# Understanding Patterns of Anorexia Manifestations in Social Media Data with Deep Learning

Ana Sabina Uban, Berta Chulvi and Paolo Rosso

Pattern Recognition and Human Language Technology (PRHLT),
Universitat Politècnica de València, València, Spain
ana.uban+acad@gmail.com, berta.chulvi@upv.es,
prosso@dsic.upv.es

## Abstract

Eating disorders are a growing problem especially among young people, yet they have been under-studied in computational research compared to other mental health disorders such as depression. Computational methods have a great potential to aid with the automatic detection of mental health problems, but state-of-the-art machine learning methods based on neural networks are notoriously difficult to interpret, which is a crucial problem for applications in the mental health domain. We propose leveraging the power of deep learning models for automatically detecting signs of anorexia based on social media data, while at the same time focusing on interpreting their behavior. We train a hierarchical attention network to detect people with anorexia, and use its internal encodings to discover different clusters of anorexia symptoms. We interpret the identified patterns from multiple perspectives, including emotion expression, psycho-linguistic features and personality traits, and we offer novel hypotheses to interpret our findings from a psycho-social perspective. Some interesting findings are patterns of word usage in some users with anorexia which show that they feel less as being part of a group compared to control cases, as well as that they have abandoned explanatory activity as a result of a greater feeling of helplessness and fear.

## 1 Introduction and Previous Work

Anorexia nervosa (AN) is a type of eating disorder that leads to multiple psychiatric and somatic complications and constitutes a major public health concern. It involves a restriction of energy intake in relation to needs, leading to significantly low body weight in relation to age, sex, developmental course and physical health. It includes among its typical symptomatology an intense fear of gaining weight or becoming fat and a distortion of one's body image (APA, 2014).

The incidence of AN, like that of other eating disorders (ED), has increased in recent decades. In a systematic literature review for the 2000-2018 period (Galmiche et al., 2019), the reported weighted means of lifetime ED (proportion of EDs at any point in life) were 8.4% (3.3–18.6%) for women and 2.2% (0.8–6.5%) for men. The authors also report that the weighted means of point ED prevalence increased over the study period from 3.5% for the 2000–2006 period to 7.8% for the 2013–2018 period. This highlights a real challenge for public health and healthcare providers.

In an attempt to understand the psychosocial origins of anorexia nervosa, some studies have investigated how body image is shaped in people suffering from this mental disorder (Giordani, 2006, 2009; Giacomozzi and da Silva Bousfield, 2011). From early research in social psychology we already know that body image, in an existential context, is the revelation of an identity that the subject constructs in the frame of concrete social relations (Goffman, 1963). From a sociological perspective, some research proposes to understand bodies attending the interaction with social forces (Turner, 2008). From anthropology, new uses of bodies (tattoos, piercings, etc) support Le Breton's idea about the study of body in modernity as an unfinished material, as "a place of self-presentation" (Le Breton, 2011). This body of research could be applied to the study of anorexia nervosa without forgetting the enormous symbolism of the act of eating. It is well established that eating with others (Dunbar, 2017) and eating the same food as the others is a major symbol of social integration (Harris, 1971; Young et al., 1971).

Mental health disorders in general, as a very significant public health matter (World Health Organization, 2012), have received attention in previous research in computational studies as well. The majority of research has focused on the study of depression (De Choudhury et al., 2013; Eich-



*Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology*, pages 224–236

staedt et al., 2018; Abd Yusof et al., 2017; Yazdavar et al., 2017), but other mental illnesses have also been studied, including generalized anxiety disorder (Shen and Rudzicz, 2017), schizophrenia (Mitchell et al., 2015), post-traumatic stress disorder (Coppersmith et al., 2014, 2015), risks of suicide (O'dea et al., 2015), and self-harm (Losada et al., 2019; Yang et al., 2016).

For anorexia, there are very few studies approaching the problem from a computational perspective. To our knowledge, the only publicly available social media dataset dedicated to anorexia is the eRisk dataset (Losada et al., 2019). The winners of eRisk's shared task on anorexia detection (Mohammadi et al., 2019) used a hierarchical attention network and obtain a state-of-the-art F1 score of 0.71. In (Cohan et al., 2018) the authors introduce a dataset annotated for multiple mental disorders including anorexia. Another study (Amini and Kosseim) on the explainability of anorexia detection models analyzes attention weights of a neural network to show that attention at the user level correlates with the importance of individual texts for classification performance.

Explainability of machine learning models, especially in the field of mental health, is a very important issue. In practice, models based on neural networks are vastly successful for most NLP applications, even though they have been only briefly explored in existing computational studies on mental disorders. Nevertheless, neural networks are notoriously difficult to interpret. While there is increasing interest in the field of explainability of machine learning models including in NLP (Gilpin et al., 2018), there are fewer such studies for mental health disorder detection.

In the name of transparency, it is essential for any automatic system that can assist with mental health disorder detection to make its decision-making process understandable. Especially in the medical domain, using black-box systems can be dangerous for patients (Zucco et al., 2018; Holzinger et al., 2017). Moreover, recently, the need of explanatory systems is required by regulations like the General Data Protection Regulation (GDPR) adopted by the European Union.

While many quantitative studies in the computational analysis of mental health use features such as lexicons (Trotzek et al., 2017; De Choudhury et al., 2014) to study the manifestations of mental disorders in user-generated data, these models are very limited computationally in comparison to deep learning models. The behavior of powerful classifiers modelling complex patterns in the data has the potential to help uncover manifestations of the disease that are potentially difficult to observe with the naked eye, and thus be useful not only as tools for the detection of disorders, but also as analysis instruments for generating insights and potentially assisting clinicians in the diagnosis process.

In our study, we propose using deep learning as a tool to aid with a deeper investigation of anorexia manifestations in social media texts. We train a hierarchical attention network to classify people with anorexia against control cases based on their social media activity. This architecture has been shown to provide good results for anorexia detection, and additionally includes in the model a series of features that encode different levels of the language (style, emotions, topics etc). To our knowledge this has not been done in previous work, and allows us the advantage of a more interpretable model. We interpret the predictions of the network as well as its hidden layers as a way to identify different patterns of anorexia symptoms in social media users, which we analyze in view of the different features, and offer hypotheses on the different patterns observed from a psychological perspective. Thus, we aim to answer the following research questions:

**RQ1.** Is it possible to leverage complex deep learning models and their encoding power in order to identify different patterns of anorexia symptoms in social media texts?

**RQ2.** Can we characterize the differences between different groups of people with anorexia based on psycho-linguistic and emotion features, and measures of personality traits?

**RQ3.** Could we identify some features to explore the hypothesis that anorexia nervosa is a way to express some degree of conflict with one's own group?

## 2 Classification Experiments

In order to explore the proposed hypotheses, we start by building a deep learning model in order to automatically classify texts belonging to users with anorexia and control cases.

### 2.1 Dataset

**eRisk Reddit dataset on anorexia.** The eRisk CLEF lab[1] is focused on the early prediction of

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

mental disorder risk from social media data, focused on disorders such as depression (Losada et al., 2018), anorexia and self-harm tendencies (Losada et al., 2019, 2020). Data is collected from Reddit posts and comments selected from specific relevant sub-reddits. Users suffering from a mental disorder are annotated by automatically detecting self-stated diagnoses. Control cases are selected from participants in the same sub-reddits (having similar interests), thus making sure the gap between healthy and diagnosed users is not trivially detectable. A long history of posts are collected for the users included in the dataset, up to years prior to the diagnosis. The dataset on anorexia, released as part of eRisk 2019 (Losada et al., 2019), contains 823,754 posts collected from 1,287 users, of which 10.4% are anorexic users.

### 2.2 Model and Features

We choose a hierarchical attention network (HAN) as our model: a deep neural network with a hierarchical structure, including multiple features encoded with LSTM layers and two levels of attention. HANs have previously shown to be successful for the detection of anorexia in social media. (Amini and Kosseim; Mohammadi et al., 2019).

The HAN is made up of two components: a *post-level encoder*, which produces a representation of a post, and a *user-level encoder*, which generates a representation of a user's post history. The post-level encoder and the user-level encoder are modelled as LSTMs. The word sequences encoded using embeddings initialized with GloVe pre-trained embeddings (Pennington et al., 2014) and passed to the LSTM are then concatenated with the other features to form the hierarchical post encoding. The obtained representation is passed to the user-encoder LSTM, which is connected to the output layer. We use the train/test split provided by the shared task organizers, done at the user level, making sure users occurring in one subset don't occur in the other. Since individual posts are too short to be accurately classified, we construct our datapoints through concatenating groups of 50 posts, sorted chronologically. We use a weighted loss function to compensate for the class imbalance. The architecture of the model is shown in Figure 1.

We represent social media texts using features that capture different levels of the language (semantic, stylistic, emotions etc.) and train the model to predict anorexia risk for each user.

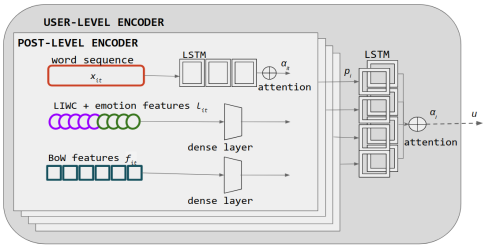



Figure 1: Model architecture.

**Content features**. We include a general representation of text content by transforming each text into word sequences, represented as embeddings.

**Style features**. The usage pattern of function words is known to be reflective of an author's style, at an unconscious level (Mosteller and Wallace, 1963). As stylistic features, we extract from each text a numerical vector representing function words frequencies as bag-of-words, which are passed through an additional dense layer of 20 units. We complement function word distribution features with other syntactical features extracted from the LIWC lexicon, as described below.

**LIWC features.** The LIWC lexicon (Pennebaker et al., 2001) has been widely used in computational linguistics as well as some clinical studies for analyzing how suffering from mental disorders manifests in an author's writings. LIWC is a lexicon mapping words of the English vocabulary to 64 lexico-syntactic features of different kinds, with high quality associations curated by human experts, capturing different levels of language: including style (through syntactic categories), emotions (through affect categories) and topics (such as money, health or religion).

**Emotions and sentiment.** We dedicate a few features to representing emotional content in our texts, since the emotional state of a user is known to be highly correlated with her mental health. Aside from the sentiment and emotion categories in the LIWC lexicon, we include a second lexicon: the NRC emotion lexicon (Mohammad and Turney, 2013), which is dedicated exclusively to emotion representation, with categories corresponding to a wider and a more fine-grained selection of emotions, containing the 8 Plutchik's emotions (Plutchik, 1984), as well as *positive/negative* sentiment categories: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*. We represent LIWC and NRC features by computing for each category the proportion of words in the input text which are associated with that category.

| Model | P | R | F1 | AUC |
|---|---|---|---|---|
| **HAN** | .60 | .63 | .60 | **.96** |
| **RoBERTa** | .64 | .69 | .70 | .83 |
| **AlBERT** | .78 | .54 | .65 | .77 |
| **LogReg** | .55 | .45 | .49 | .90 |

Table 1: Precision, recall, F1 (positive class) and AUC scores anorexia classification.

These are concatenated with the other features to form the post-level encodings, which are then stacked and passed to the final user-level LSTM which is connected to the output layer.

### 2.3 Classification Results

The results obtained with our neural network for the detection of anorexic users are shown in Table 1. As performance metrics we compute the F1-score of the positive class and the area under the ROC curve (AUC), which is more robust in the case of data imbalance. We compare the results of our model with baselines such as a logistic regression model with bag-of-word features, and transformer-based models including RoBERTa (Liu et al., 2019) and AlBERT (Lan et al., 2019) with word sequences as features.

Our HAN model achieves the best results in terms of AUC. In the following sections, we explore this model in more depth in order to explain its behavior and leverage it to discover insights on linguistic patterns associated with anorexia.

## 3 Explaining Predictions

In this section we present different analyses meant to uncover insights into how the model arrives at its predictions, first looking at the attention weights and abstract internal representations of the data in the layers of the neural network, and secondly providing several feature-focused analyses, using emotions and LIWC categories, as well as personality markers.

### 3.1 Attention Analysis

Attention is a mechanism frequently used in recurrent neural networks in order to weigh the parts of the input sequence differently according to their importance for prediction. Attention weights are learned by the network, and thus can be used as a means to interpreting its decisions. In our models we include recurrence at the user level, along with an attention layer, which can thus be used to infer the weight placed on each part in a user's post history by the neural network.

>>> oh wow thats so i saw you ve already
lost bunch too i m doing ok haven t been doing too great past few days
because of anxiety being at higher than i normally weight but i m hoping to some
progress and get back down where i was before
>>> same i also hate lower stomach too
>>> awh ok thanks
>>> i discovered vegan
butter a few ago and omg my worst binge is toast with vegan butter on i
could eat the entire loaf that stuff on it
>>>
think of this way even if you feel you look overweight if your teachers then you
must not be its to not be ready for treatment you can see a therapist without

Figure 2: Attention weights example.

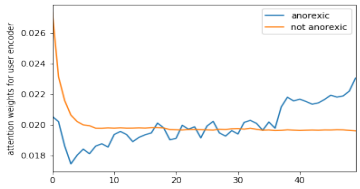



Figure 3: Attention weights over time

In Figure 2, we show an example post (with noise added in view of anonymizing the author) with words and sentences highlighted according to the attention weights provided by the neural network , showing in green the importance of words in each post, and in yellow the importance of each post. In Figure 3, we plot the attention weights for the user-level attention layer for each of the classifiers trained on the three datasets. For this experiment, we train the neural network on one datapoint for each user, so as to ensure attention weights consistently correspond to the same part of the post history for each training example. The plot shows a general increasing importance for users suffering from anorexia: posts in the end of a user's history are more heavily weighted. This is an interesting finding, since intuition, supported by findings such as those presented in the previous sections related to emotion evolution, would suggest a user's activity on social media becomes increasingly indicative of their mental state as time goes by.

Recent studies (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019) have questioned whether attention mechanisms necessarily help with the interpretability of neural network predictions. We further explore additional techniques in order to interpret the representations learned by the model.

### 3.2 User Embeddings

We continue explaining the model's behavior by analyzing the internal representations of the neural network. We can regard the final layer of the trained neural network as the most compressed representation of the input examples, which is, in terms of our trained model, the optimal representa-

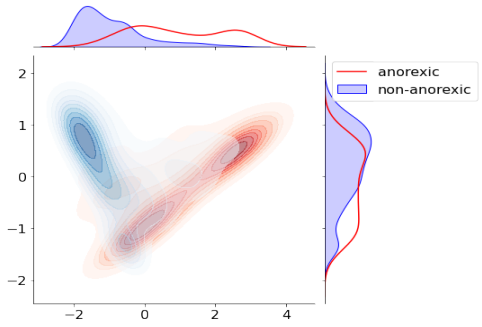
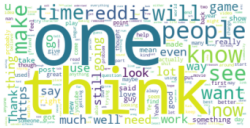



Figure 4: User embeddings in 2D.

(a) Cluster of users with anorexia **ANO1**.

(b) Cluster of users with anorexia **ANO2**.



(c) Cluster of control cases.

Figure 5: Word clouds for the different clusters of anorexic and healthy users.

tion for distinguishing between control cases and those suffering from a disorder. Thus, the final layer (the output of the 32-dimensional user-level LSTM) can be interpreted as a 32-dimensional embedding for the input points, corresponding to the users to be classified.

We analyze the output of the *user embedding* layer by reducing it to 2 dimensions using principal component analysis (PCA) and visualizing it in 2D space with a kernel density estimate (KDE) plot to show the distribution of scores across the 2 dimensions (Figure 4). We make sure to train the PCA model on a balanced set of positive and negative users, then we extract 2D representations for all users in the test set. By looking at these representations, we can gain insight into the separability of the classes, from the perspective of the trained model, and better understand where it encounters difficulties in separating between the datapoints belonging to different classes.

We notice that, even in two dimensions, the groups of people with anorexia and control cases appear as separate clusters in the user embedding space, suggesting that the encodings generated by the model while training for its objective are powerful enough to separate the two groups.

We notice an interesting bimodal aspect in the distribution of positive users, which appear clustered into two distinct groups, while control cases' representations seem to be more compact. One of the clusters of people with anorexia is more clearly distinct from control cases in this space (ANO1), while the other cluster shows a higher overlap with control cases, presumably leading to false negatives in the model's predictions (ANO2). We suggest that the two observed clusters might represent two different groups of people with anorexia and of patterns of symptoms of anorexia, which the model is able to capture during the training process.

The problem of false negative predictions in particular is important to study, since false negatives can lead to missing cases of people with high risk of anorexia, and, left undetected, the disorder might further develop. In the following sections we take a deeper dive into the three identified clusters and attempt to interpret from different perspectives the differences among the clusters of people with anorexia and control cases.

## 4 Patterns of Anorexia

### 4.1 Clustering Method

We begin the analysis of the different patterns of anorexia manifestations by attempting to automatically identify the clusters of people with anorexia in the user embedding space. We perform k-means clustering on the user embeddings for users with anorexia in the test set, and cluster the user embeddings into 2 groups. We obtain 31 users in one cluster which we denote as **ANO1**, and 43 in the second, denoted as **ANO2**.

As a first analysis, we take a look at the vocabulary of the three groups. Figure 5 illustrates word clouds corresponding to the vocabulary used by the 3 groups. As we can see, in both groups where users with anorexia are present, the word *people* has a greater presence suggesting it would be interesting to go deeper into analyzing the dynamics related to social relations among users with anorexia. In order to move in this direction, we will use different strategies that try to describe in more depth these three clusters and to identify the most common narratives in these clusters.

### 4.2 Emotions and Psycho-linguistic Features across Clusters

We analyze the three groups from the perspective of usage of emotion words and psycho-linguistic

| | ANO1 | ANO2 | Control |
|---|---|---|---|
| **ingest**$^{***}$ | 1.15 | 0.82 | 0.38 |
| **bio**$^{***}$ | 3.47 | 2.82 | 1.49 |
| **health**$^{***}$ | 1.01 | 0.81 | 0.39 |
| **body**$^{**}$ | 0.88 | 0.77 | 0.51 |
| **anx**$^{***}$ | 0.47 | 0.37 | 1.19 |
| **negemo**$^{**}$ | 2.78 | 2.53 | 1.66 |
| **disgust**$^{**}$ | 1.25 | 1.31 | 0.90 |
| **fear**$^{*}$ | 1.81 | 1.66 | 1.71 |

Table 2: Features about typical symptoms of anorexia and negative emotional states, percentage of average usage per cluster.
$^{***}$Statistically significant difference across the three clusters
$^{**}$Statistically significant difference between people suffering from anorexia and control users.
$^{*}$Statistically significant difference between **ANO1** and others

categories in the LIWC lexicon. Similarly to the way we encode these features as inputs to the deep learning model, we compute the average values of prevalence (as percentages of overall word usage) for words in each category, separately for the texts in each cluster. We then identify features where there are statistically significant differences among the groups (using a *t*-test).

We identify separately features which show significant differences in usage across all three clusters, or just between users with anorexia and control cases. In general, we observe a pattern of **ANO2** having intermediate values between users with anorexia and control cases, for most of the significant features (36 out of 75 categories show statistically significant differences among all three clusters, and for 30 of them the values for **ANO2** are situated between **ANO1** and control cases).

To obtain a deeper interpretation of the observed differences we select features which are most relevant to anorexia symptoms (see Table 2) and the features which refer to negative emotions. Interestingly, these features also show a distinct pattern from an error analysis perspective: if we select those categories which have lower values in misclassified examples in a statistically significant way, we obtain these four LIWC categories: *anx*, *health*, *bio* and *ingest*.

### 4.3 Personality Analysis

As a separate analysis, we try to analyze the different clusters from the perspective of personality types using the Five Factor Model. The Five Factor Model is a process of attributing certain psychological characteristics to an individual according to the so-called 'Big Five' taxonomy that has been developed into a laborious research paradigm initiated by the social psychologist Gordon Allport (Allport, 1937). Allport (1937) formulated *The lexical hypothesis* proposing that most of the socially relevant and salient personality characteristics have become encoded in the natural language (John et al., 1999). After decades of research, the field is approaching consensus on a general taxonomy of personality traits, the "Big Five" personality dimensions. The five factors are openness to experience (1) conscientiousness (2) agreeableness (3) extraversion (4) and neuroticism (5), as emotional stability. Exploiting this theoretical framework to extract information about users' personality from their posts means identifying such semantic associations and mapping the text around the five factors according to the words referring to them. An effective approach to do this consists in the one proposed by Neuman and Cohen (2014): the evidences of a particular personality trait are summarised into a score, which is calculated as the semantic similarity between the context-free word embedding representations respectively of the text written by the author and of the set of the benchmark adjectives (i.e., the terms empirically observed to be able to encode each of the five personality aspects according to the 'Big Five' framework). In more detail, for each trait, Neuman and Cohen (2014) define a positive and a negative sub-dimension, which correspond respectively to the possession of a sufficient degree of a given factor or, vice versa, the evidence of the exact opposite characteristic. Neuman and Cohen (2014) associate a small series of benchmark adjectives to all the 19 sub-dimensions. In the Appendix we list in full the adjectives that make up the vectors associated with each personality trait.

The set of the benchmark adjectives for the personality traits proposed by Neuman and Cohen has been successfully employed in other tasks such as profiling fake news spreaders (Giachanou et al., 2020). We use a similar approach, and measure personality scores by computing the overlap of the words in the defined vectors for each trait with the words used in each text in our dataset, normalized by text length. Following this approach we found significant differences (p-value $<.005$) between users with anorexia and control cases in three factors: *Agreeableness (+) (-)*, *Extraversion (-)*, *Neuroticism (+)*.

As we can see in Table 3, in *Extraversion (-)* the difference is statistically significant when we

| | ANO1 | ANO2 | Control |
|---|---|---|---|
| **EXT+** | 0.0037 | 0.0076 | 0.0038 |
| **EXT-**** | 0.23 | 0.82 | 0.41 |
| **AGR+**** | 0.79 | 0.82 | 0.41 |
| **AGR-**** | 0.84 | 1.07 | 0.68 |
| **NEUR+***** | 0.47 | 0.28 | 0.11 |
| **NEUR-** | 0.0081 | 0.18 | 0.10 |
| **CON+** | 0.28 | 0.24 | 0.22 |
| **CON-** | 0.10 | 0.12 | 0.14 |
| **OPN+** | 0.25 | 0.42 | 0.29 |
| **OPN-** | 0.12 | 0.20 | 0.13 |

Table 3: Personality-related words, usage per-mille across the clusters.
***Statistically significant difference across the three clusters
**Statistically significant difference between people suffering from anorexia and control users.
*Statistically significant difference between **ANO1** and others

| | ANO1 | ANO2 | Control |
|---|---|---|---|
| **work**** | 1.22 | 1.47 | 2.31 |
| **money**** | 0.41 | 0.50 | 0.86 |
| **leisure**** | 1.12 | 1.15 | 1.88 |
| **pronoun***** | 17.41 | 16.20 | 11.54 |
| **I***** | 6.95 | 5.52 | 3.49 |
| **we*** | 0.33 | 0.46 | 0.51 |
| **friend**** | 0.22 | 0.25 | 0.15 |
| **family**** | 0.27 | 0.28 | 0.22 |
| **humans**** | 0.82 | 0.92 | 0.72 |

Table 4: Features about everyday activities and social relations, percentage of average usage per cluster.
***Statistically significant difference across the three clusters
**Statistically significant difference between people suffering from anorexia and control users.
*Statistically significant difference between **ANO1** and others

compare users suffering from anorexia with control cases, but not between the two clusters of people with anorexia. It seems that more introverted personality traits characterize users who suffer from anorexia. The same applies to the factor *Agreeableness*, in positive and in negative sense, the difference in agreeableness is statistically significant among people with anorexia and control cases. Users suffering from anorexia show more characteristic traits of a pleasant personality and unplesant personality than those who do not suffer from this disorder. The explanation for this difference, in positive and in negative traits, may be related to a greater manifestation of emotions and feelings among people with anorexia.

With the factor measuring *Neuroticism (+)* we find statistically significant differences among the three clusters suggesting that users in **ANO1** are at a more severe stage of this mental disorder than **ANO2** because they have higher scores in this factor. Neuroticism speaks about emotional instability that leads to frequent experiences of negative emotions and which is said to result from a low ability to handle stress or strong external stimuli.

## 5 Identifiying Different Narratives and Cognitive Styles

In **RQ3** we raised the possibility of exploring if some features from LIWC and emotion lexicons allow us to identify among users with anorexia a higher degree of conflict with their own group and some degree of social isolation.

As we can see in Table 4, users with anorexia talk less about *work*, *money* and *leisure* (LIWC categories) than control cases. The absence of words from these three categories tells us about a certain degree of social isolation. These results connect with a pattern that we find in the use of *personal pronouns* among the different clusters.

As we can observe, users with anorexia employ significantly fewer words under the category *we* (*we*, *us*, *our*), a clear linguistic marker of a sense of belonging. Users in **ANO1** use it significantly less than users in **ANO2** and than control cases, suggesting that a higher degree of conflict with one's own group or a higher degree of social isolation may be linked to the more severe manifestations of anorexia. The opposite pattern is found in the use of the first person pronouns that LIWC collects under the category *I* (*I*, *me*, *mine*): users suffering from anorexia use it significantly more than control cases and cluster **ANO1** uses it significantly more than cluster **ANO2**.

Three other features expressing social processes, *family*, *friends* and *humans* are more present in the narratives of users with anorexia than among control cases. The greater presence of these linguistic categories may indicate a greater centrality of social relations in the identity of these subjects suffering from anorexia. We would need to design new strategies to go deeper into this interpretation, but this greater centrality of social relations categories in people with anorexia could be derived, precisely, from a greater degree of conflict with the social environment.

We also observed in Table 5 some differences among clusters in relation to some LIWC categories related to cognitive and perceptual processes. These differences may indicate the existence of different cognitive styles between users who suffer from anorexia and those who do not.

*Cognitive style* is a concept used in cognitive

psychology to describe the way individuals think, perceive, and remember information (Grigorenko and Sternberg, 1995). Research in psychology suggests that some cognitive styles are more prevalent in some patients suffering from depression and anorexia (Lo et al., 2008; Kaye et al., 1995).

As we can see in Table 5, people with anorexia use in their narratives more words that LIWC classifies as *cognitive processes* (*cogmech*) than control cases and the difference is also statistically significant between clusters **ANO1** and **ANO2**. A greater presence of these traits among users with anorexia is indicative of a special effort to reason about reality, which is also a characteristic of conflictual states. We could also consider that this effort to understand involves the subject at a personal level because we see a higher presence of words belonging to the *feel* category among users with anorexia. *Feel* is, among the perceptual process in LIWC, the one that involves the subject at a deeper level.

Within *cognitive processes* we find it interesting to analyze certain features such as *certainty*, *tentative* and *causation* where there are significant differences among clusters. *Certainty* expresses a rigid or absolute style of thinking and *tentative* expresses a more flexible or less absolute style of thinking. We find significant differences in these two categories between users with anorexia and control cases. *Certainty* is more used in narratives from users suffering anorexia and could indicate a major degree of cognitive conflict. We see the opposite pattern with *tentative*, that just expresses a flexible style. It could be expressing the two sides of the same phenomenon.

With *causation* we only found statistically significant differences between **ANO1** and the other two clusters (similarly to what we found for the expression of fear, as seen in Table 2). Reasoning about causes indicates an effort to understand the world that shows a healthier position of the subjects, and one possible interpretation is that users in **ANO1** have abandoned this explanatory activity as a result of a greater feeling of helplessness and feeling more fear.

## 6 Connection to Clinical Research

Trying to identify different groups of patients who claim to suffer from anorexia nervosa or have compatible symptomatology may be a way to develop a better understanding of this complex pathology (Clinton et al., 2004). Research such as that of Viborg et al. (2018) shows a relationship between six clusters of young adolescents suffering from anorexia and higher levels of psychological difficulties and lower levels of body esteem. However, these clusters rely exclusively on symptomatology linked to eating behavior. We think that our results show that a deep learning model applied to social media texts allows us to identify clusters of patients considering more variables than those related to the eating disorder itself.

From these initial results we can provide some insights to clinical discussion. More and more clinicians (Gutiérrez and Carrera, 2021) ask themselves about the intractability of anorexia nervosa, including the disconcerting aspect of the recovery of a significant number of patients not receiving formal treatment (Keski-Rahkonen, 2014). As Gutiérrez and Carrera (2021) state in a recent review of this issue, Bruch's proposal regarding the charactization of typical anorexia in terms of *Body Image Disturbances* (BID) (Bruch et al., 1974) and the relevance of this construct in different editions of the *Diagnostic and Statistical Manual of Mental Disorders* (APA, 2014), may have over-directed clinical practice, rendering other aspects of psychopathology invisible and increasing the numbers of patients diagnosed as atypical cases.

Analyzing the language of people suffering from anorexia, we found traces of problems that could guide new approaches indicating the existence of a conflict between the patient and his or her own group: lower use of the first person plural and a greater presence of features expressing social processes. Some research focusing on the discourse analysis of people suffering from anorexia points to the existence of a link between acting on one's own body as a mechanism to take control over their lives (Malson and Ussher, 1996).

Our results may indicate that a possible origin of this need of control could be the social conflict with one's own group and the inability to communicate this conflictual situation. In this sense, Botta and Dumlao (2002) have shown the relationship between parent-child conflict and family communication styles and the development of eating disorders. In addition, Davies et al. (2012) have demonstrated in their experimental research the relevance of a verbal expression of emotions in patients with anorexia nervosa. More research is needed, for instance it may be interesting to revisit one of Bruch's ideas in her seminal work which has not been as

| | ANO1 | ANO2 | Control |
|---|---|---|---|
| **cogmech**$^{***}$ | 16.43 | 15.88 | 14.58 |
| **feel**$^{**}$ | 0.86 | 0.86 | 0.43 |
| **certain**$^{**}$ | 1.55 | 1.69 | 1.04 |
| **tentative**$^{**}$ | 3.09 | 3.01 | 3.13 |
| **causation**$^{*}$ | 1.71 | 1.85 | 1.87 |

Table 5: Features about cognitive styles (cognitive processes and perceptual processes), percentage of average usage per cluster.
$^{***}$Statistically significant difference across the three clusters
$^{**}$Statistically significant difference between people suffering from anorexia and control users.
$^{*}$Statistically significant difference between **ANO1** and others

successful as her concept of BID: the description of anorexia as "a communicative disorder" which is experienced as a means of taking control over one's body as a pseudosolution to intra- and interpersonal difficulties (Bruch, 1978). Following this idea, we think that deep learning models applied to social media data can open an interesting avenue to explore the language of people suffering from anorexia and provide elements for further clinical discussion.

## 7 Ethical Considerations

Powerful machine learning models that can be trained to detect or predict the development of mental health disorders can be very valuable, but any deployment of a tool for mental disorder detection should take into account potential ethical concerns. If such tools are used by third parties (such as employers seeking to filter candidates based on their mental health profile), this could compromise the privacy of the subjects. We suggest that the development of an ethical standard is necessary, and that launching such tools could be accompanied by an ethical statement to constrain its use.

Moreover, it is ethically necessary, and recently even required by regulations in some countries (such as countries in the European Union) that artificial intelligence models used in the mental health domain have interpretable behavior. We hope that we were able to take a step forward in this direction, by providing an in-depth explanation of the representations generated by our neural network, and thus facilitating trust in the predictive model.

## 8 Conclusions and Future Work

In this study, we have approached the problem of detecting people suffering from anorexia in social media thorugh training a deep learning model, and taken it a step further by explaining the behavior of the model. Based on this, we identified different clusters of users suffering from anorexia with regards to the manifested symptoms, as encoded by the trained model **(RQ1)**. We presented several analyses for interpreting the decisions of the model trained to profile social media users at risk of developing anorexia, going beyond more common techniques such as attention weight analysis, and including hidden layer analysis and error analysis at different levels of the language for better understanding how mental disorders manifest in social media data **(RQ2)**.

We have shown that we can interpret the detected clusters from the point of view of the social behavior of people with anorexia **(RQ3)**, and provided in-depth interpretations of these patterns as a socio-psychological phenomenon. To our knowledge, our approach and findings are novel in the domain of the computational study of anorexia, and encourage us to go deeper into the analysis of these patterns, such as looking into the use of pronouns in relation with emotions, which could help identify more clearly the existence of a social conflict.

From a technical perspective, a more sophisticated analysis of the features included here (LIWC categories, emotions and personality vectors) could be achieved by using more semantically rich representations than bag-of-words. One such approach would be using word embeddings to identify sub-emotions (Aragón et al., 2019) starting from Plutchik's 8 emotions.

Moreover, including a temporal dimension as a variable could also reveal additional insights such as cases of patients with symptoms moving from one cluster to another over time. The possibility to categorize people with anorexia into different groups according to their symptoms might help with identifying those people at a higher risk of more serious developments of their disorder and also could give some inputs for for a more in-depth discussion of symptomatology in clinical forums.

Finally, it would also be interesting to explore the connection between anorexia and other mental disorders or manifestations such as suicide attempts.

## Acknowledgements

The authors thank the EU-FEDER Comunitat Valenciana 2014-2020 grant IDIFEDER/2018/025. The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

## References


Noor Fazilla Abd Yusof, Chenghua Lin, and Frank Guerin. 2017. Analysing the causes of depressed mood from depression vulnerable individuals. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 9–17.

Gordon Willard Allport. 1937. Personality: A psychological interpretation.

Hessam Amini and Leila Kosseim. Towards explainability in using deep learning for the detection of anorexia in social media. *Natural Language Processing and Information Systems*, 12089:225.

APA. 2014. *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. American Psychiatric Association.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486.

Renee Botta and Rebecca Dumlao. 2002. How do conflict and communication patterns between fathers and daughters contributeto or offset eating disorders? *Health communication*, 14:199–219.

Hilde Bruch. 1978. *The golden cage: The enigma of anorexia nervosa*. Harvard University Press.

Hilde Bruch et al. 1974. *Eating disorders. Obesity, anorexia nervosa, and the person within.* Routledge & Kegan Paul.

David Clinton, Eric Button, Claes Norring, and Robert Palmer. 2004. Cluster analysis of key diagnostic variables from two independent samples of eating-disorder patients: Evidence for a consistent pattern. *Psychological Medicine*, 34(6):1035.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Helen Davies, Nicola Swan, Ulrike Schmidt, and Kate Tchanturia. 2012. An experimental investigation of verbal expression of emotion in anorexia and bulimia nervosa. *European Eating Disorders Review*, 20(6):476–483.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

R. I. M. Dunbar. 2017. Breaking bread: the functions of social eating. *Adaptive Human Behavior and Physiology*, 3:198–211.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Marie Galmiche, Pierre Déchelotte, Gregory Lambert, and Marie Tavolacci. 2019. Prevalence of eating disorders over the 2000-2018 period: a systematic literature review. *The American journal of clinical nutrition*, 109:1402–1413.

Anastasia Giachanou, Esteban A Ríssola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. 2020. The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In *International Conference on Applications of Natural Language to Information Systems*, pages 181–192. Springer.

Andréia Isabel Giacomozzi and Andréa Bárbara da Silva Bousfield. 2011. Representação social do corpo de participantes de comunidades pró-anorexia do orkut. *Psicologia, Saúde e Doenças*, 12(2):255–266.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Rubia C.F. Giordani. 2006. A auto-imagem corporal na anorexia nervosa: uma abordagem sociológica. *Psicologia & Sociedade*, 18:81–88.

Rubia C.F. Giordani. 2009. O corpo sentido e os sentidos do corpo anoréxico. *Revista de Nutrição*, 22:809–821.

Ervin. Goffman. 1963. *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs, N.J: Prentice-Hall.

Elena L Grigorenko and Robert J Sternberg. 1995. Thinking styles. In *International handbook of personality and intelligence*, pages 205–229. Springer.

Emilio Gutiérrez and Olaia Carrera. 2021. Severe and enduring anorexia nervosa: Enduring wrong assumptions? *Frontiers in Psychiatry*, 11:1–19.

Marvin Harris. 1971. *Culture, man, and nature: An introduction to general anthropology*. Crowell.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Oliver P John, Sanjay Srivastava, et al. 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*, volume 2. University of California Berkeley.

Walter H. Kaye, Andrea M. Bastiani, and Howard Moss. 1995. Cognitive style of patients with anorexia nervosa and bulimia nervosa. *International Journal of Eating Disorders*, 18:287–290.

Anna et al. Keski-Rahkonen. 2014. Factors associated with recovery from anorexia nervosa: a population-based study. *The International journal of eating disorders*, 47(2):117–23.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

David Le Breton. 2011. *Anthropologie du corps et modernité*. Presses universitaires de France.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Cola Lo, Samuel M.Y.Ho, and Steven D.Hollonb. 2008. The effects of rumination and negative cognitive styles on depression: A mediation analysis. *Behaviour Research and Therapy*, 46(4):487–495.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 343–361. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2696.

Helen Malson and Jane M Ussher. 1996. Body polytexts: Discourses of the anorexic body. *Journal of community & applied social psychology*, 6(4):267–280.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Quick and (maybe not so) easy detection of anorexia in social media posts. In *L. Cappellato, N. Ferro, D. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2380.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Yair Neuman and Yochai Cohen. 2014. A vectorial semantics approach to personality assessment. *Scientific reports*, 4(1):1–6.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2017. Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In *L. Cappellato, N. Ferro, L. Goeuriot and T. Mandl (eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 1866.

Bryan Turner. 2008. *The body & society: Explorations in social theory*. SAGE Publications Ltd.

Njördur Viborg, Margit Wångby-Lundh, Lars-Gunnar Lundh, Ulf Wallin, and Per Johnsson. 2018. Disordered eating in a swedish community sample of adolescent girls: Subgroups, stability, and associations with body esteem, deliberate self-harm and other difficulties. *Journal of Eating Disorders*, 6:3–11.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

WHO World Health Organization. 2012. Depression: A global crisis. world mental health day, october 10 2012. *World Federation for Mental Health, Occoquan, Va, USA*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.

Michael W Young et al. 1971. Fighting with food. leadership, values and social control in a massim society. *Fighting with food. Leadership, values and social control in a Massim society.*

Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747. IEEE.

# Appendix

## A Hyperparameter configurations for the neural network

### A.1 Hierarchical attention network

- LSTM units (post encoder) = 128
- dense BoW units = 20
- dense lexicon units = 20
- LSTM units (user encoder) = 32
- dropout = 0.3
- $l_2$ = 0.00001
- optimizer = Adam
- learning rate = 0.0001
- early stopping patience = 20
- epochs = 20
- maximum sequence length = 256
- posts per chunk = 50

## B Adjective vectors for each personality dimension

**EXT+:** dominant assertive authoritarian forceful assured confident firm persistent
**EXT-:** nervous modest quiet forceless afraid shy calm indecisive
**AGR+:** tender gentle soft kind affectionate helpful sympathetic friendly
**AGR-:** cruel unfriendly negative mean brutal inconsiderate insensitive cold
**CON+:** organized orderly tidy neat efficient persistent systematic straight careful reliable
**CON-:** distracted unreliable incompetent wild inefficient disloyal chaotic confused messy disorganized
**NEU+:** worried stressed anxious nervous fearful touchy guilty insecure restless emotional
**NEU-:** balanced stable confident fearless calm easy-going relaxed secure comforted peaceful
**OPN+:** philosophical abstract imaginative curious reflective literary questioning individualistic unique open
**OPN-:** narrow-minded concrete ordinary incurious thoughtless ignorant uneducated common conventional restricted