Incorporating Risk Factor Embeddings in Pre-trained Transformers Improves Sentiment Prediction in Psychiatric Discharge Summaries

Xiyu Ding[1] and **Mei-Hua Hall**[2,3] and **Timothy A. Miller**[1,3]
[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA
[2]Psychosis Neurobiology Laboratory, McLean Hospital, Belmont, MA
[3]Harvard Medical School, Boston, MA

Abstract

Reducing rates of early hospital readmission has been recognized and identified as a key to improve quality of care and reduce costs. There are a number of risk factors that have been hypothesized to be important for understanding re-admission risk, including such factors as problems with substance abuse, ability to maintain work, relations with family. In this work, we develop RoBERTa-based models to predict the sentiment of sentences describing readmission risk factors in discharge summaries of patients with psychosis. We improve substantially on previous results by a scheme that shares information across risk factors while also allowing the model to learn risk factor-specific information.

1 Introduction

About 1 in 5 Medicare patients discharged from the hospital is rehospitalized within 30 days (Jencks et al., 2009). Four out of the top ten conditions with the largest number of readmissions among the Medicaid enrollees were mental health conditions or substance use disorders (Hines et al., 2014). Readmissions are harmful both in being disruptive to patients and families, and as a major driver of health-care costs in psychiatry (Wu et al., 2005; Mangalore and Knapp, 2007). Also, premature discharge of patients contributes not only to rehospitalization but to increased risk of homelessness and the possibility of violent behavior or suicide. Thus, reducing rates of early hospital readmission has been recognized and identified as a key to improve quality of care and reduce costs.

There are a number of risk factors that previous work compiled from the literature (Holderness et al., 2018) to be important for understanding readmission risk, including such factors as problems with substance abuse, ability to maintain work, relations with family. Studying readmission through the use of explicit risk factors may go against prevailing trends in machine learning (black box models that take in raw signals), but has at least two potential benefits: 1) Allowing for descriptive study of, and thus better understanding of, the factors that are important for readmission; and 2) The possibility that a risk classifier that uses explicit risk factors will be more interpretable and trustworthy to providers, and thus more likely to be put into practical use.

In this work, we make use of a publicly available data set of sentences from discharge summary that is annotated for seven risk factors, along with the "sentiment," marked as *Positive*, *Negative*, or *Neutral* (more details in background). Previous work on this dataset showed that the task was approachable with neural methods, but performance suffered because of small dataset size (Holderness et al., 2019). Here, we address this issue with two advances. First, we show that transfer learning helps dramatically. While the previous work used smaller neural models trained from scratch, we start from pre-trained transformer models (RoBERTa (Liu et al., 2019)), and improve them for the task with architectural and data augmentation strategies.

Second, we show that sharing data between the different risk factor domains is better for performance than training separately, at least with pre-trained models. Previous work trained separate classifiers for each of the risk factor domains, due to the (legitimate) belief that there are significant differences in predicting sentiment in different risk factor domains (Holderness et al., 2019). Here, we demonstrate that a method that allows for sharing of information between different risk factor domains can improve performance. Because of a mismatch in the way training and test data were annotated, we introduce a data augmentation method for the training set that allows our method to improve over the vanilla RoBERTa baseline. The improvements

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 35–40

from this latter technique point in a direction that could allow for further improvements even without additional training data.

2 Background

The dataset we use was created in previous work (Holderness et al., 2018) and made publicly available. It consists of sentences extracted from discharge summaries of patients with psychosis at Boston-area hospitals. Previous work had examined two tasks on this dataset, the identification of the risk domain represented by the sentence (*Appearance*, *Mood*, *Interpersonal*, *Substance Use*, *Occupation*, *Thought Content* and *Thought Process*), and the sentiment of the sentence given the risk domain. In the training data, each sentence has only one risk factor, and thus one sentiment, but the test data allows for sentences to have multiple risk factors, and thus multiple different sentiments for a single sentence, conditioned on the risk factor domain. Both the previous work for risk factor classification (Holderness et al., 2018) and sentiment classification (Holderness et al., 2019) used multilayer neural networks in their experiments and found them to be the best-performing.

Despite the promise of the above work, the tasks are still unsolved. Recent work in contextualized embeddings has shown great success for sentence classification tasks. Specifically, BERT-based models (Devlin et al., 2019), based on the transformer architecture (Vaswani et al., 2017), showed that pre-training deep transformer encoders on massive text datasets with a language modeling objective could lead to improvements in a variety of tasks. The best performance is typically obtained by fine tuning, where a classifier head is attached to a special sentence token, and new tasks are learned via standard supervised learning, in which the weights of the classifier head are trained from scratch while the weights of the transformer encoder are allowed to update. In this work, we make use of the RoBERTa updates to BERT (Liu et al., 2019), which use the same architecture but pre-trained on a larger dataset and for a longer time.

3 Data

The training dataset contains 3500 sentence-length texts, 500 from each of the seven readmission risk factors mentioned above. The test dataset contains 1650 texts which can involve multiple risk factors and are more variable in length compared with the training data, as described in the previous study (Holderness et al., 2019). We divided the training instances into training and development set with an 80%/20% split,[1] leaving us with 2800 training instances and 700 development instances.

Since we are focusing on sentiment prediction in this work, we take the risk factor domains as given, and for test sentences with multiple risk factor domains, we create multiple instances where the input pairs the sentence text with each domain, and the target is the gold sentiment label for that domain. This results in 2103 test instances. 750 of these are labeled as the *Other* domain, which does not have training instances nor reported results, so we discard these, leaving 1353 test instances.

4 Methods

We developed several variations of a risk factor sentiment classifier based on the RoBERTa architecture.

4.1 Baseline

We fine-tune seven independent RoBERTa models as the baseline, one for each of the risk factor domains. This follows previous work (Holderness et al., 2019), which suggested that positive or negative clinical sentiments might differ in different risk factor domains.

4.2 Plain RoBERTa

Instead of fine-tuning seven independent models, we fine-tune one RoBERTa model on all training texts to learn the shared representation of sentiments, ignoring the risk factor domain of the sentences during training and only learning sentiment labels. Since the test set allows for a sentence to have multiple risk factor domains, but this model can only make one sentiment prediction per sentence, the model is penalized on cases where a sentence has multiple risk factor domains with different sentiments.

4.3 Risk Factor Domain Embeddings

In this method, we modify the input representation to contain both the input sentence and the risk factor domain to be classified. In BERT-style models, this means using the special sentence-separating token ([SEP]) between the sentence tokens and the domain tokens. For example, the first sentence in

[1] Specifically, we use *iterative_train_test_split* from scikit-multilearn (Szymański and Kajdanowicz, 2017) to create a split that is stratified with respect to multiple labels.

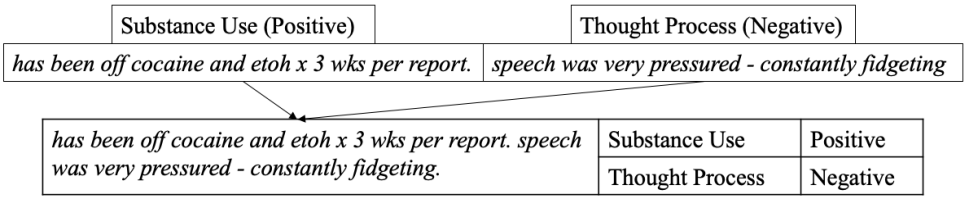

Figure 1: One example of the augmented sentences. Two randomly chosen sentences from the training data are concatenated to create two training instances with different risk factor domains.

Figure 1 would be represented as *[CLS] has been off cocaine and etoh x 3 wks per report. [SEP] substance use [SEP]*. Giving the domain as the second sentence allows the model to potentially learn what part of the input the classifier should focus on, and previous work has shown similar methods to be effective (Shi and Lin, 2019). In contrast to another possible approach with a separate input stream for domain feature, this method allows us to utilize the pre-trained contextual embeddings of the domain words and let the model learn to use its attention mechanism to relate the risk factor domains with the corresponding part of the sentences and make sentiment predictions about only that part.

4.4 Data Augmentation

The training data only has one risk factor domain and sentiment annotation per sentence, and so the above approach alone may fail on test data, because at training time the model may never need to use the domain embedding, because the whole sentence is relevant because of the way the data was constructed. Then when applied to the test data the risk factor domain embedding could be useful, but the model has never been trained to use it.

Therefore, we developed a data augmentation scheme that creates new synthetic training instances from pairs of existing instances, such that these synthetic instances more closely resemble test instances with multiple risk factor domain and sentiment labels. We concatenate two randomly sampled sentences in the training data from two different domains, and create two training instances from that concatenated input, with the two different domain and sentiment labels from the original instances. We add a period and space between sentences missing them to look more natural. Figure 1 shows an example of the augmented sentences. These synthetic instances are surely lacking discourse coherence that the real instances in the test set will have, but our hypothesis is that they will at least force the model to associate domain embeddings with specific parts of the input at training time and not just learn the sentiment for the whole input sequence.

We created 2800 new training instances with this sampling procedure (1400 unique texts), for an augmented training data size of 5600 instances (this method is only use for augmenting training and is not applied to the test data). We then fine-tune the model from the previous section with this augmented data set, using the domain risk factor as the second sentence as above.

4.5 Training Details

We use the HuggingFace Transformer library (Wolf et al., 2019) for our RoBERTa implementations. We use grid search to find the optimal hyper-parameters (batch size, learning rate and max sequence length) for the RoBERTa fine-tuning process. We monitor the training and validation loss for each training epoch and save the model with the highest Macro-F1 score on the development set before testing on the test set. We use pandas for data processing (Wes McKinney, 2010) and scikit-learn for model evaluation (Pedregosa et al., 2011).

Model	Neu F1	Pos F1	Neg F1	Acc	Macro F1
Baseline (Indep.)	0.375	0.623	0.698	0.591	0.565
Roberta	0.445	0.727	**0.789**	0.690	0.653
Roberta+D	0.456	0.723	0.778	0.680	0.652
Roberta+D+Aug	**0.468**	**0.743**	**0.789**	**0.699**	**0.666**

Table 1: Modeling results for the four architectures with the highest score on each performance metric in bold.

5 Evaluation

We evaluated the four different architectures on this task to explore the importance of sharing information between different risk factor domains as well as learning the domain specific information.

We first evaluate at the instance level, computing precision and recall for each sentiment label, and combining to get an F1 score for each sentiment label, as well as overall accuracy (proportion of instances correctly predicted regardless of sentiment or risk factor domain). Table 1 summarizes the accuracy F1 score results, with Macro F1 also

Model	Domain	Pos P	Pos R	Pos F1	Neg P	Neg R	Neg F1	Neu P	Neu R	Neu F1
Holderness et al. (2019)	**Average**	0.62	0.416	0.478	0.67	0.652	0.658	0.289	0.437	0.329
Roberta	**Average**	0.760	0.671	0.709	0.802	**0.785**	**0.790**	0.385	0.477	0.418
	Interpersonal	0.665	0.716	0.690	0.779	0.782	0.779	0.619	0.580	0.597
	Mood	0.846	0.738	0.787	0.750	0.818	0.780	0.510	0.528	0.518
	occupation	0.791	0.590	0.674	0.688	0.759	0.719	0.409	0.514	0.451
	Substance Use	0.646	0.605	0.620	0.808	0.765	0.789	0.392	0.468	0.423
	Appearance	0.854	0.647	0.735	0.852	0.868	0.857	0.395	0.565	0.465
	Thought Content	0.768	0.697	0.730	0.892	0.728	0.800	0.067	0.237	0.108
	Thought Process	0.748	0.706	0.728	0.842	0.772	0.807	0.306	0.447	0.362
Roberta+D+Aug	**Average**	**0.779**	**0.708**	**0.735**	**0.860**	0.713	0.768	**0.404**	**0.581**	**0.459**
	Interpersonal	0.654	0.822	0.724	0.896	0.646	0.748	0.653	0.720	0.682
	Mood	0.873	0.714	0.785	0.750	0.883	0.810	0.480	0.456	0.464
	Occupation	0.838	0.656	0.733	0.896	0.607	0.722	0.475	0.769	0.585
	Substance Use	0.680	0.556	0.606	0.833	0.824	0.829	0.415	0.489	0.446
	Appearance	0.882	0.776	0.824	0.893	0.491	0.626	0.313	0.762	0.438
	Thought Content	0.737	0.636	0.680	0.882	0.732	0.799	0.114	0.410	0.179
	Thought Process	0.787	0.798	0.790	0.872	0.811	0.840	0.381	0.463	0.416

Table 2: Results of Roberta and Roberta fine-tuned on augmented dataset (Roberta+D+Aug) with the highest score on each performance metric in bold. The top row is reported results of the "Fully Supervised MLP" system of Holderness et al. (2019). "Average" scores are computed by taking the average of 7 risk factor domains in the same column.

Example input text	Domain	GOLD	Roberta	Roberta+D+Aug
A. Pt.'s **affect appeared** slightly brighter, **but** remains **flat** at baseline. Her **mood** appears **slightly improved** as well.	Mood	Positive	Positive	Positive
	Appearance	Neutral	Positive	Neutral
B. Discussed pt.'s **job** and recent episode of "**crying** but not knowing why."	Mood	Neutral	Neutral	Negative
	Occupation	Neutral	Neutral	Neutral
C. discussed **work dynamics**, **MI** for **substance abuse**, processing the episode, medication planning	Occupation	Neutral	Neutral	Neutral
	Substance Use	Negative	Neutral	Neutral

Table 3: Three examples of the sentiment extraction results in the test data

reported as the average across sentiment labels. Table 2 shows detailed results of the plain RoBERTa and the augmented (RoBERTa+D+Aug) model, including precision and recall, and broken down by risk factor domain. The average across risk factor domains is substantially higher than the results reported in previous work.

The overall best performance was obtained by using domain embeddings with data augmentation (RoBERTa+D+Aug). Although fine-tuning seven independent RoBERTa models for each risk factor domain is the worst performing model, its baseline scores are still higher than the previous study which did not use pre-trained models (Holderness et al., 2019). Plain RoBERTa is surprisingly strong in this task despite the fact that it ignores the risk factor domain and is forced to make the same predictions for texts with multiple sentiment labels.

The improvement obtained by fine-tuning one single RoBERTa model instead of seven independent models suggests that the benefits of sharing information between risk factor domains outweigh the potential risks that the model will learn conflicting information. However, simply using domain as the second sentence during RoBERTa fine-tuning does not lead to improvement in the overall model performance. Augmenting the training data to look more like the test data was necessary in order for the domain embedding input to show benefits.

We tested the significance of the improvements in accuracy and Macro-F1 between RoBERTa and RoBERTa+D+Aug by fine-tuning with 20 randomly selected seeds, and performing a one-tail t-test, and results were found to be significant ($p<0.05$, $p<0.005$ for accuracy and Macro-F1).

6 Discussion and Conclusion

We selected instances from the test set where the system trained on augmented data (RoBERTa+D+Aug) did well, and others where it did not (see Table 3). Example A shows that the system is able to distinguish between the positive mood and neutral appearance, where plain RoBERTa was forced to select a single sentiment label. Example B shows where RoBERTa+D+Aug still makes mistakes – plain RoBERTa actually does better by picking the single sentiment for the sentence that fits best, while RoBERTa+D+Aug tries to pick two different sentiments and gets one wrong. In fact, in 66.5% of the 221 test sentences with multiple risk factor domains, the sentiment labels are the same, which means plain RoBERTa

is usually not penalized for picking a single sentiment label. Example C shows an example where both models make errors, probably due to missing the complex inference that the patient has a substance abuse issue that requires treatment (MI=motivational interviewing).

Overall, our new approach shows major gains in performance over the existing state of the art for this problem. The biggest gains come from simply using large pre-trained models. However, the modified architecture and data augmentation technique lead to further gains, and have the ability to separate out multiple sentiments for a single sentence on new data. Future work may see larger benefit with methods for creating augmented training data that create more natural-looking sentence pairs. The source code used to fine-tune the model will be made publicly available [2].

Acknowledgments

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Numbers R01LM012918 and R01LM012973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anika L. Hines, Marguerite L. Barrett, Joanna Jiang, and Claudia A. Steiner. 2014. Conditions With the Largest Number of Adult Hospital Readmissions by Payer, 2011. HCUP Statistical Brief #172, Agency for Healthcare Research and Quality, Rockville, MD.

Eben Holderness, Philip Cawkwell, Kirsten Bolton, James Pustejovsky, and Mei-Hua Hall. 2019. Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 117–123, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eben Holderness, Nicholas Miller, Kirsten Bolton, Philip Cawkwell, Marie Meteer, James Pustejovsky, and Mei Hua-Hall. 2018. Analysis of risk factor domains in psychosis patient health records. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 129–138, Brussels, Belgium. Association for Computational Linguistics.

Stephen F. Jencks, Mark V. Williams, and Eric A. Coleman. 2009. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428. PMID: 19339721.

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*.

Roshni Mangalore and Martin Knapp. 2007. Cost of schizophrenia in England. *The journal of mental health policy and economics*, 10(1):23–41. Place: Italy.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling.

P. Szymański and T. Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Eric Q. Wu, Howard G. Birnbaum, Lizheng Shi, Daniel E. Ball, Ronald C. Kessler, Matthew Moulis,

[2] https://github.com/Machine-Learning-for-Medical-Language/psychosis-sentiment-data-augmentation

and Jyoti Aggarwal. 2005. The economic burden of schizophrenia in the United States in 2002. *The Journal of clinical psychiatry*, 66(9):1122–1129. Place: United States.