Integrating UMLS for Early Detection of Sings of Anorexia

Flor Miriam Plaza-del-Arco, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, plubeda, mcdiaz, laurena, maite}@ujaen.es

Abstract. Mental disorders are one of the main concerns of today's society. Early detection of symptoms can greatly help people who suffer from these illnesses. Nowadays, social media play an important role in peoples mental health. Therefore, the treatment of this information using NLP technologies can be applied to automatically detect mental problems such as eating disorders. In this paper, we describe our participation at CLEF eRisk 2019. In particular, we have participated in Task 1: Early Detection of Signs of Anorexia. We have developed three systems based on machine learning. Our main contribution is the use of external knowledge in our systems such as UMLS and similarity embeddings. Our results shown that the use of biomedical ontologies improve the accuracy of the systems.

Keywords: Anorexia · SVM · TF-IDF · Similarity Embeddings · UMLS

1 Introduction

Mental disorders are one of the diseases that most concern society today. They embrace a wide range of problems with different symptoms. However, they are usually characterized by a combination of abnormal thoughts, perceptions, emotions, behaviour, and relationships with others. Examples are anxiety, dissociative identity, depression, bipolar, schizophrenia or anorexia nervosa.

According to a study of World Health Organization, 450 million people suffer from a mental or behavioural disorder, one in four families has at least one member affected by a mental disorder and about 1 million people commit suicide each year. Mental disorders often influence other diseases such as cancer or cardiovascular disease. Therefore, people with this type of problem have disproportionately high rates of disability and mortality.

Nowadays, social media play an important role in people's mental health [17, 4]. The language and vocabulary that users use to express themselves in social media may indicate feelings of guilt, helplessness, hatred or contempt for themselves, which are some symptoms of depression [3]. People suffering from eating disorders, such as anorexia and bulimia, can often be identified through the use of certain keywords that characterize and promote these disorders [1, 19].

The burden of mental disorders continues to grow with significant impacts on health and major social, human rights and economic consequences in all countries of the world. Technology can be applied to develop systems for detect mental disorders in social media. These models use features or variables that have been extracted from labeled user-generated data [13]. To collect the data, the most popular platforms are usually Twitter, Facebook or Reddit [6, 15, 19].

The most common features used to build predictive models are those related to the user texts such as: topics, frequencies of each word or multiple words, features based on sentiment analysis to measure the subjectivity of a sentence and features derived from lexicons like LIWC to measure the usage of self references, social words and emotions.

In this paper, we present the different systems we have developed as part of our participation at CLEF eRisk 2019: Early risk prediction on the Internet [12]. It gives three task. Task 1 is about early detection of signs of anorexia, task 2 is about early Detection of Signs of Self-harm and the last one is about measuring the severity of the signs of depression. Particularly, we have participated in Task 1. This task was introduced in 2018 and consists of sequentially processing pieces of evidence and detect early traces of anorexia as soon as possible. The source of data is also the same used for eRisk 2017 [10] and 2018 [11]. It is a collection of writings (post or comments) from a set of social media users. There are two categories of users, anorexia and non-anorexia, and, for each user, the collection contains a sequence of writings (in chronological order).

The rest of the paper is structured as follows. In Section 2 we explain the data used in our methods. Section 3 presents the details of the proposed systems. In Section 4, we discuss the analysis and evaluation results for our systems. We conclude in Section 5 with remarks and future work.

2 Dataset

The dataset used in the eRisk 2019 early detection of signs of anorexia task has the same format as the collection described in Losada [9]. The dataset for this year contains the training and test data used in 2018 [11]. The collection consist of writings obtained from the social media platform Reddit.

This task takes into account the timeline, so it is an early detection of signs of anorexia. For that, we will obtain the writings of users by chunks and we must go sent the answers to obtain the next writings. For example, in the first step, they will give us the first writing of each user, we will send our answers for each user and we will obtain the second set of writings, and so on.

The training phase consists of all writings of all users explicitly indicating which users are diagnosed with anorexia. On the other hand, the test collection for 2019 is composed of 849 users and 2000 chunks of writings, and the messages have dates after January 2011.

We have obtained some statistics from the training corpus before starting to develop the systems. These statistics are shown in Table 1, to obtain the tokens of sentences and words we have used Natural Language Toolkit (NLTK) library in Python.

Table 1. Statistics obtained from the training corpus.

Statistics	***Total***
Number total of user	472
Number of user with anorexia	61
Number of user non-anorexia	411
Mean writings per user	536.73
Number of writing with tittle	61.5
Number of writing with text	210.34
Mean tokens in title	2.91
Mean tokens in text	32.97
Mean sentences in title	0.28
Mean sentences in text	2.35
Vocabulary size	168,400

3 Methodology

In this section we will expose the systems created for this task. All our systems are based on machine learning approaches, specifically Support Vector Machine (SVM).

The architecture of the experiments carried out is shown in Figure 1. We can see that we make use of external resources such as the Spacy library[1] and UMLS explained in Section 3.3 and Section 3.4, respectively.

3.1 Pre-processing

In order to carry out the three experiments in the same way, we first carry out a pre-processing of the text using Natural Language Processing (NLP) tools and techniques. Pre-processing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process [18].

[1] https://spacy.io/

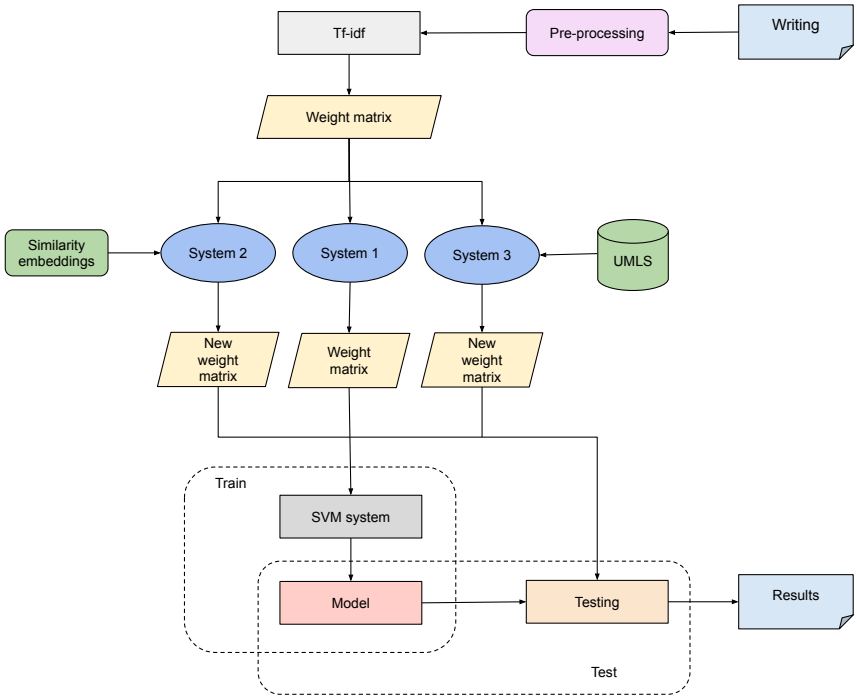

Fig. 1. Architecture followed in the systems.

For all our systems, we took into account the title and the text and we created a new document joining the title and the text. Pre-processing for this new document was as follows:

1. Change all words to lowercase.
2. Remove empty multi-lines from text.
3. Remove URLs from text.
4. Treat only words that contain alphanumeric characters.

3.2 Baseline system

In the first system each sentence is represented as a vector of uni-grams choosing the Term frequency - Inverse document frequency (TF-IDF) scheme and it is used as feature for the classification using the SVM algorithm.

SVM are supervised learning models with associated learning algorithms that analyze data used for binary classification analysis. Many researchers have reported that this classifier is perhaps the most accurate method for text classification [14] and specifically in signs of anorexia there are several studies [5]. In our case, we try to predict whether a text suggests signs of anorexia or not.

TF-IDF is a numerical statistic which shows that a word is how important to a document in a collection. This statistics is often used as a weighting factor

in text mining. The value of TF-IDF increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the dataset.

The parameters used for TF-IDF are shown below:

- min_df = 3
- max_df = 0.
- sublinear_tf = True
- stop_words = english stopwords
- use_idf = True
- tokenize = we use Spacy tokenizer with *en_core_web_md* module
- lowercase = True
- ngram_range = (1, 1)

The next approaches described in the Section 3.3 and Section 3.4 are based on the Baseline (SVM + TF-IDF) adding more relevant information to each document.

3.3 Similarity embeddings

For the second system, we employ word embeddings for measuring similarity. Semantic similarity is a measure of conceptual distance between two objects, based on the correspondence of their meanings [8]. Distributional word vector models capture some aspect of word co-occurrence statistics of the words in a language [7]. Therefore, word embeddings which are trained on word co-occurrence counts can be used to capture semantic word similarity.

The idea in this system was to modify the values of TF-IDF matrix for each document taking into account the similarity of the word *anorexia* with the rest of the words in the corpus. This idea arises because in the corpus vocabulary we have observed many words related to anorexia, such as *vomiting*, *appetites*, *mismanagement*, *nutrition*, *illness*, *thinness*, *calories*, *bulimia*, among others.

Table 2. Examples of concepts related to *anorexia* in UMLS.

Word₁	***Word₂***	***Spacy similarity score[0,1]***
bulimia	anorexia	0.99
disorder	anorexia	0.99
illness	anorexia	0.69
undernutrition	anorexia	0.57
game	anorexia	0.084
computer	anorexia	0.13
work	anorexia	0.19

To calculate the semantic similarity between two words, we employ word vectors from the Spacy library available for Python language. Specifically, we use

the available pre-trained statistical models for English "en_core_web_md" wich version is 1.2.0. It is composed of 685k keys, 20k unique vectors (300 dimensions) and it was trained on OntoNotes, with GloVe vectors trained on Common Crawl. To modify the TF-IDF matrix, in this system we apply the following steps:

1. Load the spacy model "en_core_web_md".
2. Load the task dataset.
3. Pre-process the dataset following the pre-processing explained in the Section 3.1.
4. Get the similarity of each word in the document with the word anorexia using the spacy model.
5. Modify the TF-IDF matrix for each document by multiplying the TF-IDF value of a word by its similarity to the word anorexia.
6. Finally, we use as classifier the SVM.

3.4 Related concepts in UMLS

For the third experiment, we use external knowledge source related to the medical domain to add new features to each word of the message.

In this case, we will use Unified Medical Language System (UMLS) [2]. UMLS is formed by three components: Metathesaurus, specialist lexicon and semantic network [13].

Metathesaurus consists of terms and codes from many vocabularies, including ICD-10-CM, LOINC, MeSH or SNOMED CT. The lexicon is large syntactic lexicon of biomedical and general English and tools for normalizing strings, generating lexical variants, and creating indexes. Last, the purpose of the semantic network is to provide a consistent categorization of all concepts represented in the Metathesaurus and to provide a set of useful relationships between these concepts.

With UMLS we can obtain the concepts related to the concept *anorexia*. In UMLS the concept *anorexia* has the identifier C0003123, in this way, we just have to extract all the relationships with that identifier.

In English we get 285 relationships for the concept *anorexia*, each of these concepts also has synonyms that we will also take into account. Some examples are shown in Table 3 in it we can see the concept identifier, the term and its synonyms.

These words that we find in the concepts and their synonyms are taken into account to modify the TF-IDF matrix. The words of the concepts are preprocessed in the following way:

1. Obtain tokens using the TweetTokenizer of NLTK library.
2. Change tokens to lowercase.
3. Remove tokens that are digits.
4. Remove tokens that are stopwords.
5. Remove tokens that are punctuation marks.
6. Remove tokens with length equal to 1

Table 3. Examples of concepts related to *anorexia* in UMLS.

Code	***Concept***	***Synonyms***
C0162429	Malnutrition	Nutritional deficiency (disorder), malnourished, nutritional deficiency state, undernutrition (disorder), etc.
C1971624	Loss of appetite	Appetite lack of, appetite impaired, loss of appetite (finding), loss of appetite, appetite lost, anorexia, no appetite, appetite absent, etc.
C2267227	Bulimia Nervosa	Bulimia nervosa, bulimia, bulimia nervosa (disorder), bulimia nervosa (diagnosis), eating disorder bulimia nervosa, etc.
C0689452	Megestrol Acetate	Megestrol acetate 20mg tablet, megestrol Acetate 20 mg oral tablet, etc.

This process will help to obtain only the relevant words from the biomedical concepts giving them a greater weight in the matrix. A total of 525 tokens were obtained after pre-processing and stored in a dictionary for later reference. Some saved example tokens are: *food, bulimia, disease, anemia, abdomen, weight, appetite, anorexic, loss, appetites, mismanagement, nutrition, illness, toxic, metabolism*, etc.

As we can see in the example of our dictionary, there are words that are more related to anorexia, so we will try to give more attention.

In the TF-IDF matrix all the weights of the words that are included in our dictionary of relevant words by UMLS will be modified. Finally, we will obtain a new matrix where the tokens included in our dictionary will have a value equal to 1.

4 Results analysis

This section discusses the results we have obtained by our different systems. During the pre-evaluation phase we carried out several experiments with the training set using the 10-fold cross validation to evaluate our approaches. During the evaluation phase, we used the training set to train our systems and the test set to evaluated them.

The official competition metric included in the experimental report are the standard measures such as Precision (P), Recall (R) and the F-measure (F) together with ERDE and latency. ERDE is the Early Risk Detection Error measure proposed in [9]. Latency is an alternative evaluation metric for early risk prediction is done by Sadeque and colleagues [16]. The latest measure taken into account is speed. Speed is computed as follows:

$$speed = (1 - median\{penalty(ku) : u \in U, du = gu = 1\}) \quad (1)$$

The results we have obtained by the three systems we carried out are shown in Table 4. The 1 run refers to our baseline system described in Section 3.2, the 2 run is related to our similarity embeddings systems described in Section 3.3 and the 3 run is associated to our related concepts in UMLS described in Section 3.4.

Table 4. Decision-based evaluation.

Run	P	R	F1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	Speed	Latency-weighted F1
1	0.12	0.97	0.21	0.11	0.07	5	0.98	0.21
2	0.11	0.99	0.20	0.11	0.07	5	0.98	0.20
3	0.18	0.95	0.30	0.09	0.05	8	0.97	0.30
Mean all participants	0.38	0.54	0,39	0.09	0.06	54.15	0.95	0.39

The results obtained by our team are not as expected. However, in Table 4 it should be noted that related to our systems, the third run has achieved the best results obtained a 30% of F1-score outperforming our baseline system (21% F1-score). We can also notice that the Recall measured obtained in all of our runs is remarkably high in compared of the average achieved by the participants. Nonetheless, the precision of our systems is very low so it penalizes the F1 score.

As regards to the system corresponding to the run 2, it has not outperform the results obtained by the baseline system. Perhaps, this is because the vocabulary used in the embeddings is not appropriate for this task and can introduce noise when obtaining the similarity between two words. For this reason, the 3 run could be obtained better results with a specialized vocabulary related to anorexia vocabulary. In this experiment, we can see that adding new sources of external biomedical domain knowledge is a good option as we get better results. This is because the terminology used this run is enriched with different ontologies. These ontologies are made up of medical words providing extra information to the message. In this way, we have obtained greater precision and improve the final result.

5 Conclusion and future work

In this paper, we presented our first participation at CLEF eRisk 2019: Early risk prediction on the Internet. Specifically, we have participated in Task 1 called Early Detection of Signs of Anorexia.

All of our systems are based on machine learning approaches (SVM) taken into account the TF-IDF weight matrix. The main hypothesis considered in the experiments 2 and 3 was to modify the TF-IDF matrix with extra knowledge obtained by similarity embeddings from a model of spacy and UMLS.

In the evaluation phase, we realized that our systems were not computationally fast. For this reason, we could only run 317 chunks of 2000.

As regards to our results, we have not managed to surpass the average of the results obtained by the other participants. However, we have succeeded in overcoming our baseline system with a 19% of F1 in the case of the third system.

A problem that we have found is that the training dataset contains many messages from the same user diagnosed with anorexia, but not all messages written by that user refer to this disease. Therefore, to improve the systems, we consider it is very important that the dataset contain information about the moment in which the user refers to anorexia.

In order to perform a complete analysis of our systems, we will wait for the task organizers to release the complete test dataset with its corresponding labels.

As future work, we plan to improve the speed of our systems in order to evaluate all the possible chunks. Also, we will explore other systems based on deep learning and we will continue studying some resources for the purpose of improve our results incorporating external knowledge.

Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

References

1. Arseniev-Koehler, A., Lee, H., McCormick, T., Moreno, M.A.: # proana: pro-eating disorder socialization on twitter. Journal of Adolescent Health **58**(6), 659–664 (2016)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media (2013)
4. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences **18**, 43–49 (2017)
5. Guo, Y., Wei, Z., Keating, B.J., Hakonarson, H.: Machine learning derived risk prediction of anorexia nervosa. BMC medical genomics **9**(1), 4 (2015)
6. Hwang, J.D., Hollingshead, K.: Crazy mad nutters: the language of mental health. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. pp. 52–62 (2016)
7. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics **3**, 211–225 (2015)
8. Lin, D., et al.: An information-theoretic definition of similarity. In: Icml. vol. 98, pp. 296–304. Citeseer (1998)
9. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 28–39. Springer (2016)

10. Losada, D.E., Crestani, F., Parapar, J.: Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In: CLEF (Working Notes) (2017)
11. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk: Early risk prediction on the internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 343–361. Springer (2018)
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)
13. McCray, A.T.: The umls semantic network. In: Proceedings. Symposium on Computer Applications in Medical Care. pp. 503–507. American Medical Informatics Association (1989)
14. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between svm and ann. Expert Systems with Applications **40**(2), 621–633 (2013)
15. Prieto, V.M., Matos, S., Alvarez, M., Cacheda, F., Oliveira, J.L.: Twitter: a good place to detect health conditions. PloS one **9**(1), e86191 (2014)
16. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 495–503. ACM (2018)
17. Seabrook, E.M., Kern, M.L., Rickard, N.S.: Social networking sites, depression, and anxiety: a systematic review. JMIR mental health **3**(4), e50 (2016)
18. Vijayarani, S., Ilamathi, M.J., Nithya, M.: Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks **5**(1), 7–16 (2015)
19. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 91–100. ACM (2017)