Model	Features	Precision	Recall	F_1
SVM	Words	53.3	38.6	44.8
SVM	DepWords	**77.3**	20.5	32.4
SVM	DepWords+Metamap	49.0	30.1	37.3
SVM	DepEmbed	69.4	30.1	42.0
SVM	DepEmbed+DepWords	66.7	45.8	54.3
SVM	DepEmbed+DepWords+Metamap	53.9	66.3	59.5
GRU	Words	72.8	51.8	60.6
GRU	DepWords	62.0	68.7	65.1
GRU	DepWords+Metamap	67.0	**75.9**	**71.2**
GRU	DepEmbed	65.8	60.2	62.9
GRU	DepEmbed+DepWords	60.0	61.4	60.7
GRU	DepEmbed+DepWords+Metamap	60.0	62.7	61.2

Table 3: Comparison of different models and feature sets in five-fold cross-validations on the training set when considering the entire posting history (window=∞).

which model/feature combinations look most promising, so that those can be evaluated on the test set. We focus for now on the simpler setup where the model observes a user's entire posting history (window=∞), and is evaluated just in terms of precision and recall.

Table 3 shows the cross-validation performance of a variety of models on the training data. The best F_1, 71.2, is achieved by the sequential (GRU) model with depression words (DepWords) and UMLS medical concept (MetaMap) features. Comparing across types of models, the sequential models are the clear winners: even the worst sequential (GRU) model had a higher F_1 than the best non-sequential (SVM) model (60.6 vs. 59.5). This finding is intuitive, given that early detection is a sequential prediction problem. Comparing across types of features, adding medical concepts (MetaMap) always improved F_1, but the results for other types of features were more mixed. Depression embeddings (DepEmbed) always improved the non-sequential (SVM) models, but always hurt the sequential (GRU) models. And using all words (Words) was better than just the depression words (DepWords) for the non-sequential (SVM) model, but the reverse was true for the sequential (GRU) model.

Looking across all the models, we selected two models for evaluation on the test set: the best non-sequential (SVM) model (DepEmbed+DepWords+ Metamap) and the best sequential (GRU) model (DepWords+Metamap). For each of these models, we apply a risk window as described in Section 5, considering all possible risk windows between 0 and the maximum number of posts, and optimizing the window size to maximize cross-validation F_{latency} on the training set. For the SVM model, an 11-post risk window yields the highest F_{latency} (67.1, with an F_1 of 82.0), while for the GRU model, a 23-post risk window yields the highest F_{latency} (52.6, with an F_1 of 65.7).

6.3 Evaluation

Table 4 evaluates the best models on the eRisk 2017 test set. For contrast, we also show each model with a risk window of 0 (i.e., the first '+' or '-' prediction is final) and a risk window of ∞ (i.e., the model always waits for all of a user's posts and decides at the final post).

Model	Risk window	$ERDE_5$	$ERDE_{50}$	F_1	Latency	F_{latency}
SVM	0	13.1	9.7	51.3	63.5	38.9
SVM	11 (best)	13.6	10.1	51.4	75	36.8
SVM	∞	13.2	11.7	45.4	199	16.0
GRU	0	12.5	9.4	33.5	9	32.3
GRU	23 (best)	15.2	11.5	44.4	69.5	32.7
GRU	∞	15.0	13.6	45.0	199	15.8

Table 4: Comparison of the top non-sequential and sequential models (SVM:DepEmbed+DepWords+Metamap and GRU:DepWords+Metamap) on the test set. For contrast, the same models are also shown with risk windows of 0 and ∞.

Comparing ERDE to F_{latency}, we see that F_{latency} is better at discriminating between models. For example, the non-sequential (SVM) and sequential (GRU) models with risk window 0 have given very similar values for ERDE, with their $ERDE_5$s differing by only 0.6 points and their $ERDE_{50}$s differing by only 0.3 points. Yet these two models have hugely different performance characteristics: the GRU is extremely fast (latency 9) at a significant cost to accuracy (F_1 of 33.5), while the SVM is much more cautious (latency 63.5) and much more accurate (F_1 of 51.3). Table 4 also shows the challenge of setting the ERDE o parameter: with $o = 5$ as in eRisk 2017, ERDE can't distinguish (only a 0.1 point difference) between a non-sequential (SVM) model that sees a median of 63.5 posts (window=0) and one that sees a median of 199 posts (window=∞), despite the latter being much, much slower to make predictions. We see these empirical results as a strong indication that F_{latency} better captures the important evaluation characteristics of early detection problems.

We found that the models with risk windows optimized on the training set (SVM:window=11 and GRU:window=23) did not always outperform other simple choices of risk window (window=0 or window=∞) on the test set. While the 23-window GRU model indeed outperformed the F_{latency} of the other GRUs (GRU:0 and GRU:∞), the 11-window SVM model did not have a better F_{latency} than the 0-window SVM; the tiny improvement in F_1 achieved by SVM:11 over SVM:0 was outweighed by its larger jump in latency.

Despite the training set results where sequential models substantially out-performed non-sequential models, on the test set the no-risk-window non-sequential (SVM) model outperformed all sequential (GRU) models, in terms of both F_{latency} and F_1. But note that on the training set, we compared systems with access to the entire posting history (window=∞), and, as can be seen in Table 4, the performance of the SVM model is much worse with such a large risk window. Probably the simple way that the non-sequential model aggregates feature vectors makes it easy to lose the signal of a single depressed post in a sea of many non-depressed posts.

7 LIMITATIONS

First, while it would have been ideal to compare ERDE to F_{latency} on the actual systems submitted to eRisk 2017, since neither code nor predictions were publicly available for any of the top systems, we had to approximate such systems by exploring combinations of the most common models and features in the task. We believe this still results in a nice contribution, as the models and features can be more

directly compared, but since our models are re-implementations, their performance on the task may differ somewhat from the systems submitted to the task.

Second, during model selection we first selected model architectures and feature sets under the maximal risk window, and then searched over all possible risk window sizes to select the best model in terms of F_{latency}. This two-stage procedure was not ideal for our non-sequential models, where it turned out that using no risk window, instead of the maximal one, resulted in the best models on the test set. A better approach would be to optimize risk windows simultaneously with feature sets and model architectures. There are also probably further gains to be had by directly optimizing for F_{latency} (e.g., instead of accuracy) during model training.

Third, F_{latency} combines F_1 and latency under the assumption that systems generally want to optimize F_1. However different applications may need to optimize different evaluation measures. For example, if the goal is to have a human intervene when a risk of depression is detected in a social media user, then probably a high recall even at the expense of precision would be preferred, so that the human would be able to intervene wherever possible. On the other hand, if the goal is to have an automatic intervention when a depression risk is detected, then probably a high precision is needed so that the automatic intervention is only applied when the model is very certain of the depression risk. Future work may need to extend F_{latency} to such scenarios, perhaps by including something like F_β's parameter for trading off between precision and recall.

Finally, F_{latency} is a general metric, applicable to any problem where systems must examine a sequence of items associated with an object, and make a prediction about that object's class as rapidly as possible. However, in the current paper, we only explore F_{latency} as applied to early detection of depression on social media. Future work will need to investigate the utility of F_{latency} on other kinds of problems: detecting drug discontinuation, churn prediction, etc.

8 CONCLUSION

We introduced latency and F_{latency} as evaluation metrics for early detection tasks, and showed that the theoretical behavior of these metrics is preferable to the current state-of-the-art, early risk detection error (ERDE). We replicated common models and features from the eRisk 2017 shared task on early detection of depression in social media, and showed empirically that our metrics are better than ERDE at capturing important differences between models. We also introduced the concept of risk windows for helping models find an acceptable trade-off between precision and latency, and showed that it successfully improves the performance of sequential prediction models on the eRisk 2017 test set. We believe our proposed metrics can be useful in the broad range of applications where models need to make fast user-level predictions from a sequence of the user's social media interactions.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grant R01GM114355 from the National Institute of General Medical Sciences (NIGMS). The computations were done in systems supported by the National Science Foundation under Grant No. 1228509. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or National Science Foundation.

REFERENCES

[1] Hayda Almeida, Antoine Briand, and Marie-Jean Meurs. 2017. Detecting Early Risk of Depression from Social Media User-generated Content. In *8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017.*

[2] Alan R Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. In *Journal of the American Medical Informatics Association 17(3).* 229–236. https://doi.org/10.1136/jamia.2009.002733

[3] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32, suppl_1 (2004), D267–D270. https://doi.org/10.1093/nar/gkh061

[4] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8).*

[5] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29.

[6] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media.. In *ICWSM.* 2.

[7] Roger Detels. 2009. *The Scope and Concerns of Public Health.* Oxford University Press Inc., New York.

[8] Chris Ellis, Syed Zain Masood, Marshall F. Tappen, Joseph J. Laviola, Jr., and Rahul Sukthankar. 2013. Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition. *Int. J. Comput. Vision* 101, 3 (Feb. 2013), 420–436. https://doi.org/10.1007/s11263-012-0550-7

[9] Marcelo L. Errecalde, Ma. Paula Villegas, Dario G. Funez, Ma. JosÃĺ Garciarena Uelay, and Leticia C. Cagnina. 2017. Temporal Variation of Terms as concept space for early risk prediction. In *8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017.*

[10] Frederick K. Goodwin and Kay Redfield Jamison. 1990. *Manic-Depressive Illness: Bipolar Disorder and Recurring Depression.* Oxford University Press Inc., New York.

[11] Aron Halfin. 2007. Depression: the benefits of early and appropriate treatment. *The American journal of managed care* 13, 4 Suppl (November 2007), S92âĂŤ7. http://europepmc.org/abstract/MED/18041868

[12] Minh Hoai and Fernando De la Torre. 2014. Max-Margin Early Event Detectors. *International Journal of Computer Vision* 107, 2 (01 Apr 2014), 191–202. https://doi.org/10.1007/s11263-013-0683-3

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. International Conference on Learning Representation.

[15] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). http://arxiv.org/abs/1405.4053

[16] David Losada, Fabio Crestani, and Javier Parapar. 2017. CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum.*

[17] David E Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language Use. In *International Conference of the Cross-Language Evaluation Forum for European Languages.* Springer International Publishing, 28–39.

[18] Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. 2016. Data61-CSIRO systems at the CLPsych 2016 Shared Task.. In *CLPsych@ HLT-NAACL.* 128–132.

[19] Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting Post Severity in Mental Health Forums. In *The 3rd Workshop on Computational Linguistics and Clinical Psychology.* 133–137.

[20] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 Shared Task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology.* Association for Computational Linguistics, San Diego, CA, USA, 118–127. http://www.aclweb.org/anthology/W16-0312

[21] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm* 20 (2011), 265–272.

[22] Greenberg PE, Kessler RC, Birnbaum HG, Leong SA, Lowe SW, Berglund PA, and Corey-Lisle PK. 2003. The economic burden of depression in the United States: how did it change between 1990 and 2000?. In *J Clin Psychiatry 64(12).* 1465–1475.

[23] Brian A. Primack, Ariel Shensa, CÃĺsar G. Escobar-Viera, Erica L. Barrett, Jaime E. Sidani, Jason B. Colditz, and A. Everette James. 2017. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults. *Computers in Human Behavior* 69 (2017), 1 – 9. https://doi.org/10.1016/j.chb.2016.11.013

[24] Farig Sadeque, Ted Pedersen, Thamar Solorio, Prasha Shrestha, Nicolas Rey-Villamizar, and Steven Bethard. 2016. Why do they leave: Modeling participation in online depression forums. In *Proceedings of the 4th Workshop on Natural Language Processing and Social Media*. 14–19.

[25] Farig Sadeque, Dongfang Xu, and Steven Bethard. [n. d.]. UArizona at the CLEF eRisk 2017 Pilot Task: Linear and Recurrent Models for Early Depression Detection. ([n. d.]).

[26] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[27] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012).

[28] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 173–180.

[29] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2017. Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. In *8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017*.

[30] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. 2017. Understanding and Discovering Deliberate Self-harm Content in Social Media. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 93–102.

[31] World Health Organization WHO. 2001. 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.

[32] World Health Organization WHO. 2003. Global Burden of Disease (GBD) 2000: version 3 estimates. http://www.who.int/entity/healthinfo/gbdwhoregionyld2000v3.xls?ua=1. (2003). Last Accessed: 2016-04-08.

[33] Ian H Witten and Eibe Frank. 1999. Data mining: practical machine learning tools and techniques with Java implementations. (1999).

[34] Liu yi Lin, Jaime E. Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B. Colditz, Beth L. Hoffman, Leila M. Giles, and Brian A. Primack. 2016. Association between Social Media Use and Depression among U.S. Young Adults. In *Depression and Anxiety, 33(4)*. 323–331. https://doi.org/10.1002/da.22466

A APPENDIX: METAMAP

Metamap tries to identify UMLS concepts within a text, but has been designed primarily for biomedical use and not for social media. Below is an example of MetaMap applied to the following post from the depression subreddit:

> Nobody really gives a shit how depressed you are as long as you don't kill yourself, but if you did then they wonder why you didn't ask for help. This is a messed up world.

Running Metamap on this post will produce something like:

> Nobody really [gives: PREFERRED NAME='GIVE - DOSING INSTRUCTION IMPERATIVE', CUI='C1947971', SEMTYPES='[FTCN]', TRIGGER='["GIVE - DOSING INSTRUCTION IMPERATIVE"-TX-1-"GIVES"-VERB-0]'] a shit how [depressed: PREFERRED NAME='DEPRESSED MOOD', CUI='C0344315', SEMTYPES='[FNDG]', TRIGGER='["DEPRESSED MOOD"-TX-1-"DEPRESSED"-VERB-0]'] you are as [long: PREFERRED NAME='LONG', CUI='C0205166', SEMTYPES='[QLCO]', TRIGGER='["LONG"-TX-1-"LONG"-ADV-0]'] as you don't [kill: PREFERRED NAME='KILLING', CUI='C0162388', SEMTYPES='[SOCB]', TRIGGER='["KILLING"-TX-1-"KILL"-VERB-0]'] yourself, but if you did then they wonder why you didn't ask for [help: PREFERRED NAME='ASSISTED (QUALIFIER VALUE)', CUI= 'C1269765', SEMTYPES='[QLCO]', TRIGGER='["ASSI-STED (QUALIFIER VALUE)"-TX-1-"HELP"-VERB-0]']. This is a messed up world.

Metamap finds trigger words (i.e. depressed) and then maps it to a UMLS concept that has a preferred name (Depressed mood), a Concept Unique Identifier or CUI (C0344315), a semantic type (fndg or Finding) and some properties of the trigger itself.