channels of information, could give a broader view that helps to detect signs of mental disorders and obtain better classification results that using single information.

3.4 Research Questions

Due the increasing popularity of social media platforms, the opportunity of detecting mental disorders have increased through the analysis of linguistic styles, thematic content, emotions and other activity traces of different users (e.g., Facebook, Twitter, Reddit, etc.). The information shared by users in social media (a.k.a. posts) seen in different new channels, could be useful for the detection of mental disorders and creates the following questions:

1. *Which information presented in the posts of the users could be helpful to detect that a user has a mental disorder problem?*
   In a post, it can be analyzed different kinds of views from the same source (e.g., semantic aspects, style, emotions) and combine them to get more information of the users and create a profile that determinate if they have a problem.

2. *How to automatically combine the data presented in the posts to improve the representation for the detection of mental disorders?*
   Different views of the information could give us a full look of the data, but this does not mean that all information provides the same value for the detection. Looking for a way to automatically combine each information channel could improve the detection in users that have these mental disorders.

3. *How relevant is the temporality or sequentiality of information presented in the user's posts?*
   Users that present mental disorders have more unstable behaviors, and using the temporal information to capture these changes through the time in the post could help to improve the representation of the information.

3.5 Main Objective

Design a method applying traditional NLP techniques combined with deep learning techniques to automatically learn a Multichannel 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.

3.6 Specific Objectives

1. Design methods that learn new representations of the different channels in the post history of the users: the context, the style of the author, the emotions used and phonetic information, that improves the representation of the users for the detection of mental disorders.

2. Design 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.

3.7 Expected Contributions

Through this doctoral research are expected to obtain the following contributions, where the first and second contribution are the most important:

1. Different methods that use different views of the information to create separate channels that are available in the post history of the users and evaluate each of these channels to identify if a user has a mental disorder.
2. A representation of a user's profile that was learned automatically by a model combining the information of different channels, that improves the detection of a mental disorder.
3. A method that takes advantage of the existing sequentiality in the texts to enrich the representation.
4. A detailed study of the utility of using different channels to detect mental disorders in users of social media.

4 Methodology

This section presents in detail the methodology to reach the proposed objectives. The proposed methodology consists of four stages, where stage 2, 3 and 4 have the major contributions of the dissertation proposal. Section 9 contains some concepts related to the techniques in this section.

1. **Identify and Obtaining datasets related to mental disorders.** We plan to obtain datasets like depression detection, anorexia detection, PTSD detection. The purpose is to find different datasets collection that is related to the detection of conditions that affect the thinking, feeling, mood, and behavior of people using their post history. These conditions can affect the ability to relate to others and function each day. The following part presents some of the criteria to guided the selection of these datasets collections.
   (a) *Social Media Platforms:* Identified datasets where the post of the users are using a Social Media Platform like Facebook, Twitter, Reddit, etc.
   (b) *Mental Disorder related:* Obtain datasets related to the detection of mental disorders, either between users without a mental disorder or users with different types of mental disorders.
2. **Develop methods that extract information in different channels.** In this step, it is necessary the analysis of different kinds of information presented in the posts to extract and create separate channels. There exist different information depending on the complexity of the desired process, for example, if it is wanted to study the internal structure of words and a core part of linguistic, the description of how words are grouped and connected to each other in a sentence, the understanding of the meaning of words or complex tasks. Some possible channels could be:

(a) *Emotion Channel:* Design a method that creates a representation of different emotions presented in the text that help to detect people with a mental disorder problem. Most of the works focus in the extraction of positive and negative sentiments. The analysis of emotion-related expression could be important to reveal symptoms or insights of people that have some psychological distress state. Emotions have been widely studied in different research areas like psychology and neuroscience because they are an important part of human nature [4]. Some psychological studies found a correlation between mental disorders and emotions and have been explored using social media platforms [5].

(b) *Semantic Channel:* Design a method that is based on the semantic analysis to create a representation that could capture the connections between understanding and relation of words. Semantic Analysis provides the meaning of words and also their relationship with other words. To create a good semantic analysis of the data, it is important to know the context of the surrounding words, phrases, and objects, to extract the relevant parts and compare them to deepen the understanding of the content [51].

Some popular techniques to obtain this analysis are: the usage of Latent Semantic Analysis (LSA) [52] to extract relationships between a set of documents and the terms that are contained to produce a set of concepts related to terms and documents. Another popular technique is the usage of Ontologies to extract structure information from the unstructured data.

(c) *Style Channel:* We plan to design a method that creates a style representation of the user, for example, the usage of passive voice, questions, and personal expressions help to identify the usage of formal language, understanding the readability and connection between the expressions used in their posts. The usage of style analysis could give hints for identification and verification of users in social media and help to categorize their posts finding similarities between the people that could have a mental disorder [53].

(d) *Phonetic Channel:* Similar to the style channel, create a representation using the properties of the sound of the words and create relations between ”slang” and common words. In social media due to the way of people writing more informally, they tend to change the words to adjust it in how they speak, this creates a lot of vocabulary that normally is harder to process.

Phonetic analysis is related to how the sounds of the words are produced when someone speaks. It has different ways of study the sounds, for example, using the acoustic phonetics that deals with the waves of sound that a human-produced, the auditory phonetics that concentrate on how the brain and ear process the sounds, the articulatory phonetics that study the movement of various parts of the vocal tract when someone speaks [54].

3. **Develop a model to create a representation that combines the different channels automatically.** This step involves the development of a model that automatically combines the different channels obtained in the step before, and creates a new representation. For example, traditional algorithms based on concatenation of the features or ensemble of classifiers

tend to learn some of the hierarchical structure of the information but did not capture well the relationship between the different kind of features. To overcome this problem using models inspired in Deep Neural Networks that learn to combine and or give importance to a different type of information. A comparison between traditional and Deep Networks for combining information is needed to determine the relation of the various channels, some examples of these are:

(a) *Early Fusion:* Develop a method to early fusion the information of different channels. The information from the different channels is taken as one vector, and then using a classifier to learn this representation [55].

(b) *Late Fusion:* In this part each group of features are represented as a vector, and are used to train an ensemble of classifiers, where the obtained results are weighted and mixed [55].

(c) *Autoencoder:* Design a model inspired in an encoder to combine the different channels and compress into a short representation, then use the decoder to transform this representation in the desired output.

(d) *Attention Models:* Develop a model with the attention mechanism that could learn the important features of the different channels, extracting the most important parts from the channel combined or learn from each one the relevant information. A big advantage of attention is that it gives us the ability to interpret and visualize what the model is doing for an easier analysis of the results.

(e) *Transformers:* Similar to the previous part (attention models), using a model inspired in a Transformer to extract the most important parts of the features from each channel to generate a new representation of the data.

4. **Design an approach that effectively incorporates sequential information in the representation.** Due to the nature of the information that is created involving the sequencing of actions, where a user writes a post one after another. In this step of the work is proposed analysis and exploration of the usage of the sequentially presented in the user's posts. For example, hand-crafted temporal features and deep learning models like Recurrent Neural Networks that take time and sequence into account. This type of artificial neural network is designed to recognize patterns in sequences of data, and are often used in text analysis, spoken word, numerical time series or handwriting. There are different types of Recurrent Neural networks that can be used to analyze the sequences of the posts, some examples:

(a) Hand crafted Features: Design a method that extract the temporal information like statistical features like mean, standard deviation or variance, that could help to analyze the information as a signal made of features.

(b) *Recurrent Neural Network:* Design a method inspired in Recurrent networks. RNN take as their input, not just the current input example they see, but also what they have received previously. With this process, the network creates a memory of what they previously learn and it finds correlations between events separated by many moments.

(c) *Long Short-Term Memory Units (LSTMs):* Similar as the previous part, design a method that use a LSTM. This neural network is a variation of recurrent networks, that contains information outside the normal flow in a gated cell. This information can be stored in, written to, or read from a cell. The cell decides what store and what forget, this allows bigger retention of information than normal recurrent networks.

(d) *Gated Recurrent Unit (GRU):* Design a method that use a GRU to incorporate temporal information. This network is a variation of an LSTM without an output gate, this cell fully writes the contents from its memory at each time step to the larger net.

5 Work Schedule

This section presents in Figure 2 a general work schedule for the next three years, and it includes the most relevant activities that are planned.

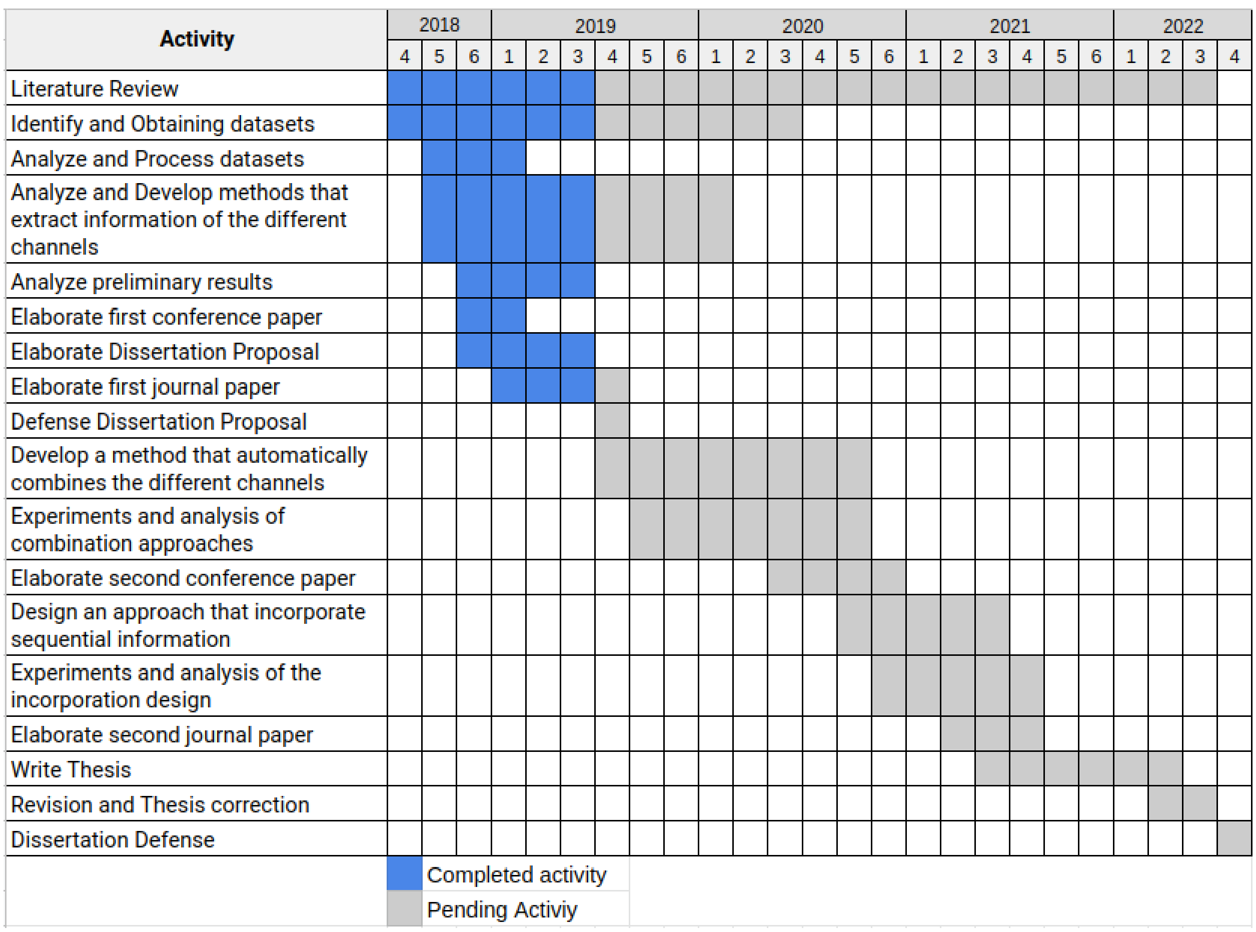

Figure 2. Work Schedule for the completed and pending activities divided by bimester.

6 Preliminary Work

This section presents the preliminary work that has been done that supports our hypothesis and research proposal. The following points are a resume of the preliminary work:

1. Identify and Obtaining first datasets (part of the first step in the methodology). For this step, we obtained datasets from eRisk evaluation task.

2. Our first experimental approach consists of the usage of the emotions channel (part of the second step in the methodology); we proposed a new representation called Bag of Sub-Emotions (BoSE). This channel represents social media documents using a set of fine-grained emotions that are automatically generated using lexical resources based on emotions and sub-word embeddings. To evaluate this representation, we used two different tasks: depression and anorexia detection. The results are promising; the usage of these fine-grained emotions improved the results from a representation based on traditional methods and based on the core emotions. The results obtained are also competitive in comparison to state of the art approaches.

3. Temporal analysis for the emotion channel (part of the fourth step in the methodology). A first exploration of the temporal information that is presented in the emotion channel. For this experiment we explore the usage of handcrafted temporal features.

4. An early and late fusion of the temporal features with the original BoSE (part of the third step in the methodology). A first exploration in combining different information from the same channel. This approach obtains a little increase in the results that using the information separated.

6.1 Identify and Obtaining first datasets for Depression and Anorexia detection

Our first step was to evaluate our approach to the tasks of depression and anorexia detection. For this, we obtained the datasets from eRisk 2017 and 2018 evaluation tasks [1, 2].

These datasets contain Reddit posts for several users. The users who explicitly mentioned that were diagnosed with depression and anorexia were automatically labeled as positive, the rest of them were labeled as negative. Table 1 shows some numbers from these datasets.

6.2 A new representation for the Emotion Channel

Figure 3 describes the first approach using the emotion channel. It has two general steps: in the first step, it used a lexical resource described in [66] and compute a set of fine-grained emotions for each broad emotion presented. In the second step, it uses the generated fine-grained emotions to mask the texts and then represent them using a histogram of their frequencies. This new representation is named **BoSE** (Bag of Sub-Emotions). In the next subsections, it further explains these two main steps.

Data set	Training		Test	
	NC	C	NC	C
dep eRisk'17	83	403	52	349
dep eRisk'18	135	752	79	741
anor eRisk'18	20	132	41	279

Table 1. Mental dissorders datasets used for experimentation. Each dataset have two classes (No Control (have mental disorder) = NC, Control (do not have mental disorder) = C).

6.2.1 Generating Fine-Grained Emotions

To generate these fine-grained emotions, first we use a lexical resource based on eight emotions [57] and two sentiments[3]; Anger, Anticipation, Disgust, Fear, Happiness, Sadness, Surprise, Trust, Positive and Negative.

These emotions are represent in a formal way as $E = \{E_1, E_2, ..., E_{10}\}$, where E is the set of emotions presented in the lexical resources and $E_i = \{w_1, .., w_n\}$ is the set of words that are associated to the emotion E_i.

Then, we computed a word vector for each word using a pre-trained sub-word embedding of size 300 from FastText [58]. These vectors were pre-trained using Wikipedia. After the vectors are computed, we create subgroups of words by emotion using the *Affinity Propagation* clustering algorithm [59]. This algorithm chooses the number of clusters based on the data provided, does not employ artificial elements to create the clusters. Table 2 presents the length of the vocabulary for each emotion in the lexical resource and the number of clusters created using *Affinity Propagation*.

After this process, we have a set of fine-grained emotions that represent each broad emotion as $E_i = \{F_{i1}, ..., F_{ij}\}$, where each F_{ij} is a subset of the words that were computed from E_i and is represented by the average vector of their respective embeddings.

Creating these subgroups of words allows to separate each broad emotion by topics, these topics help to identify and capture more specific emotions used or expressed by the user in their posts.

Figure 4 presents some examples of groups of fine-grained emotions that were automatically obtained using this approach. If the figure is analyzed, for each column we can appreciate that words with similar context tend to group. We can also notice that even in the same emotion each group of words shows very different topics. For example, in the Anger emotion, it has one group that is related to fighting and battles and another group with loud noises or growls. In the Surprise emotion it has some interesting examples; one group express surprise related to art and museums, in another group the emotion is related to accidents and disasters, and in other group presents surprise related to magic and illusion.

[3] In the rest of the document it refers to these sentiments as emotions as well.

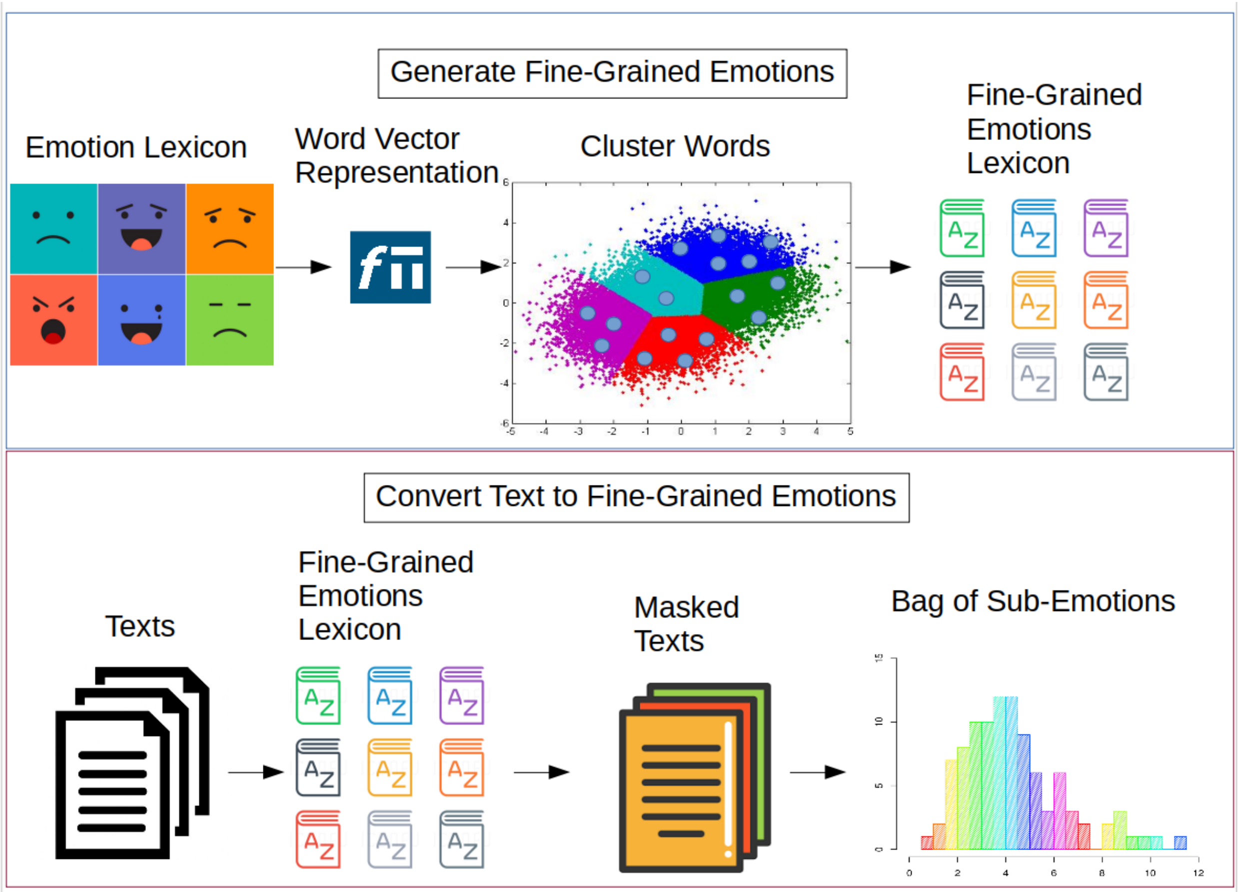

Figure 3. Diagram that represents the creation of the Bag of Sub-Emotions (BoSE) representation. First, Fine-Grained Emotions are generated from a given Emotion Lexicon; then, texts are masked using these fine-grained emotions and their histogram is build as final representantion.

6.2.2 Building the BoSE Representation

To build the representation we need to perform the next two steps:
Text masking: First, mask the documents replacing each word with a label that represents the closest fine-grained emotion. For this, we computed the vector representation of each word in the document using the sub-word embeddings obtained from FastText. Then, measure the distance between each word vector and all fine-grained emotions using the cosine similarity. Then, change each word by the label of the closest fine-grained emotion. Consider this example to illustrate the process: for the text "*Leave no stone unturned*", the sentence will be masked as "*negative8 anticipation29 anger10 anticipation3*".

Text representation: Once the documents are masked; we built their **BoSE** representations computing a frequency histogram of their fine-grained emotions. We have two different approaches to build these representations: *i*) a histogram is created counting the number of occurrences of each

Emotion	Vocabulary	Clusters
anger	6035	444
anticipation	5837	395
disgust	5285	367
fear	7178	488
joy	4357	318
sadness	5837	395
surprise	3711	274
trust	5481	383
positive	11021	740
negative	12508	818

Table 2. Lenght of the vocabulary for each emotion and the number of clusters created for each one.

fine-grained emotion presented in the text, this process is similar to the Bag-of-Words representation. We named this representation as **BoSE-unigrams**. *ii*) similar to the previous approach, a histogram is created counting the occurrences of sequences of fine-grained emotions in the document, and refer to this representation as **BoSE-ngrams**. For the latter representation, we tested different sizes and combinations of sequences.

6.2.3 Experimental Settings

Preprocessing: For these experiments, first the texts are normalized lowercasing all the words and removing special characters. Once the text is preprocessing the text is masked using the created fine-grained emotions.

Classification: After built the **BoSE** representation, we select the most relevant features of the sequences of fine-grained emotions, using the chi^2 distribution X_k^2 [60]. Once the best features are selected, to classify the documents we use a Support Vector Machine (SVM) with a linear kernel and $C = 1$.

Baselines: To evaluate the relevance of the created fine-grained emotions in the detection of mental disorders, the representation is compared with a representation based on the occurrences of the broad emotions combined with the words that do not have an associated emotion. This approach is named Bag-of-Emotions (BoE). Also, the results are compared to traditional Bag-of-Words representation. Both representations were created using word unigrams and n-grams; these are common baseline approaches for text classification. Additionally, the representation results are compared against the participants of the eRisk 2017 and 2018 evaluation tasks [1, 2], considering the f_1 over the positive class.

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. Examples of Fine-Grained Emotions corresponding to four different broad emotions.

6.2.4 Experimental Results

For this work, different evaluation metrics are needed and are described in the following part:

1. **Precision:** Precision in pattern recognition and information retrieval is also called positive predictive value. When a program retrieves instances that are predicted as the ones of interest, the precision calculates the correct instances among the predicted instances. The formula could be express as: $Precision = \frac{TP}{TP+FP}$ where TP are the right predictions from the program and FP are the wrong predictions selected as right.

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 that exist. The formula could be express as: $Recall = \frac{TP}{TP+FN}$ where TP are the right predictions from the program and FN are the right predictions that were not selected as right.

3. **F-measure:** Also, know as F1 score or F-score. Is an evaluation measure of the test accuracy, where it considers the precision an the recall to give the score of the evaluation. The F-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 could be express as: $F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$.

In our first experiment, we evaluate the effectiveness of the **BoSE** representation to identify mental disorders in users. To analyze this, we compared its performance against the results obtained using BoE representation and a traditional BoW representation. Table 3 presents the f_1 performance over the positive class for the BoW, BoE and BoSE approaches. It can appreciate that the BoSE representation outperforms both baseline results, especially when are considered sequences of fine-grained emotions for the representation. To further expand our exploration, it also used BoSE representation without positive and negative sentiments (BoSE8). In the results, it can be appreciated that the performance drops without the usage of the two sentiments; this demonstrates that these sentiments give important information to identify mental disorders.

Table 3. F1 results over the positive class against baseline methods

Method	Dep'17	Dep'18	Anor'18
BoW-unigrams	0.60	0.58	0.69
BoE-unigrams	0.57	0.60	0.50
BoSE8-unigrams	0.56	0.6	-
BoSE-unigrams	0.61	0.61	**0.82**
BOW-ngrams	0.59	0.60	0.69
BoE-ngrams	0.61	0.58	0.58
BoSE8-ngrams	0.57	0.59	-
BoSE-ngrams	**0.64**	**0.63**	0.81

To further evaluate the relevance of the **BoSE** representation, Table 4 compares its results against those from the first three places at the eRisk 2017 and 2018 evaluation tasks, respectively:

Table 4. F1 results over the positive class against top performers at eRisk

Method	Dep'17	Dep'18	Anor'18
first place	**0.64**	**0.64**	**0.85**
second place	0.59	0.60	0.79
third place	0.53	0.58	0.76
BoSE	**0.64**	0.63	0.82

For these tasks, the participants create more complex models than our proposed approach. They employed different types of data, inspired by traditional representation and deep learning models. They employ for example linguistic meta-data from user-level, word embeddings, the combination of different models including convolutional neural networks, sentence-level analysis, different linguistic features, domain-specific vocabularies, and psychology-based features.

From the obtained results it can highlight the following observations:

1. The approach outperformed the traditional BOW representation in both datasets, indicating that considering emotional information is quite relevant for the detection of depression and anorexia in online communications.

2. The use of fine-grained emotions as features helps improve the results from a representation that only considers broad emotions. This result confirms our hypothesis that users with a mental disorder tend to express their emotions in a different way than users without them.

3. The approach obtained comparable results to the best-reported approaches in both datasets. It is essential to highlight that the participants of these tasks tested different complex models with a wide range of features and sophisticated approaches based on traditional and deep

learning representation of texts, whereas ours only relies on the usage of the fine-grained emotions as features.

For further analysis, in Figure 5 we can appreciate a 3D plot using the t-sne algorithm [65]. In the first column there are the graphics for the Bag of Words (BoW) representation of the users, and in the second column are the graphics for the BoSE representation. We can see the depression detection task in the first row, where for BoSE the red dots that represent the depressive user are more clearly than in BoW representation. In the second row, we can be appreciate the anorexia task, where BoSE has a more clear separation between the users than using BoW.

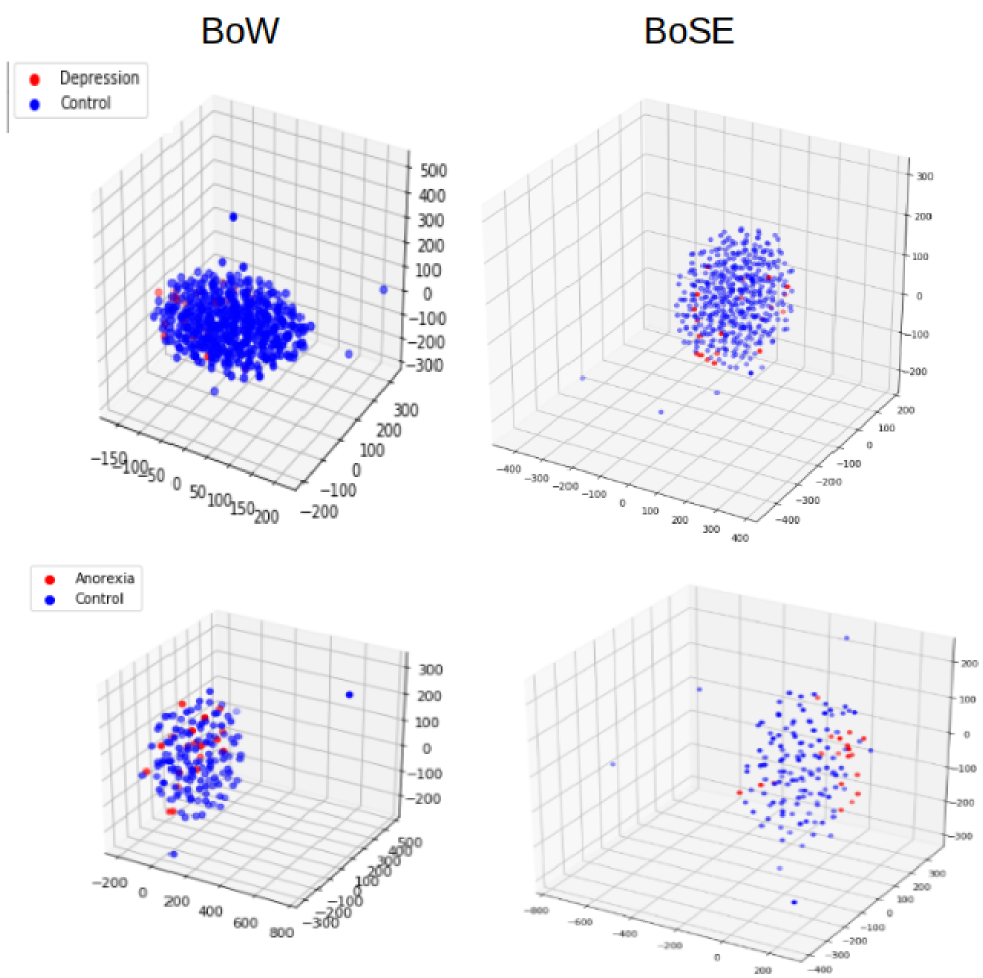

Figure 5. Plot of the BoW and BoSE representation for the detection of Depression an Anorexia.

6.2.5 Analysis of the Fine-Grained Emotions

To further understand what fine-grained emotions capture, the most relevant sequences are selected for the detection of depression and anorexia according to the chi^2 distribution. Table 5 shows some relevant sequences of fine-grained emotions as well as some examples of the words that correspond to these sequences in the depression task. Table 6 shows some of the relevant sequences of fine-grained emotions for the detection of anorexia task and also some examples of the words corresponding to these sequences.

Most of the fine-grained emotions that present high relevance for the detection of depression is related to negative topics, for example, the anger emotion is associated to the feeling of abandonment or unsociable, and the disgust emotion is related to dilution, insecurity, and desolation. These fine-grained emotions capture the way a depressed user expresses about themselves and their environment.

Table 5. Examples of words that create the fine-grained emotions

Examples of relevant sequences	
”anger1” ”anger11-anticipation10” ”disgust16-anger11” ”disgust11-fear17”	
anger1	abandoned, deserted, unattended
anger11	unsociable, crowd, mischievous
anticip10	disappointed, inequality, infidelity
disgust16	unsatisfactory, dilution, influence
disgust11	insecurity, desolation, incursion
fear17	hysterical, immaturity, injury

Table 6. Examples of sequences relevant to the anorexia detection

”anger4”- ”negative65” ”disgust32” - ”anticipation12” ”anticip10” - ”fear19” ”disgust21” - ”anticip12”	
anger4	bruising, contusion, bleeding, fracture
disgust32	breakdown, fight, crushed, abandoned
disgust21	stomach, intestinal, bile, esophagus
negative65	bathroom, toilet, washroom
anticip10	hurting, refused, anxious, afraid
anticip12	ashamed, embarrass, upset, disgust
fear19	food, eating, eat, consume

For the anorexia detection the fine-grained emotions that present higher relevance is related to embarrassment, self-harm and eating topics, for example, the disgusts emotions are associated to mental states of defeated and internal organs related to eating, and anticipation emotions that are related to self-harm, fear and shame. These fine-grained emotions capture the essence of the problems that are presented in a person that have anorexia and how they are expressed.

To analyze the fine-grained emotions used for each task, Figure 6 presents the distribution of the 1000 most important fine-grained emotions obtained using chi^2 and are group by their general emotion. It can be appreciated that the emotions have different distribution depending on the task, this demonstrates that the representation captures the emotions that different persons with mental disorders tend to express when they post in their social media platform.

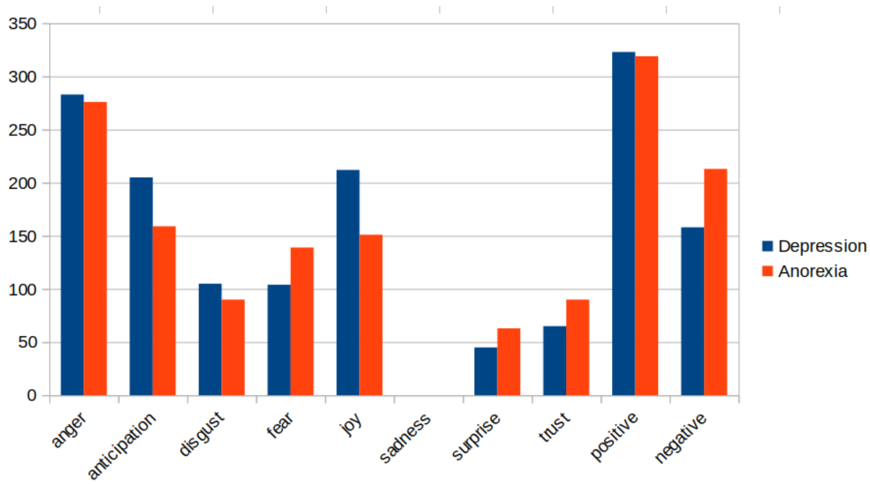

Figure 6. Distribution of the 1000 most relevant fine-grained emotions for each task.

In Figure 7 we present different word clouds created from the datasets. We can see in the word clouds that different emotions are predominant for each task, similar as the previous analysis of emotions, the representation captures different important topics related to emotions depending on the task.

6.2.6 BoSE in early Predictions

For this experiment, the main idea is to see how much information is needed to have an adequate detection performance. We divided the post history of the users in 10 parts. Figure 8 show the results of the proposed representation against the BoW and n-grams competitors. We can appreciate that the BoSE representation, in general, gets better performance than the traditional representations, in the depression task having approximately the 70-80% of the history post is enough to get better results consistently. In the anorexia task BoSE outperforms the other traditional representations with a more extensive advantage, and only need approximately 40% of the post history to get a better result than the best result of the others representations with the 100% of the data.

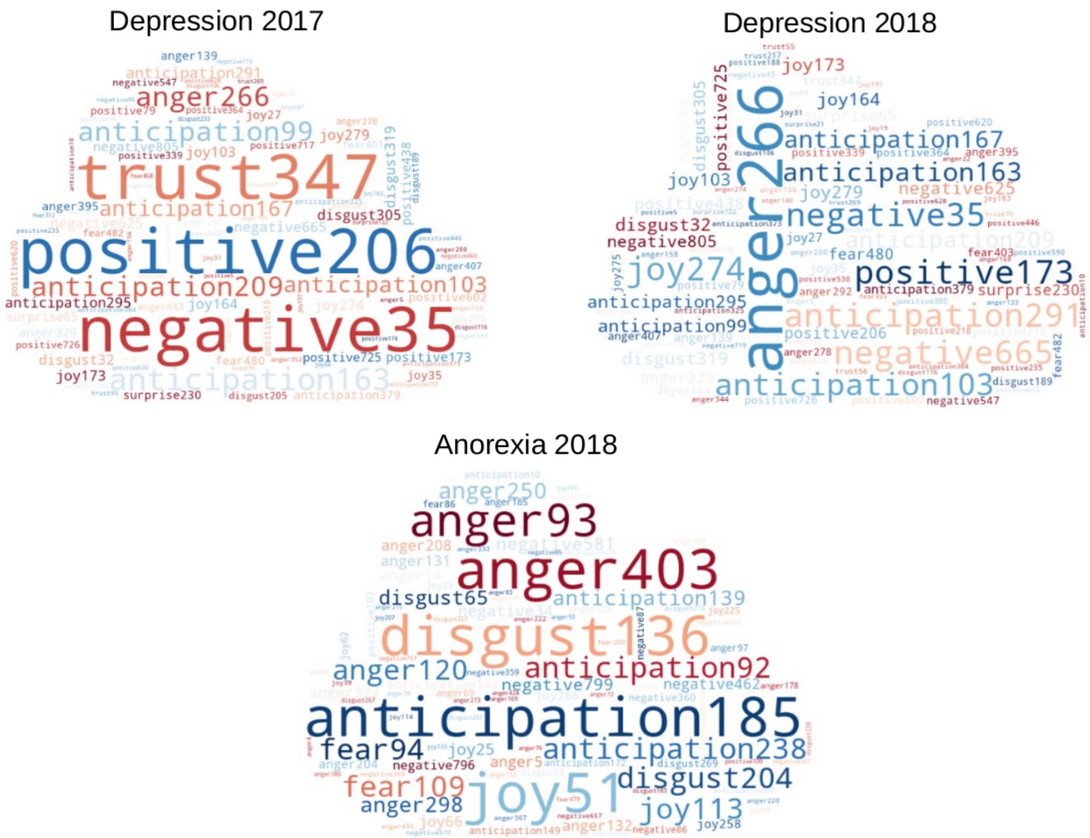

Figure 7. Word Cloud distribution of relevant fine-grained emotions for each task.

6.3 Temporal Analysis for Fine-Grained Emotions

To further analysis the use of this representation based on fine-grained emotions, a new approach is proposed, using the sequentially presented in the data. The hypothesis behind this approach is that a user that has a mental disorder tend to present more instability in their emotions than a user without a mental disorder. For this, the post history of the users is divided into ten parts, for each part is calculated the BoSE representation creating a vector of the fine-grained emotions, and finally two different strategies are used: 1) Calculate the difference between the vectors each time, this creates nine new vectors that consists in the difference of each fine-grained emotion in each different time. 2) Use the ten vectors directly without the need of calculating the differences.

Once the vector of each time in the post history is created, an statistical analysis is performed. The information created by the fine-grained emotions through the time is treated as a signal and eight different features are extracted from each fine-grained emotion: mean, sum, max value, min value, standard deviation, variance, average, and median. This creates a temporal feature vector, that captures the changes through the time of each fine-grained emotion and is uses to classify the users.

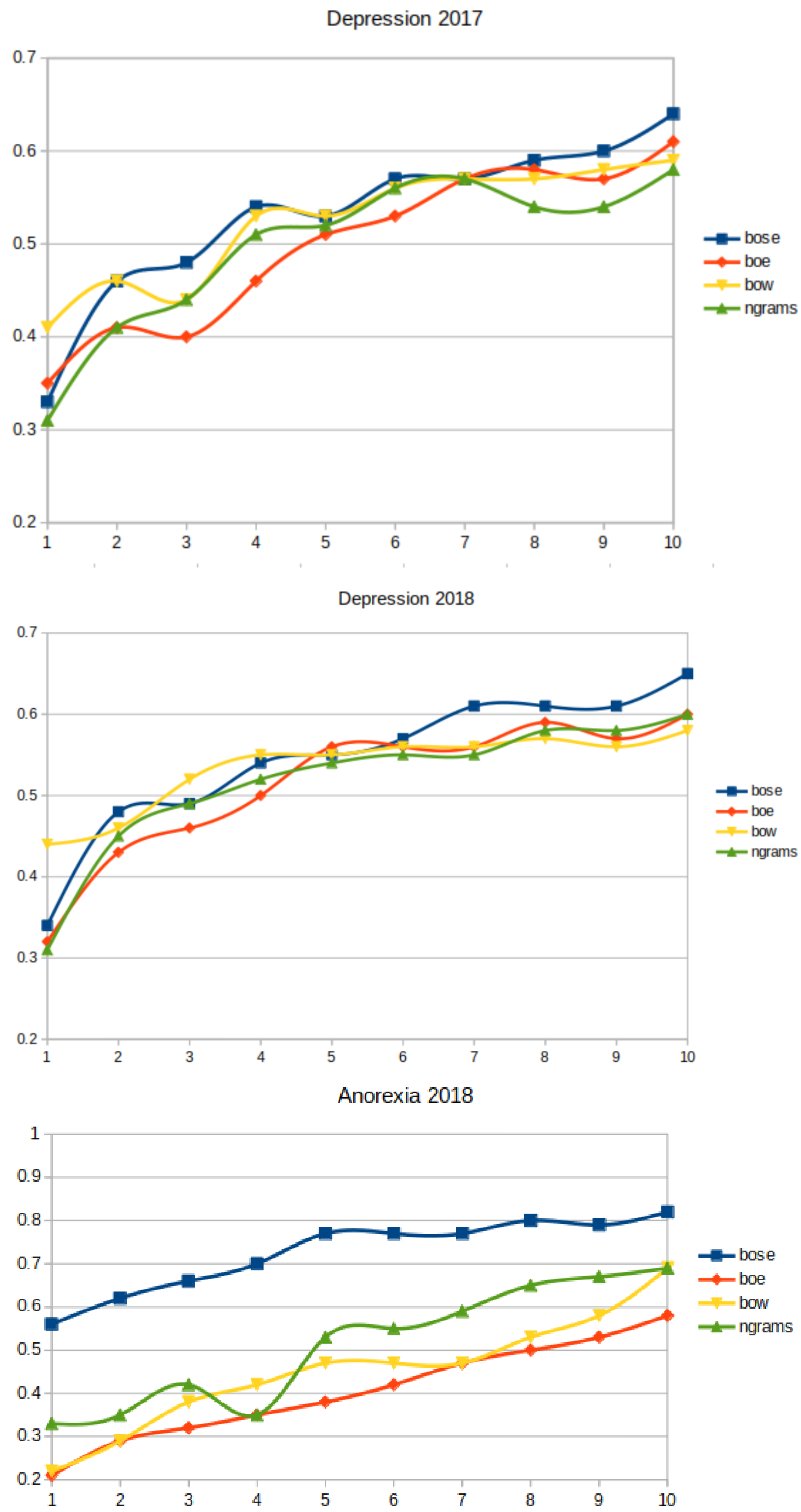

Figure 8. Results by chunk in the datasets. X-axis represent the chunks and Y-axis the F1 result.

Table 7. Results for Temporal, NonTemporal and Fusion

	Depression'18	Anorexia'18
NonTemporal	0.63	0.82
Temporal Diff	0.59	0.77
Early Fusion	**0.64**	0.81
Temporal Abs	0.53	0.79
Early Fusion	0.62	0.77
Late Fusion	**0.64**	**0.84**

After the creation of the temporal vector, the next step is to combine the temporal and nontemporal vector, to improve the results. Two different strategies were proposed for this combination: 1) concatenation of the vectors, and 2) a vote of the classifiers. These approaches are also known as early fusion and late fusion respectively. We present the results in Table 7, we can appreciate that the late fusion performs well for both task, and obtain an improvement in the results.

For more in-depth analysis of the temporality of the emotions, Figure 9 presents some examples of these fine-grained emotion signals. In this figure, we compared the control group colored in orange against the mental disorder group colored in blue (depression is in the left part and anorexia is in the right part). As we can see, the control group present fewer changes or peaks through time than the mental disorder group, this proves our hypothesis of users that have more emotional instability when they present signs of a mental disorder.

6.4 INAOE-CIMAT at eRisk 2019

In this section, we described the joint participation of INAOE-CIMAT using Bag of Sub-Emotions (BoSE) at eRisk 2019. The 2019 Early Risk Prediction on the Internet (eRisk@CLEF) is a forum of evaluation that has the objective of dealing with problems related to health and safety risks detection on the Internet. The main goal of the task that the organizers present, is to identify if a user presents signs of anorexia as soon as possible, processing as pieces of evidence their post history. Applying sequentially monitoring of the user's interactions in their social media platforms, post are processed in the order they were created and then a prediction is send.

This shared task is a continuation of eRisk 2018 T2 task [2]. This task consists in detecting signs of anorexia as soon as possible in users of Reddit. This detection is done by sequentially processing the users' posts. This year task, the organizers modified the way the posts are released, which was variable-chunk-length in 2017 and 2018, where users that wrote more would have more information per chunk. Now the posts are released item-by-item and a server iteratively provides user post in chronological order and using a token identifier for each team. For each iteration that the server provides a post, we need to respond with a prediction to continue the next round of posts; otherwise, the server won't provide the next iteration.

The strategy for the shared task is to decide if a user presents signs of anorexia making a prediction every five posts, where the posts are preprocessed and a classification procedure is made to generate the labels for each user. Lastly, we used five different strategies to sent the predictions.

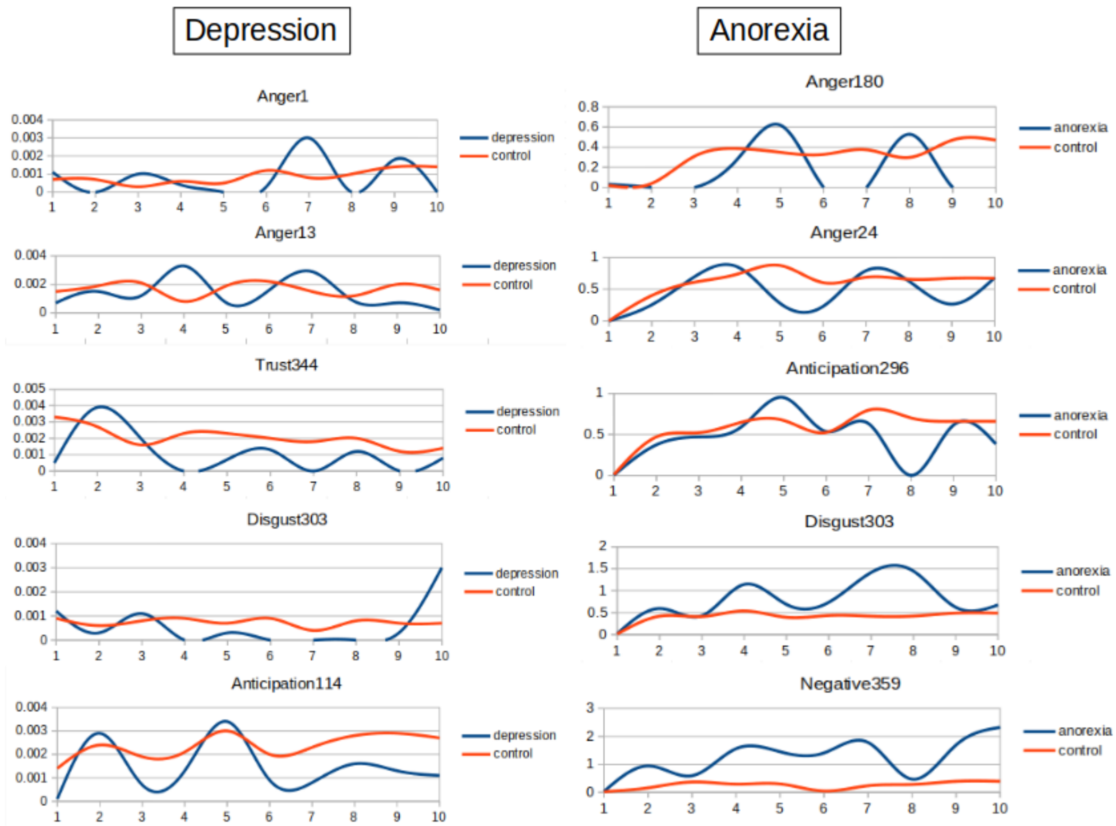

Figure 9. Emotional signals comparison between the control group and the mental disorder group. X-axis represent the number of parts that each document was split and Y-axis represent the value of the sub-emotion in that time.

We explained the whole process below.

Prediction making: For each post that the server provides, we need to predict if the user presents signs of anorexia or not. The main idea is to make a correct detection as soon as possible. We proposed to tackle the task by using the following five strategies: *i*) considering the label obtained directly from the classifier; *ii*) using the probability of the label, assigned as positive if the chance is higher than 60% of belonging to the anorexia class; *iii*) similar to the first strategy, consider the label obtained directly from the classifier, but only assigned the label 1 if the user is detected as positive in the previous and current predictions; *iv*) the user is classified as positive if the probability of the classifier is higher than 60% in the actual and previous predictions; *v*) similar to the fourth strategy but the classification probability needs to be higher than 70%.

6.4.1 Experimental Results

First evaluated the model with the previous dataset provided in 2018, and determine the parameters for the model and then send the prediction in the server. For this dataset, there are two categories of users: with anorexia and control (users without anorexia). We measured the F1 over the positive class using the whole post history of the users. In Table 3 we present the obtained results over the 2018 dataset.

For the test dataset, we trained the model using all the users in the training dataset and then we determined if the users present or not signs of anorexia using the five different strategies mentioned in the previous subsection. Table 8 show the results obtained by the five strategies over the test dataset. Is important to note that on these results: run1 did not work on the server, and we still do not know the reason for this, therefore, their results are not included in the table. The fourth strategy obtained the best results (named as run3); this strategy consists in classifying the user as positive if the probability is higher than 60% in the current and previous prediction. This strategy involves the temporal stability obtained by the classifier where we get two consecutive positives predictions over the user.

Method	F1	$ERDE_5$	$ERDE_50$	latency-weighted F1
run 0	0.66	0.09	0.04	0.62
run 2	0.66	0.09	0.09	0.50
run 3	**0.68**	0.09	0.05	**0.63**
run 4	0.66	0.09	0.05	0.61

Table 8. Results over the positive class in the Testing Dataset

To present further analysis of the results, in Figure 10, we present a boxplot of all the results obtained for F1 measure and Latency-weighted F1. In the figure, the green **X** mark represents the position of our results. We can appreciate that our results for both evaluation metrics are in the highest quartile, indicating the good results obtained for this task.

In the second part of Figure 10, we present the boxplots of the results of all participants under the ERDE5 and ERDE50 evaluation metrics. The results are placed in the middle quartile (note: is better a lower value in ERDE). These are expected results, since the strategy does not focus on fast prediction, but instead on the temporal stability of the users. In [3] they present the overall results of the task as well as a complete analysis of every team.

7 Conclusions

In this document we describe part of the work that has been made and the future work that is planned to do during the Ph.D. program. The main objective of this research is to focus on the detection of mental disorders in users by publishing on various social media platforms. The work will focus on the detection of these users improving the state of the art results, using a new multichannel representation that exploits traditional natural language process methods combined with deep learning methods. For example, extracting from different channel features like semantics,

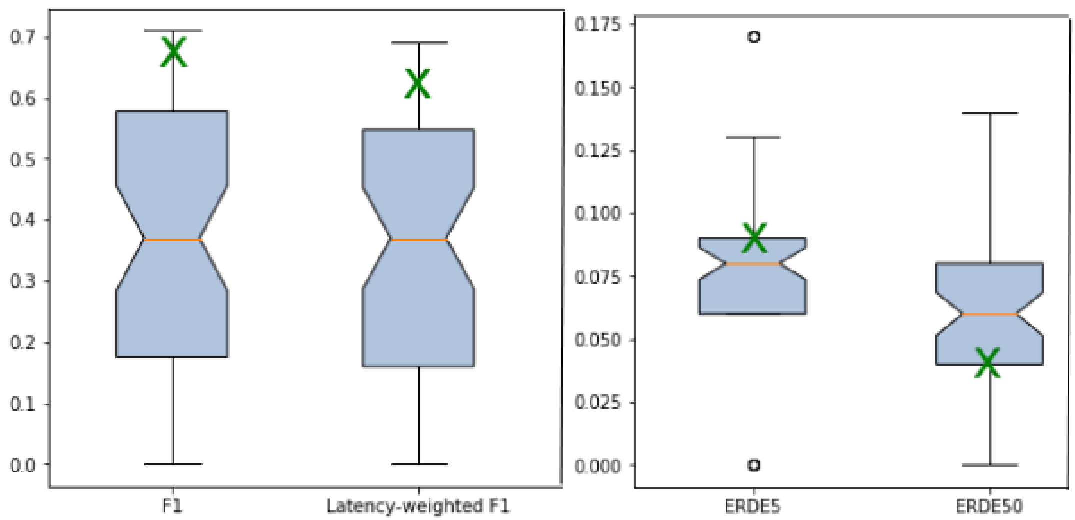

Figure 10. Boxplot for the results in F1, Latency-weighted F1, ERDE5, and ERDE50, where the green X mark represents our obtained results.

emotions or phonetic to feed a deep neural network that automatically learns how to combine these features and extract the most relevant information from it.

In the preliminary work, we proposed a new representation that creates fine-grained emotions that were automatically generated using a lexical resource of emotions and sub-word embeddings from FastText. Using these fine-grained emotions, it can automatically capture more specific topics and emotions that are expressed in the documents by users that have depression and anorexia. The emotional channel present useful information that helps the detection of mental disorders. BoSE obtained better results than the proposed baselines and also improved the results of only using broad emotions. Incorporating temporal analysis over the emotion channel and combine it with the previous representation demonstrate that helps the detection of users that presents signs of mental disorders. It is worth mentioning the simplicity and interpretability of the representation, creates a more straightforward analysis of the results.

8 Published Papers

Some of the preliminary results that are contained in this dissertation proposal are published in:

1. Detecting Depression in Social Media using Fine-Grained Emotions. Mario Ezra Aragón, A. Pastor López-Monroy, Luis C. González-Gurrola and Manuel Montes-y-Gómez. 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), Minnesota, USA. (June, 2019).

2. INAOE-CIMAT at eRisk 2019: Detecting Signs of Anorexia using Fine-Grained Emotions. Mario Ezra Aragón, A. Pastor López-Monroy and Manuel Montes-y-Gómez. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (September, 2019).

9 Background Concepts

This section describes an overview of the different techniques and core concepts needed for this dissertation proposal. For example, it introduces Text Classification, different networks related to deep learning and applied to Natural Language Processing. This section is divided as follows: first, a description of text classification and techniques for text representations like bag of words and word embeddings. Then, some of the main ideas that are behind deep learning, with some relevant neural networks useful for the representation of the data.

9.1 Text Classification

Text Classification (TC) is the process of assigning categories or tags to a text or a document according to its content, TC can be 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 [61].

Text classification can work in two different ways: i) manual, where a human annotator review the text and categorize it accordingly to how interprets the content, and 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 [62].

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 11 described a general process for Text Classification; the model receive an input text and return a label as an output.

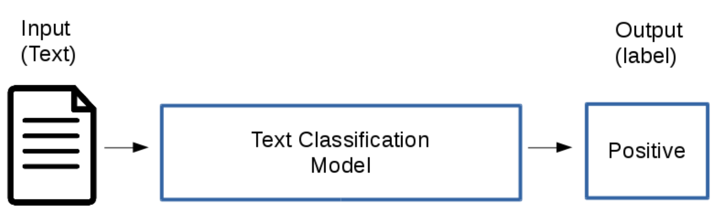

Figure 11. Text Classification General Process

9.2 Text Representation

9.2.1 Bag of Words

The bag of words (BoW) is the most simple and well know technique for text representation and classification, where the text is described by the occurrence of words within a document; first is the creation of a vocabulary $\mathbf{w}$ from training data, and then the presence of the words is measured by its frequency [64]. 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 12 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 really simple and flexible, it can be used to extract features from documents in an easy way. The intuition behind is that documents or texts present similar content if they are from 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 12. Example of a BoW Histogram Vector for the text: "John likes to watch movies. Mary likes movies too"

9.2.2 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 that are presented in the vocabulary are transformed in 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 that is presented, thus two words 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 important for solving most NLP problems and a common practice is to use these pre-trained word representations to be adjusted on the data.

There are different techniques to obtain the word embeddings, some of them are done using neural networks [22, 23, 24, 25, 26] or matrix factorization [27, 28]. One of the most common word embeddings is Word2Vec, where a neural network is used to predicts the target word from the context, for example, word(w) = ”playing” and the context = ”the musician is w the guitar”, where w is the target word that the network model learns, this model is named Continuous Bag-of-Words (CBOW) [24]. The other model names Skip-Gram (SG)[25], where the model learns the inverse prediction, it learns the word and predicts the context of the given word. 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.

Another traditional approach is Glove [28], which are embeddings trained using nonzero entries of a global word to word co-occurrence matrix.

9.3 Deep Learning

Deep learning, is a group of methods to learn representations that are known as deep architectures [11]. These methods consist of multiple layers of nonlinear units that process the data for the feature extraction and transformation. The first layers that are closer to the input data learn simple features, and the next layers learn sophisticated features extracted from the first layers. These architectures are known as hierarchical representation and are able to learn without the need for an expert in feature extraction and selection from the original data.

Conventional machine learning techniques are limited to process data in raw form. These techniques required the construction of a pattern recognition system with the 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 [12]. Using the higher layers to amplify relevant aspects of the input data for discrimination between irrelevant information and important variations. The important aspect of deep learning is that the layers of features learned from the data using a general learning procedure, instead of the designed by human experts in the domain.

In the past years, deep learning produces state of the art result in many domains; for example, they start in computer vision, speech recognition and more recently in natural language processing.

9.3.1 Recurrent Neural Network

Recurrent Neural Networks (RNN) are distinguished 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 has a purpose to preserve in the hidden state of the network the sequential information, and affect the processing of each new example to find correlations between events that are separated for different moments. RNNs are very good 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 [12].

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 affect 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 is modified by a weight matrix W and is 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 that are 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 13 presents a simple example of a RNN unit.

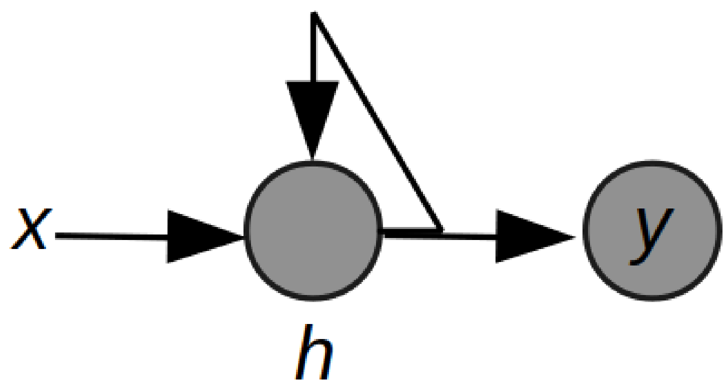

Figure 13. A simple example of a RNN unit.

9.3.2 Long Short Term Memory

Recurrent Neural Networks suffers to learn to store information for very long sequences. To solve this problem, the use of explicit memory is proposed. A long short-term memory (LSTM) networks that use special hidden units and learn to remember inputs of the sequence for a long time. These hidden units are called memory cells, a gated neuron that leaks information through the time. Each memory cell has a connection to itself at the next time step, where it copies the value of the actual state and accumulates the new values, and have a multiplicative gate by another memory cell that learns to decide to clear or keep the content of the memory [12].

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 output a number between zero and one that describes how much of each component should pass. A value closer of one means more information to let pass.

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.

9.3.3 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 designed and produce equally results in some cases [67].

GRUs use two gates name update gate and reset gate. These gates are two vectors that decide what information and the amount of information that should be pass 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 pass 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 change for a tanh function.

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 14 presents a general diagram of the different cell units of the recurrent networks. It shows the differences between the units and the way that each network let information pass.

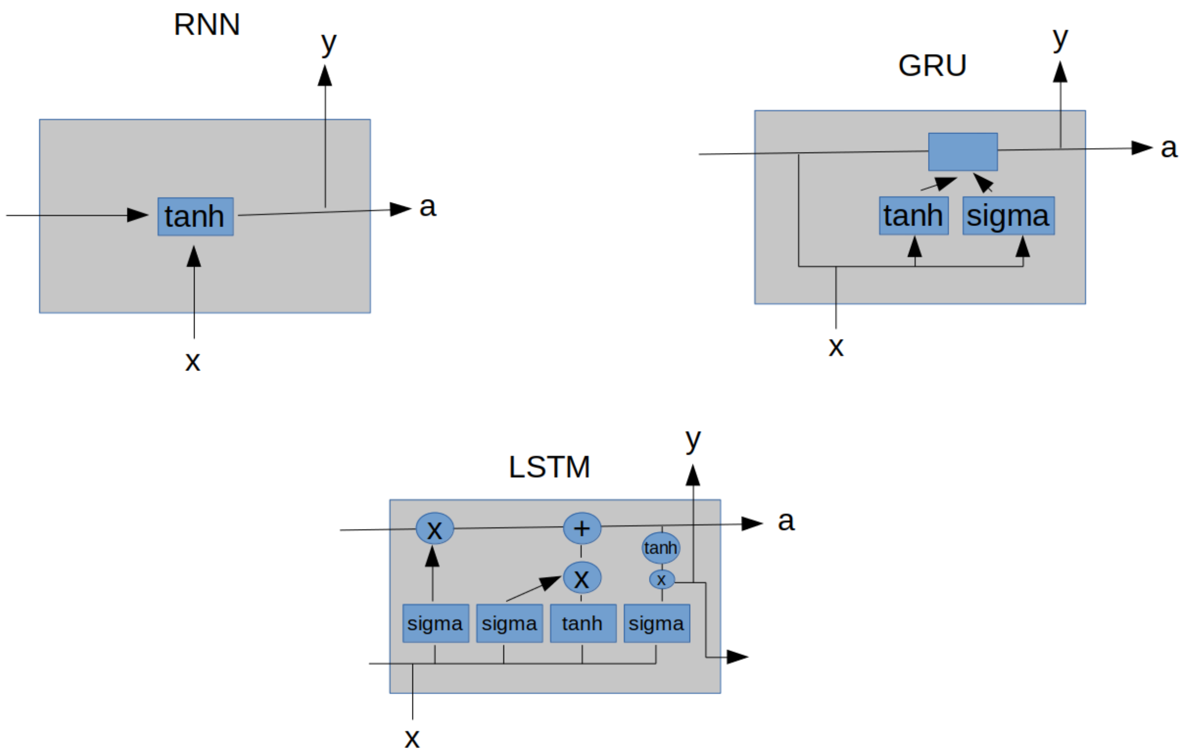

Figure 14. General Diagram of the different cell units of the RNN, LSTM and GRU.

9.4 Representation Learning

The performance of any machine learning method is mostly dependent on the choice to represent the data, also known as features. For this reason, a lot of the effort is applied in the development of designs of preprocessing and data transformation that helps in creating a representation of the data that can support the machine learning methods [68].

Learning representations of the data could make it easier to extract useful information, and make it easier to perform classification or prediction task. In deep learning, representation learning is formed by the combination of multiple non-linear transformations of the data, with the objective of creating more abstract and useful representations.

9.4.1 Autoencoder

An autoencoder is a type of unsupervised neural network. The main objective of an autoencoder is to learn a representation training to reconstruct an input data [6]. The autoencoder learns how to compress the data using the input layer (encoder) and converting it into a shortcode, and then the output layer (decoder) uncompress that shortcode into a representation that is closely matched to the original data. This helps to reduce the dimensionality of the input data, making the autoencoder to learn how to ignore the noise. Figure 15 shows a general structure of an autoencoder. Autoencoder reduces data dimension by learning how to ignore the noise in the data and learns the correlation of the input data, and perform well when compressing the features.

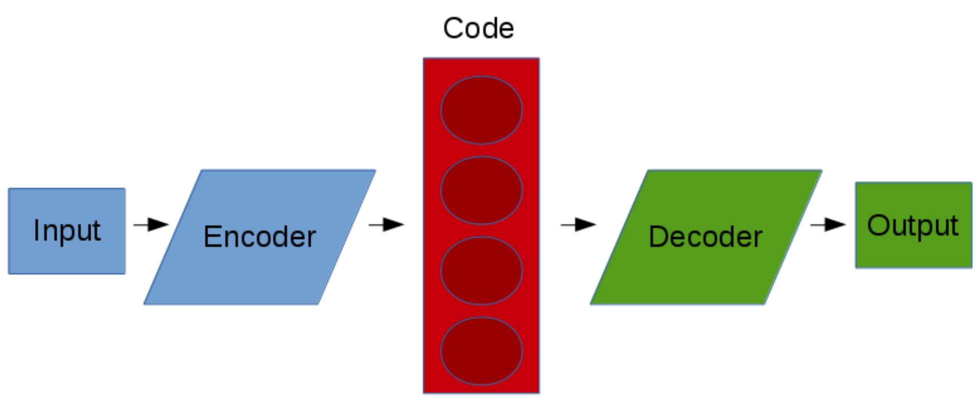

Figure 15. Diagram structure of an Autoencoder.

9.4.2 Attention Models

Attention models are networks similar to short-term memory, but these models allocate attention over the input data that they not long ago seen. The attention mechanisms are parts of the networks that learn to access memory that was storage externally instead of learning the sequences of the hidden states like the recurrent neural networks [29].

The external data that is storage works like an embedding for the attention mechanism, and this data can be altered, writing the new information that is 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, is like performing a max-pooling on all the hidden states of the network.

9.4.3 Transformers

Transformers is a neural network architecture based on self-attention mechanism, dispensing the usage of recurrence and convolutions [30]. This architecture transforms one sequence into another one using an Encoder and Decoder (discussed in a previous subsection). 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 were one of the best ways to do it. However, in recent works [31], using this kind of architectures improves the results in sequence related tasks. Figure 16 shows the general model architecture of the Transformer; the Encoder is on the left, and the Decoder is the right part. Both of the modules can be stacked on top of each other multiple times as needed (as is refer 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 that are learned during the training, and 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 first 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 is done since the model does not have recurrence to remember how the sequence was feed.

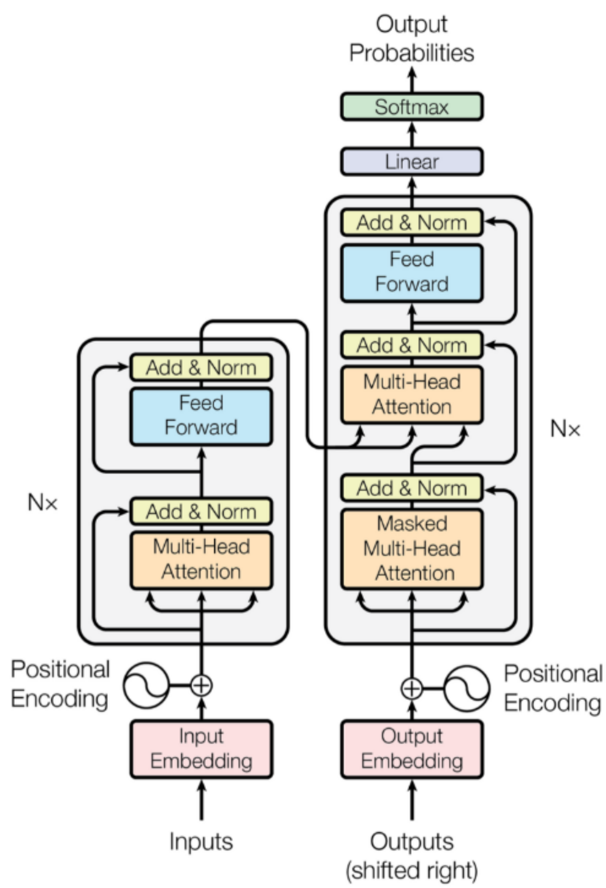

Figure 16. Transformer Model Architecture [30]

References

[1] Losada, DE., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland. (2017)

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

[3] Losada, DE., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)

[4] Canales, L., Martnez-Barco, P.: Emotion Detection from text: A Survey. Processing in the 5th Information Systems Research Working Days (JISIC) (2014)

[5] Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational. Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. (2014)

[6] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In Advances in neural information processing systems. (2007)

[7] Merikangas, KR., He, J., Burstein, M., Swanson, SA., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., Swendsen, J.: Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Study-Adolescent Supplement (NCS-A). Journal of the American Academy of Child and Adolescent Psychiatry. (2010)

[8] Canadian Reference Group: Executive Summary Spring 2016. American College Health Association. American College Health Association-National College Health Assessment II. (2016)

[9] Pestian, JP., Nasrallah, H., Matykiewicz, P., Bennett, A., Leenaars, AA.: Suicide Note Classification Using Natural Language Processing: A Content Analysislin Heidelberg. Biomed Inform Insights. (2010)

[10] Guntuku, SC., Yaden, D., Kern, M., Ungar, L., Eichstaedt, J.: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. (2017)

[11] Bengio, Y.: Learning deep architectures for AI. Foundations and trends in Machine Learning. (2009)

[12] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, no. 7553 (2015)

[13] Coppersmith, G., Harman, C., Dredze, M.: Measuring Post Traumatic Stress Disorder in Twitter. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. (2014)

[14] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V., BoydGraber, J.: The University of Maryland CLPsych 2015 shared task system. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, North American Chapter of the Association for Computational Linguistics. (2015)

[15] Preotiuc-Pietro, D., Sap, M., Schwartz, A., Ungar, L.: Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, North American Chapter of the Association for Computational Linguistics. (2015)

[16] Pedersen, T.: Screening Twitter users for depression and PTSD with lexical decision lists. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, North American Chapter of the Association for Computational Linguistics. (2015)

[17] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., Bao, Z.: A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. Springer Berlin Heidelberg. (2013)

[18] Huang, X., Zhang, L., Liu, T., Chiu, D., Zhu, T., Li, X.: Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops. (2014)

[19] Xue, Y., Li, Q., Jin, L., Feng, L., Clifton, D., Clifford, G.: Detecting Adolescent Psychological Pressures from Micro-Blog. IJCNLP. (2013)

[20] Kessler, R., Bromet, E., Jonge, P., Shahly, V., Marsha.: The Burden of Depressive Illness. Public Health Perspectives on Depressive Disorders. (2017)

[21] Mathers, C., Loncar, D.: Projections of global mortality and burden of disease from 2002 to 2030. PLOS Medicine, Public Library of Science. (2006)

[22] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research. (2003)

[23] Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In Proceedings of the International Workshop on Artificial Intelligence and Statistics. (2005)

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

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

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

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

[28] Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In Proceedings of the Conference on Empirical Methods on Natural Language Processing. (2014)

[29] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, Conference Track Proceedings. (2015)

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

[31] Devlin, J., Chang, MW., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. (2018)

[32] Tausczik, YR., Pennebaker, JW.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. (2010)

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

[34] Schwartz, HA., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.: 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. (2014)

[35] Coppersmith, G., Harman, C., Dredze, M.: Measuring post traumatic stress disorder in Twitter. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. (2014)

[36] Coopersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. Workshop on Computational Linguistics and Clinical Psychology. (2014)

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

[38] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: 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. (2015)

[39] Coppersmith, G., Ngo, K., Leary, R., Wood, A.: Exploratory analysis of social media prior to a suicide attempt. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. (2016)

[40] Benton, A., Mitchell, M., Hovy, D.: Multi-task learning for mental health using social media text. In Proceedings of European Chapter of the Association for Computational Linguistics. (2017)

[41] De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. (2013)

[42] De Choudhury, M., Counts, S., Horvitz, EJ., Hoff, A.: Characterizing and predicting postpartum depression from shared Facebook data. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing. (2014)

[43] Reece, AG., Reagan, AJ., Lix, KLM., Dodds, PS., Danforth, CM., Langer, EJ.: Forecasting the Onset and Course of Mental Illness with Twitter Data. arXiv:1608.07740. (2016)

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

[45] Ramiandrisoa, F., Mothe, J., Farah, B., Moriceau, V.: IRIT at e-Risk 2018. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[46] Ortega-Mendoza, RM., Lopez-Monroy, AP., Franco-Arcega, A., Montes-Y-Gómez, M.: PEIMEX at eRisk2018: Emphasizing Personal Information for Depression and Anorexia Detection. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[47] Ramírez-Cifuentes, D., Freire, A.: UPF's Participation at the CLEF eRisk 2018: Early Risk Prediction on the Internet. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[48] Liu, N., Zhou, Z., Xin, K., Ren, F.: TUA1 at eRisk 2018. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[49] Ragheb, W., Moulahi, B., Aze, J., Bringay, S., Servajean, M.: Temporal Mood Variation: at the CLEF eRisk-2018 Tasks for Early Risk Detection on The Internet. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[50] Wang, YT., Huang, HH., Chen, HH.: A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text. Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)

[51] Rajani S., Hanumanthappa, M.: Techniques of Semantic Analysis for Natural Language Processing A Detailed Survey. International Journal of Advanced Research in Computer and Communication Engineering. (2016)

[52] Reidy, P: An Introduction to Latent Semantic Analysis. (2009)

[53] Khosmood, F., Levinson, RA.: Automatic Natural Language Style Classification and Transformation. BCS Corpus Profiling Workshop. (2008)

[54] Eisenstein, J.: Phonological Factors in Social Media Writing. Proceedings of the Workshop on Language in Social Media. (2013)

[55] Kuncheva, L.: Combining pattern classifiers. Wiley Press, New York, 241259. (2005)

[56] Qianli, Ma., Lifeng S., Enhuan, C., Shuai, T., Jiabing, W., Garrison, C.: WALKING WALKing walking: Action Recognition from Action Echoes. Twenty-Sixth International Joint Conference on Artificial Intelligence. (2017)

[57] Ekman, PE., Davidson, RJ.: The nature of emotion: Fundamental questions. New York, NY, US: Oxford University Press. (1994)

[58] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. (2016)

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

[60] Walck, C.: Hand-book on Statistical Distributions for experimentalists. University of Stockholm, Internal Report SUFPFY/9601. (2007)

[61] Aggarwal, C.C., and Zhai, C.: A survey of text classification algorithms. In Mining text data. Springer. (2012)

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

[63] Duong, C. Lebret, R., Aberer, K.: Multimodal Classification for Analysing Social Media. arXiv:1708.02099. (2017)

[64] Goldberg, Y.: Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Graeme Hirst. (2017)

[65] Van der Maaten, L.J.P., Hinton, G.E.: Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. (2008)

[66] Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence. (2013)

[67] Cho, K., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing. (2014)

[68] Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL.35. (2014)

[69] Ocampo, M.: Salud mental en Mexico. NOTA-INCyTU NMERO 007. (2018)