Early detection of anorexia using RNN-LSTM and SVM classifiers

Akshaya Ranganathan, Haritha A, Thenmozhi D, Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai {akshaya16009, haritha16038}@cse.ssn.edu.in {theni_d, aravindanc}@ssn.edu.in

Abstract. Social Media text analysis has engendered a variety of applications in the medical domain, a major example being the detection and cure of deleterious mental disorders. Anorexia is a deadly, psychiatric eating disorder with typical characteristics of alarmingly low body weight conditions and distorted body image, with an unreasonable sense of being overweight. With developments in the field of Natural Language Processing, such highly lethal disorders can be identified and mitigated in their rudimentary stages, saving the victim a lot of mental and physical abuse. The Task 1 of CLEF 2019's eRisk lab focuses mainly on the early prediction of anorexia, analysed by posts which are sourced from social media platforms. Our team, SSN-NLP has used variations of two major models for sentiment classification, a deep learning RNN-LSTM, and a traditional SGDC Classifier. User-specific data from consequent posts that were extracted from Reddit was released by CLEF eRisk, which was used in its entirety for our training, testing, evaluation and scoring process. With the help of RAKE (Automated keyword extraction), numeric scores were obtained to identify the level of anorexia/self-harm.SSN-NLP submitted 5 variant models to the server that repeatedly accepted submissions and gave user writings to the participating teams. According to the ERDE-50 and F1 scores, our 2-layer LSTM with normed-bahdanau attention, performed the best having scores of 0.07 and 0.33 respectively.

Keywords: Anorexia · early detection · natural language processing · deep learning · machine Learning · LSTM · SVM

1 Introduction

Anorexia Nervosa is a potentially life-threatening psychiatric disorder characterized by very extreme unhappiness over one's body image and intense desire to lose weight even if it's lower than what's considered normal. In the age of Instagram celebrities showing off their perfectly toned bodies, internet culture has created harsh rules that people, especially teenagers are expected to adhere

to. According to a study by National Eating Disorders Association[1], irrespective of the time, 0.3-0.4% of women and 0.1% men test positive for anorexia nervosa. DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) gives definitions and diagnostic material for mental disorders. According to DSM-5, Anorexia Nervosa is characterized by the following criteria: [2]

1. Restriction of energy intake relative to requirements leading to significantly low body weight in the context of age, sex, developmental trajectory, and physical health.
2. Intense fear of gaining weight or becoming fat, even though underweight
3. Disturbance in the way in which oneś body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight.

However, another serious type of anorexia is called Atypical Anorexia where a person maintains a healthy weight despite consistent loss in weight.Types of anorexia include:

1. **Binge/purge type:** A person tries to purge by over-exercising or even vomiting after eating in an attempt to compensate for the weight gained by eating.
2. **Restrictive type:**: A person levies harsh restrictions on the quantity of food consumed, which in most cases is barely sufficient for survival.

eRisk 2019 primarily focuses on early detection of risk on the internet. The primary goal is to use text mining solutions for early detection in various areas like detection of people with suicidal tendencies, tendency to fall prey to criminal organizations, etc [4]. The aim of Task 1 under Erisk 2019 is to detect symptoms of anorexia as early as possible. Early detection technologies using text processing can be employed in different areas, particularly those related to health and safety. A few applications of early detection include the areas of sexual predators, mental disorders and cyber-bullying. Prediction is broadly classified into two stages: - the training stage and the test stage. In the training stage eRisk released chunks of training data as well as test data of eRisk 2018. The chunks consisted of user writings posted on Reddit as well as classification results. Users are classified as Anorexic and Non-Anorexic. During the testing stage, an automatic server repeatedly accepted our submissions and released test data batch by batch. The task evaluates the earliness of predictions in addition to their correctness. The task aims to obtain a scoring system based on the level of alert.

2 Related Works

Extant methods to detect Anorexia can be categorized into two types. One method is the analysis of change in behavioral patterns by general physicians

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

as well as friends and family of the patient through structured mental analysis. According to a study that weighs the importance of a primary physician in detecting eating disorders, a series of questions are used to detect the presence of anorexia which is done by examining the answers to each of these questions [16]. A few examples include What did you eat yesterday?, Do you ever binge eat (eat more than you want) or use laxatives, diuretics, or diet pills?, Do you think you are thin (too thin), etc. The second method [9] involves the use of Sentiment Analysis on Social Media posts. For example, a research work showcased that students with signs of depression use more personal pronouns like 'I' and negative valence possessing words (eg: gloomy, sad). Erisk aims at early detection of anorexic tendencies by analyzing posts of users on Reddit. One such approach involves the Bag of Words (BoW) model [13] that uses a vocabulary comprising of all the unique words in the text and performs vectorization assigning a specific weight to each word. The term weighting for the BoW model has been split into 3 components: a term frequency component, a document frequency component, and a normalization component. Yet another approach involves UMLS based MetaMap [9] assistance for keyword detection. Further, Traditional Learning algorithms were applied to the information collected by the methods mentioned above (eg. SVM, logistic regression, RF). Yet another approach involved the use of TF-IDF similar to the works mentioned before. However, this research adopted a deep learning approach using CNN-LSTM [3]. Our work involves the usage of Recurrent Neural Networks with Long short term memory (LSTM) to analyze patterns and make predictions on sequences of texts. Rapid automated keyword extraction (RAKE) was implemented to identify the most frequently occurring keywords relating to anorexia in the training data. The results were combined to devise a prediction and risk-based scoring system.

3 Dataset Analysis

3.1 Dataset analysis - Task 1

This year's Task 1 was an extension of CLEF eRisk 2018's Task 2, the training data [14] of this years task was a combination of both the test and training data of the previous year. *Reddit*, much like twitter offers a python supporting API that can be used to scrape required data effectively. Twitter sentiment analysis [2] has proven to be a powerful indicator of mental illnesses like depression and PTSD. While the training data was categorized into negative and positive examples, the labels of test data had to be extracted from the file **risk-golden-truth-test.txt** and mapped on to the actual writings of the users. Each document had an XML tree structure comprising of the tags : *INDIVIDUAL ,ID ,WRITING ,TITLE ,DATE* and *TEXT*. For the training-examples, a total of 152 user writings were given in comparison to 320 users for the test-examples all out of which only the TEXT and TITLE attributes were separated to be fed as training data. Table 1 gives a summary of all heading levels.

Table 1. Training and Test 2018 data set

Attributes	2018 Train	2018 Test
Number of users	152	320
Positives/Negatives	20/132	41/279
Number of documents	84,834	1,68,507
Avg documents per user	558.26	526.89
Avg words per document	184.54	197.28

3.2 Data extraction and cleaning

The data [14] given was a consolidation of the test and training data of CLEF 2018. Data were represented as positive-examples and negative-examples chunks, each containing XML files of writings done by a certain subject. Using XML ElementTree library of Python, the given TEXT elements of each file were consolidated as follows: (see Fig. 1) To flatten out the discrepancies in the data set,

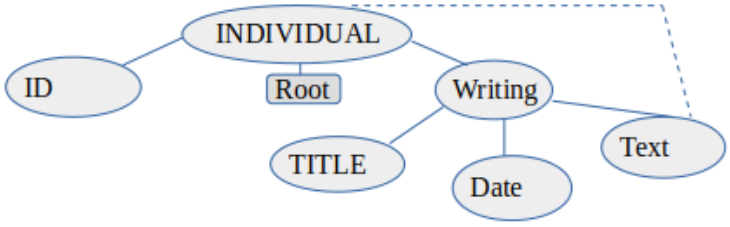

Fig. 1. XML DOM Tree structure of the released data

all special characters, erroneous blank spaces and empty strings (NULL) were removed using Regular Expressions. The cleanup of data was done in accordance with the input expected by the Neural Machine Translation model. Cleaned text and respective labels were stored in the form of comma-separated values using FileWriter of python. A vocabulary file comprising of all unique words in the training set was built to be fed into the Deep learning model.

3.3 Data augmentation

Due to the sparse characteristics of positive examples in the training set, Data Augmentation had to be done using the mentioned mechanism: **Synonym generation using POS Tagging:** Using the *POSTagger* module [3], various parts

[3] https://github.com/nltk/nltk

of speech were identified from each positive example of the text. Post identification, the *NLTK WordNet* [4] module identified the synonyms for adjectives(JJ) and adverbs(RB) and populated the dataset with replaced text which led to a significant increase of tuples in our dataset. As shown in the figure, (Fig. 3) the POS Tagger splits each sentence into relevant parts of speech, and the wordnet (Fig. 2) generates synonyms for each word. Multiple sentences of anorexia positive users were augmented to the dataset by replacing each adverb and adjective in a sentence with their respective list of most relevant synonyms. Take an example sentence: **My body is so heavy that I actively need to exercise every moment of the day**. The POS Tagger identifies **heavy** and **actively** as adjective and adverb respectively. Synset identifies synonyms for heavy as weighty, hefty, big, massive and synonyms for actively as effectively, usefully, productively. Now, sentences with combinations of these synonyms are generated. Nearly 45,000 sentences were added to our dataset through the mentioned methodology.

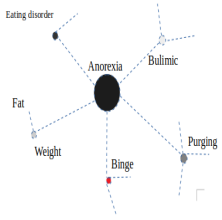

Fig. 2. Word net relating to anorexia

```
[('Each', 'DT'),
 ('of', 'IN'),
 ('us', 'PRP'),
 ('is', 'VBZ'),
 ('full', 'JJ'),
 ('of', 'IN'),
 ('stuff', 'NN'),
 ('in', 'IN'),
 ('our', 'PRP$'),
 ('own', 'JJ'),
 ('special', 'JJ'),
 ('way', 'NN')]
```

Fig. 3. POS-tagging

4 Proposed methodologies and Implementation

4.1 Deep learning approach -Neural Machine Translation

Task 1's primary goal was to classify the user as anorexia-positive or anorexia-negative. We have used a Deep Learning based approach for our implementation using Neural Machine Translation to solve the classification problem. Basic Architecture of Neural Machine Translation is a Sequence to Sequence model (Seq2Seq). NMT is built based on the concept of an Encoder- Decoder [15]. The encoder converts the input sequence to a thought vector while the decoder maps it to a target language. In our case, the decoder maps the input sequences to two classes- **positive and negative** indication of anorexia. The TensorFlow code based on tutorial code released by Neural Machine Translation[5] [7] that was developed based on Seq2Seq models [12, 1, 6] was used to implement our

[4] https://github.com/wordnet/wordnet
[5] https://github.com/tensorflow/nmt

deep learning approach for sentiment classification. Neural Machine Translation (NMT) was implemented with LSTM. LSTM is expanded as Long Short Term memory which is used to remember only the important parts of each input sentence and is trained to forget the rest. Thus, the output is a combination of the current input sentences predictions as well as the memory of previous important parts of sentences. LSTM captures Long Term Dependencies using 3 gates

- **Forget Gate:** Decides what part of previous cell state must be forgotten.
- **Input Gate:** Responsible for the addition of information to the cell state.
- **Output Gate:** Responsible for selecting useful information to output at current cell state

$$i_t = \sigma(w_x^{(i)} x + w_h^{(i)} h_{t-1} + b^{(i)}) \tag{1}$$

$$f_t = \sigma(w_x^{(f)} x + w_h^{(f)} h_{t-1} + b^{(f)} + 1) \tag{2}$$

$$o_t = \sigma(w_x^{(o)} x + w_h^{(o)} h_{t-1} + b^{(o)}) \tag{3}$$

$$\widetilde{c}_t = tanh(w_x^{(c)} x + w_h^{(c)} h_{t-1} + b^{(c)}) \tag{4}$$

$$c_t = f_t \circ \widetilde{c}_{t-1} + i_t \circ \widetilde{c}_t \tag{5}$$

$$h_{b/f} = o_t \circ tanh(c_t) \tag{6}$$

where w_s are the weight matrices, h_{t-1} is the hidden layer state at time $t-1$, i_t, f_t, o_t are the input, forget, output gates respectively at time t, and $h_{b/f}$ is the hidden state of backward, forward LSTM cells. Four different NMT variations have been implemented for runs 1-4 of our submissions.

- **Model 1:** 2 layer bidirectional LSTM with Scaled Luong attention
- **Model 2:** 4 layer bidirectional LSTM with Scaled Luong attention
- **Model 3:** 2 layer bidirectional LSTM with Normed Bahdanau attention
- **Model 4:** 4 layer bidirectional LSTM with Normed Bahdanau attention

4.2 Traditional Learning Approach

TF-IDF is used to assign weights to words to find out important words. TF stands for term frequency. It is a measure of the number of times a word occurs in a given document [10]. It is calculated by dividing the number of occurrences of a given word by the total number of words in a document. However, words like **a**, **the** occur a lot of times and are not very significant. So, we calculate the Inverse Document Frequency.

$$Weights = TF * IDF \tag{7}$$

Stochastic Gradient Descent[1] is essentially Gradient Descent with a batch size of 1 and works effectively when redundant data is present. SGD Classifier of sklearn performs Stochastic Gradient Descent Optimization on SVM Classification Model. Stochastic Gradient Descent is proven to be useful especially for large datasets and has found increased usage in several text mining applications [10]. After data augmentation, the dataset was cleaned and fed to the model. The accuracy of the model while training was found to be 90%.

- **Model 0:** SVM Classifier with SGD optimization using TF-IDF

4.3 Automated Keyword Extraction

The motive behind Task 1 of eRisk 2019 was to facilitate the early prediction of anorexia. This year, they added another feature to the submissions called a score of positivity or negativity. Score is a numeric estimation of the level of anorexia/self-harm. Using this score, CLEF 2019 adapts ranking based measures for the evaluation of participants. The module Rapid Automated Keyword Extraction (RAKE) [11] was used to identify the most frequently occurring keywords in our training set, and to calculate the score based on these keywords. The input parameters for RAKE comprise a list of stop words (or stoplist) usually provided by NLTK for the English language, a set of phrase delimiters, and a set of word delimiters. RAKE uses stop words and phrase delimiters to segment the chunk of text into candidate keywords. The number of times each word occurs in the document gives the frequency score, and the number of times each keyword occurs with each other keyword is found as the co-occurrence score.

$$final_{score} = co - occurrence_{score}/frequency_{score} \tag{8}$$

RAKE eliminates words that occur very frequently in the document but are of trivial relevance. Using co-occurring keywords, we successfully mined out pairs like body-mass, anorexia-nervosa,purge-eating, binge-eating. The fundamental difference between RAKE and TF-IDF scores is that RAKE finds word phrases in a single document and assigns relevance scores, while TF-IDF uses multiple documents to assign a single word score. Since our work required a single but voluminous training document, RAKE outperformed its TF-IDF counterpart. To achieve stable prediction scores, we used a function that checks the following :

- If a user classified as anorexia positive has stopped posting altogether, the score was significantly increased, causing a high level of alert.
- If a user was classified positive both in the current and previous runs, the score was boosted so as to confirm the decision of positive anorexia, as early as possible.
- If a user was classified positive in the previous run, but the current run is negative, the score was balanced out, waiting for further writings to make the ultimate decision.

5 Results and Evaluation

5.1 Decision based evaluation

According to the task, several methods of evaluation were considered [5]. Evaluation of results was initially based on Early Risk Detection Error (ERDE). ERDE gives a measure of correctness of decision as well as delay taken to arrive at a decision.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{11}$$

However, ERDE has certain drawbacks. For example, a system that detects all the true positive writings still does not get an error of zero. Alternatively, a modification $ERDE_o$%was suggested. This method considers the percentage of writings of the users seen before making a decision as opposed to the number of user writings. However, this method has a major flaw as in real life the total number of user writings may not be known.Another method based on $F_{latency}$ was proposed. For a user $u \in U$, k_u writings are seen before making a decision d_u. g_u stands for the ground truth of decisions. Delay in finding true positives are considered as

$$latency_{TP} = median\left\{k_u : u \epsilon U, d_u = g_u = 1\right\} \tag{12}$$

Standard measures of precision, recall, F-measure are calculated as follows:

$$P = \frac{|u \epsilon U : d_u = g_u = 1|}{|u \epsilon U : d_u = 1|} \tag{13}$$

$$R = \frac{|u \epsilon U : d_u = g_u = 1|}{|u \epsilon U : g_u = 1|} \tag{14}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{15}$$

A penalty factor is introduced. A penalty is assigned to every true positive decision taken after k_u writings.

$$penalty(k_u) = -1 + \frac{2}{1 + exp^{-p \cdot (k_u - 1)}} \tag{16}$$

Yet another factor for evaluating performance is the speed of a system. A speed of 1 indicates that the system predicted true positives in the first writing as opposed to 0 if the system predicts only after a few hundred writings.

$$speed = (1 - median\left\{penalty(k_u : u \epsilon U, d_u = g_u = 1)\right\}) \tag{17}$$

Based on the speed and F1 score, latency weighted F1 score is calculated.

$$F_{latency} = F * speed \tag{18}$$

The maximum precision attained by our system is **0.48**, whereas the overall [5] maximum among all systems is 0.71. Maximum recall of our system is 0.26, as opposed to overall maximum of all systems which is 1. Maximum F1 score is 0.34, whereas maximum of all systems id 0.71. ERDA5 is relatively low with a value of 0.08 as opposed to least value of 0.06 amongst all systems. Least ERDA50 is 0.07 for our system, while overall least is 0.03. Speed of a system is 1 if it detects true positive in the first writing of a user. Systems speed is 1.

Table 2. Decision based evaluation

team	run	P	R	F1	ERDE5	ERDE50	latencyTP	speed	latency-weighted F1
SSN-NLP	0	.32	.16	.22	.08	.08	2	1	.22
SSN-NLP	1	.30	.22	.25	.08	.07	**1**	**1**	.25
SSN-NLP	2	.47	.22	.30	.08	.07	2	1	.30
SSN-NLP	3	.48	.26	.34	.08	.07	2	1	.33
SSN-NLP	4	.32	.15	.21	.08	.08	**1**	**1**	.21

5.2 Ranking based evaluation

Along with the decision, a score which is an estimate of the level of risk, was also calculated for each user. The evaluation algorithm assigns ranks to users based on decreasing level of risk. The ranks are re-calculated after each set of writings. The rankings are evaluated with P@10 and NDCG metrics. The relatively long duration between submissions of various runs can be attributed to the offline processes used by our system(*6 days,22 hs*)

Table 3. Ranking based evaluation

Name	run	1 writing		
		P@10	NDCG@10	NDCG@100
SSN-NLP	0	.6	.64	.29
SSN-NLP	1	.3	.28	.15
SSN-NLP	2	.5	.48	.29
SSN-NLP	3	.6	.64	.30
SSN-NLP	4	.3	.33	.15

From the released evaluation results, it can be inferred that our models performed extremely well with respect to early prediction (*speed*), as the true positives were correctly classified within the first few sets of user writings. Our $F_{latency}$ however, was not up to standards, in comparison with a few of the best functioning systems, such as **CLAC**, which achieved a weighted F1 score of **0.69**. Model 1 : **2 Layer BLSTM with scaled luong attention** and Model 4: **4 Layer BLSTM with normed bahdanau attention** have shown the best performance and this could be explained by taking the concept behind these attention mechanisms. As mentioned in [6], the luong mechanism simply uses hidden states at the top LSTM layers in both the encoder and decoder, thus explaining why for a lesser number of layers (2 layers) scaled - luong attention worked better. The reason why bahdanau attention worked for a deeper number of layers (4 layers) can be justified, as a hidden state in Bahdanau goes through a deep-output and a max-out layer before making predictions [1].

6 Conclusions and Future work

In this paper, we have presented the participation of our team, SSN-NLP at the eRisk 2019 task of early detection of signs of anorexia. Early risk prediction on the internet is vital to the development in the field of mental health and safety. We have treated this as a classification problem and presented 4 variations of Deep learning approaches, and one Traditional learning model using Neural Machine Translation (NMT) and SVM with SGD optimizer. The future scope for our model includes complete automation, devoid of any kind of online processing and research on other algorithms that could improve our model accuracy.

References

1. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.
2. Coppersmith, Glen, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. "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, pp. 31-39. 2015.
3. Liu, Ning, Zheng Zhou, Xin Kang, and Fuji Ren. "TUA1 at eRisk 2018." (2018)
4. Losada, David E., Fabio Crestani, and Javier Parapar. "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, Cham, 2018.
5. Losada, David E. and Crestani, Fabio and Parapar, Javier.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.(2019)
6. Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
7. Luong, T., Brevdo, E. and Zhao, R., Neural machine translation (seq2seq) tutorial. 2017. URL: https://www. tensorflow. org/tutorials/seq2seq (17.02. 2018).
8. Robbins H, Monro S. A stochastic approximation method. The annals of mathematical statistics. 1951 Sep 1:400-7.
9. Paul, Sayanta, Jandhyala Sree Kalyani, and Tanmay Basu. "Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks." (2018)
10. Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.
11. Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic keyword extraction from individual documents." Text mining: applications and theory (2010): 1-20.
12. Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).
13. Trotzek, M., Koitka, S. and Friedrich, C.M., Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia.(2018)

14. Trotzek, Marcel, Sven Koitka, and Christoph M. Friedrich. "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences." IEEE Transactions on Knowledge and Data Engineering (2018).
15. Verma, A. A., and Bhattacharyya, P. Literature Survey: Neural Machine Translation.
16. Walsh, J M et al. Detection, evaluation, and treatment of eating disorders the role of the primary care physician. Journal of general internal medicine vol. 15,8 (2000): 577-90.