UArizona at the CLEF eRisk 2017 Pilot Task: Linear and Recurrent Models for Early Depression Detection

Farig Sadeque, Dongfang Xu, and Steven Bethard

School of Information, University of Arizona
1103 E 2nd St, Tucson, AZ 85721
{farig, dongfangxu9, bethard}@email.arizona.edu

Abstract. The 2017 CLEF eRisk pilot task focuses on automatically detecting depression as early as possible from a users' posts to Reddit. In this paper we present the techniques employed for the University of Arizona team's participation in this early risk detection shared task. We leveraged external information beyond the small training set, including a preexisting depression lexicon and concepts from the Unified Medical Language System as features. For prediction, we used both sequential (recurrent neural network) and non-sequential (support vector machine) models. Our models perform decently on the test data, and the recurrent neural models perform better than the non-sequential support vector machines while using the same feature sets.

1 Introduction

Depression is responsible for almost 12% of all Years Lived with Disabilities (YLDs), with nearly 350 million people suffering from it worldwide [20]. As of 2000, depression also comes with an annual economic burden of 83 billion US dollars [8]; and a 1990 study by Goodwin and Jamison [7] suggested that depression is also the leading cause of suicide, as 15-20% of all major depressive disorder patients take their lives. Early detection of depression can help mitigate these threats, but most studies on early detection rely on diagnoses from patients' self reported surveys and experiences[9], which are extremely costly in terms of both time and money, and a major portion of countries with primary health care service do not have the support for these diagnoses. Fortunately, social media may provide us with a solution to this problem, as many studies have successfully leveraged the contents of social media to analyze and predict users' mental well-being [6, 13, 15, 16, 19]. Unfortunately, none of these studies focuses on the importance of the temporal aspect of these detection tasks; hence the introduction of the pilot task on early risk detection of depression in the CLEF eRisk 2017 workshop on Early risk prediction on the Internet: experimental foundations [11].

In this paper we present the five early risk detection models we submitted for the pilot task. We tried to leverage external knowledge sources beyond the provided training data. We incorporated the depression lexicon created by [6] and used Metamap [1] to obtain Unified Medical Language System (UMLS)-identified concept unique identifiers related to mental and behavioral dysfunction from user texts. We used these features with both sequential and non-sequential learning models: we used support vector machines as the non-sequential linear model and recurrent neural networks with multiple layers of gated recurrent units as the sequential models. Our results demonstrate the superiority of sequential over non-sequential models on this task.

2 Task Description

The pilot task on early risk detection of depression focused on sequential processing of contents posted by users in Reddit[1] [10]. The data was a collection of writings posted by users, divided into two cohorts: depressed and non-depressed. The text collection for each user is sorted in a chronological order and then divided into 10 chunks, where the first 10% of a user's writings is in chunk 1, the second 10% is in chunk 2 and so on. Details of this data are in section 3.

The task was divided into two stages: training stage and testing stage. The training stage started on November 30, 2016, when the entire text collection of 486 users was released, with their user-level annotations of depression. Participants then had a little over two months to develop their systems. The testing stage started on February 6, 2017 when the first 10% of texts written by 401 previously unobserved users were released. For the next 9 weeks, new chunks were released, with each chunk including the next 10% of each user's text. After each release, and before the next release, systems had to make a three-way decision for each user: tag the user as depressed, tag the user as non-depressed, or wait to see the next chunk of data. If a user was tagged as either depressed or non-depressed, this decision was final and could not be changed for future chunks of data. After the release of the 10th (and last) chunk of the data, the decision was two-way: tag the user as depressed, or tag the user as non-depressed. The prediction model was then evaluated based on its correctness and how early in the series of chunks was able to make its predictions.

3 Data

A number of social media websites were considered as potential data sources for this shared task. Twitter[2] was discarded because it provided little to no context about the user, is highly dynamic and did not allow them to collect more than 3200 tweets per user, which, in 140-character microblogs, represents only a small amount of text. MTV's A Thin Red Line (ATL)[3], a platform designed to empower distressed teens to respond to issues ranging from sexting to cyberbullying, was also considered, but discarded as there were concerns about redistribution and

[1] `http://www.reddit.com`

[2] `http://www.twitter.com`

[3] `http://www.athinline.org`

problems regarding obtaining user history. Eventually, Reddit, a social media and news aggregation website, was selected because of its organization of contents among specific subreddits, and the ease of collecting data using the API provided by Reddit itself.

For each user, the organizers collected the maximum number of submissions they could find and were allowed to download through the API (maximum 2000 posts and comments per user). Users with less than 10 submissions were discarded. Original redditor IDs were replaced with pseudo user IDs for anonymization, and published along with the title, time and text of the posts.

After the data collection, the users were divided into two cohorts: an experimental depressed group and a control (non-depressed) group. For the depressed group, the organizers searched for phrases associated with self-declaration of depression, such as *diagnosed with depression*, and then manually examined the posts to filter down to just those redditors who explicitly said they were diagnosed with depression by a physician. These self declaration posts were omitted from the dataset to avoid making the detection trivial. For the non-depressed group, organizers collected redditors who had participated in depression forums but had no declaration of depression, as well as redditors from other random subreddits. Their final collection contained 531,453 submissions from 892 unique users, of which 486 users were used as training data, and 401 were used as test data. Statistics for that dataset are shown in Table 1.

	Train		Test	
	Depressed	Control	Depressed	Control
# of subjects	83	403	53	349
# of submissions	30,851	264,172	18,706	217,665
Avg. # of submissions/subject	371.7	655.5	359.7	623.7
Avg. # days from first to last submission	572,7	626.6	608.3	623.2
Avg. # of words per submission	27.6	21.3	26.9	22.5

Table 1. Summary of the task data

4 Features

We considered two types of features that could serve as inputs to our classifiers.

4.1 Depression Lexicon

This is a set of unigrams that has high probability of appearing in depression-related posts. The list was collected from [6], where the authors compiled a list of words that are most associated with the stem "depress" in the Yahoo! Answers *Mental Health* forum using pointwise mutual information and log-likelihood ratio and kept the top words based on their TF-IDF in Wikipedia articles. We used the top 110 words that were presented in the paper. For each post, we generated 110 features: the counts of how many times each word occurred in the post.

4.2 Metamap Features

We used Metamap [1], which is a highly configurable tool to discover concepts from the Unified Medical Language System (UMLS) Metathesaurus[4]. In our preliminary experiments we found out that Metamap produced a lot of incorrect concept matches in social media texts (as it was mainly built to run on clinical texts), but with some tuning, it was possible to use this effectively on social media. We restricted Metamap to only one source (SNOMEDCT-US) and to only two semantic types (Mental or Behavioral Dysfunction, Clinical Drugs). We passed each post through the restricted Metamap and collected all the predicted concept unique identifiers (CUIs). We ended up with a set of 404 CUIs. We generated 404 features for each post: the counts of how many times each CUI occurred in the post.

5 Classifiers

We considered both sequential and non-sequential learning models for the prediction task. For the non-sequential model, we used a support vector machine that observes the user's entire post history at once. For the sequential model, we used a recurrent neural network model that observes each of a user's posts, one at a time. We also used an ensemble model to combine both the sequential and non-sequential models.

As per the shared task definition, classifiers were given the user's history in chunks (the first 10% of the user history, then the first 20%, etc.) and after each chunk, the classifiers were asked to make a prediction of "depressed", "not depressed", or "wait". All our classifiers were trained to make two-way predictions, "depressed" vs. "wait", and if a classifier predicted a user as depressed after seeing the first $n\%$ of the history, that prediction was considered final and the remaining $100 - n\%$ of the history was ignored. On the final (10th) chunk, all users who had only ever been predicted as "wait" were classified as "not depressed". Note that our models never made post-by-post decisions; they always observed the entirety of the $n\%$ of the history they were given and then made a single prediction for the entire $n\%$.

5.1 Support Vector Machine

A support vector machine [5] or SVM is a machine learning technique that is used for binary classification tasks. In this technique, input vectors are non-linearly mapped to a high dimensional feature space, where a linear decision surface is constructed to classify the inputs. It is one of the most popular non-sequential machine learning techniques because of its high generalization ability.

[4] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/`

For the support vector machine (SVM) models we used, the feature vectors needed to summarize the entire history of the user. We converted the post-level raw count features to user-level proportion features (e.g., converting the number of times *depression* was used in each post to the proportion of all words in a all of a user's posts that were *depression*).

We used two out-of-the-box implementations of support vector machines:

- Weka's implementation of the sequential minimal optimization algorithm [14] for training support vector machine classifiers [21]. The model was set to output probability estimates and it normalizes all attributes by default. Other parameters were set to their defaults. We used a degree-1 polynomial kernel and a cache size of 250007 as it performed better in preliminary experiments on the training data.
- LibSVM's implementation of support vector machines [3] using C-support vector classification [2]. Apart from tuning the model for probability estimate outputs, we used the default parameter settings. We used the radial basis function kernel for this one as it performed better in preliminary experiments on the training data.

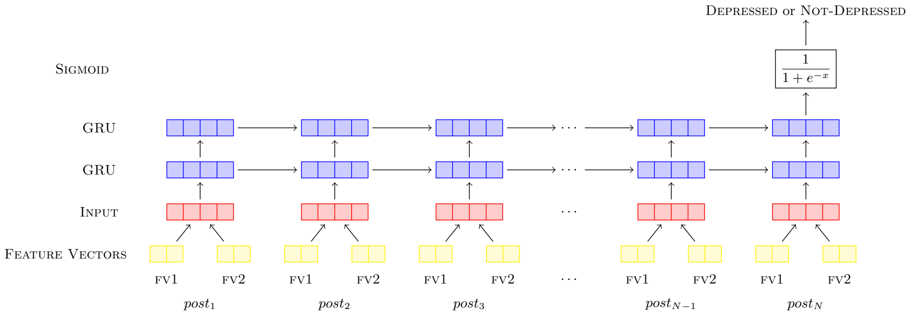

Fig. 1. Architecture of the model for reading the sequence of a user's posts and predicting the user's depression status.

5.2 Recurrent Neural Network

Due to the sequential property of the data, we opted for machine learning techniques that take advantage of this. We used Recurrent Neural Networks (RNN) which have been successful in other natural language modeling problems [12]. Recurrent neural networks are a form of artificial neural network where neurons are connected to form a directed cycle, allowing the network to exhibit temporal behavior, and thus be used as a sequential learning model.

We trained recurrent neural networks that take a sequence of feature vectors, each representing a single post, and predict whether the user is depressed or not.

Figure 1 shows the general architecture of the model. We used Gated Recurrent Units (GRU) [4] to build recurrent layers. Feature vectors representing each post are first concatenated, and then fed as input to the first recurrent layer. A second GRU layer is stacked on top of the first one for more expressive power, and its output is fed through a sigmoid to produce binary output. To make the experiments with different input features comparable, we fixed the size of the GRU units to 32 for all experiments. To avoid overfitting, we used dropout [17] with probability 0.2 on the first input-to-hidden layer. Models were trained with RMSProp optimization [18] on mini-batches of size 200, with all hyperparameters set to default except the learning rate, which was set to 0.001. Each model is trained for at most 800 epochs. The training time for each experiment was around two hours using two Graphics Processing Units (GPUs).

5.3 Ensemble

An ensemble learning model takes the outputs of a set of other machine learning algorithms and uses them as inputs for classification. They are typically used to improve the performance of individual machine learning techniques by leveraging the different strengths of multiple approaches. For this task, We implemented an ensemble learning technique using the probability outputs of the nine individual models (3 from Weka, 3 from LibSVM and 3 from RNN: models used as features either the depression lexicon, Metamap outputs, or both). We used 5-fold cross validation for each model to calculate the probability of each user being depressed and then fed these probabilities to a Naive Bayes classifier, which served as the ensemble classifier. We used Weka's naive Bayes implementation with the default parameter settings.

6 Evaluation and Analysis

We submitted five different models for the task:

- UArizonaA: An SVM model trained using LibSVM with the depression lexicon and Metamap outputs as features.
- UArizonaB: An SVM model trained using Weka with the depression lexicon as features.
- UArizonaC: An RNN model with both the depression lexicon and Metamap outputs as features.
- UArizonaD: The ensemble model.
- UArizonaE: An RNN model with the same structure as UArizonaC, but that always predicts "wait" until 60% of the test data is released.

All of these models were selected for their individual properties. UArizonaA was our most restrictive model, as it vigorously tried to not tag someone depressed, whereas UArizonaC was the most open as it tagged more users as depressed than any other models. The other 3 sat in between these 2. To make UArizonaA a

Model	Brief description	E_5	E_{50}	F_1	P	R
FHDO-BCSGA	Ranked #1 for F_1, #2 for E_5, #2 for E_{50}	12.82	9.69	0.64	0.61	0.67
UArizonaA	LibSVM + lexicon + UMLS	14.62	12.68	0.40	0.31	0.58
UArizonaB	WekaSVM + lexicon	13.07	11.63	0.30	0.33	0.27
UArizonaC	RNN + lexicon + UMLS	17.93	12.74	0.34	0.21	0.92
UArizonaD	Ensemble	14.73	10.23	0.45	0.32	0.79
UArizonaE	RNN + lexicon + UMLS + 60%-wait	14.93	12.01	0.45	0.34	0.63

Table 2. Performance of the models. E_5 and E_{50} are the shared-task-defined Early Risk Detection Error (ERDE) percentages, P is precision, R is recall, and F_1 is the harmonic mean of precision and recall.

little more open towards depression tagging, we combined its 10th chunk output with UArizonaE's 10th chunk output.

The performance of our models are given in table 2, along with the performance of the model that achieved the highest F_1 in the shared task, FHDO-BCSGA. The models were evaluated based on 5 performance measures: precision, recall and F_1, and 2 Early Risk Detection Error (ERDE) [10] variants, with cutoff parameter o set to 5 and 50 posts. ERDE penalizes systems that take many posts to predict depression. For precision, recall, and F_1, a high score is good, while for ERDE, a low score is good.

Our models were competitive with others in the shared task. UArizonaC ranked 1st out of 30 for recall, UArizonaD ranked 3rd for $ERDE_{50}$, and UArizonaB ranked 4th for $ERDE_5$. For precision and F_1, our models were less impressive; both UArizonaD and UArizonaE ranked 11th for F_1 and UArizonaE ranked 14th for precision. Overall, UArizonaD is the best of our models: it has the highest F_1, the lowest $ERDE_{50}$, and the second-best recall.

7 Limitations

Our models fell short of the best system in the task for two main reasons. First, we attempted to predict depressed users from the beginning, even though the number of posts varies dramatically from user to user (from only 1 post per chunk to over 200 per chunk). A better strategy would have been to start making predictions after observing some threshold n posts, allowing us to predict early for users with many posts, while waiting until we have more information for users with few posts. Second, we did not sufficiently explore the broad range of possible features. For example, we could have built a domain-specific depression lexicon and used it instead of a previously collected lexicon, or we could have used more sophisticated techniques to represent posts as post-level feature vectors.

8 Conclusion and Future Works

In this paper, we described the techniques for the University of Arizona submissions to the 2017 CLEF eRisk early risk detection of depression pilot task. We used features based on a depression lexicon and on the Unified Medical Language System (UMLS). We implemented sequential and non-sequential models and used ensemble methods to leverage the best of each model. We found that the ensemble model works better than the individual models, and waiting for more data before making a decision improves the traditional performance measures like precision, recall and F_1. Whether it is acceptable to wait a decent amount of time to have better performance is still an open question – and we would like to work on that. We would like to establish a timeframe that can be deemed acceptable before making a decision so that the tradeoff between correctness and speed of the decision is minimized.

Acknowledgements

We would like to express our gratitude to the organizers of the pilot task for their effort towards building the first annotated user-level depression dataset.

References

1. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. In: Journal of the American Medical Informatics Association 17(3). pp. 229–236 (2010)
2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152. ACM (1992)
3. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
4. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) (2014)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
6. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: ICWSM. p. 2 (2013)
7. Goodwin, F.K., Jamison, K.R.: Manic-Depressive Illness: Bipolar Disorder and Recurring Depression. Oxford University Press Inc., New York (1990)
8. Greenberg, P., Kessler, R., Birnbaum, H., Leong, S., Lowe, S., Berglund, P., Corey-Lisle, P.: The economic burden of depression in the united states: how did it change between 1990 and 2000? In: J Clin Psychiatry 64(12). pp. 1465–1475 (2003)
9. Halfin, A.: Depression: the benefits of early and appropriate treatment. The American journal of managed care 13(4 Suppl), S927 (November 2007), `http://europepmc.org/abstract/MED/18041868`
10. 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 International Publishing (2016)
11. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2017. Dublin, Ireland (2017)

12. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech. vol. 2, p. 3 (2010)
13. Moreno, M., Jelenchick, L., Egan, K., Cox, E., Young, H., Gannon, K., Becker, T.: Feeling bad on facebook: depression disclosures by college students on a social networking site. In: Depression and Anxiety 28(6). pp. 447–455 (2011)
14. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998), `http://research.microsoft.com/\~jplatt/smo.html`
15. Sadeque, F., Pedersen, T., Solorio, T., Shrestha, P., Rey-Villamizar, N., Bethard, S.: Why do they leave: Modeling participation in online depression forums. In: Proceedings of the 4th Workshop on Natural Language Processing and Social Media. pp. 14–19 (2016)
16. Schwartz, H.A., Sap, M., Kern, M.L., Eichstaedt, J.C., Kapelner, A., Agrawal, M., Blanco, E., Dziurzynski, L., Park, G., Stillwell, D., et al.: Predicting individual well-being through the language of social media. In: Biocomputing 2016: Proceedings of the Pacific Symposium. pp. 516–527 (2016)
17. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1), 1929–1958 (2014)
18. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4(2) (2012)
19. Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., Chang, Y.: Understanding and discovering deliberate self-harm content in social media. In: Proceedings of the 26th International Conference on World Wide Web. pp. 93–102. International World Wide Web Conferences Steering Committee (2017)
20. WHO: The world health report 2001- mental health: New understanding, new hope. `http://www.who.int/whr/2001/en/whr01_en.pdf?ua=1` (2001), last Accessed: 2016-04-02
21. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations (1999)