The 35th Canadian Conference on Artificial Intelligence
DOI: 0
Canadian AI 2022
30 May–3 June 2022, Toronto

# On the Influence of Annotation Quality in Suicidal Risk Assessment from Text

Fanny Rancourt[†], Diego Maupomé[†], Marie-Jean Meurs[†,*]
[†] Université du Québec à Montréal

**Abstract**

In applying Natural Language Processing to support mental health care, gathering annotated data is difficult. Recent work has pointed to lapses in approximative annotation schemes. While studying gaps in prediction accuracy can offer some information about these lapses, a more careful look is needed. Through the use of Influence Functions, quantification of the relevance of training examples according to their type of annotation is possible. Using a corpus aimed at suicidal risk assessment containing both crowdsourced and expert annotations, we examine the effects that these annotations have on model training at test time. Our results indicate that, while expert annotations are more helpful, the difference with respect to crowdsourced annotations is slight. Moreover, most globally helpful observations are crowdsourced, pointing to their potential.

**Keywords:** natural language processing, mental health, influence functions, explainability, annotation

## 1. Introduction

Recently, there has been increased research interest in utilizing Natural Language Processing (NLP) techniques in the service of mental health care. Mental health is a major public health issue, accounting for 13% of the global burden of disease [1], leading to higher rates of morbity and mortality [2]. Thus, early intervention in mental health has become a key issue in service reform [3, 4]. NLP can play an important part in this aspect of care by analyzing textual content from social media, widely used by at-risk persons to discuss their mental state [5]. NLP research efforts will attempt to infer models that predict the mental health status of a person given their writings on various online social media. This assessment will usually pertain to specific mental health disorders, symptoms or harmful behaviors. However, there is large variation in the nature of these assessments. Clinically grounded assessments are costly. Where annotation in other NLP tasks can be carried out on the documents to be analyzed, in the context of mental health, annotation pertains to the author, and clinically grounded annotation requires access to this person. This makes the data collection process difficult. To boot, sophisticated prediction models often require large amounts of training data. This has given rise to a variety of what could be understood as approximations to clinical truth. These vary from the use of affiliation to specific fora or groups [6] and mentions of diagnosis [7] to the use of clinical self-report tools [8].

However, there is concern about the validity of these approaches. That is, while models can be developed to accurately predict these assessments on unseen data, it is possible that these assessments capture a different construct than the desired ones, *i.e.* the clinically actionable aspects of mental health of interest [9]. Evaluating the predictive performance of models issuing from these annotation schemes on clinically grounded data can offer some insight into this phenomenon [10]. Nonetheless, a closer examination of the inner workings of these models may offer richer information as to whether these annotations remain disjointed from clinically sounder ones.

In addition to explanations aiming at external stakeholders [11], explainability techniques can provide useful insights for the development and deployment of machine learning models [12, 13]. In NLP, this can take the form of feature importance [14, 15] and saliency

[*] meurs.marie-jean@uqam.ca

maps [16] techniques, which can give insights regarding words or spans from the input text. Alternatively, methods explaining with examples [17] such as Influence Functions (IF) can give valuable insights on such complex tasks. In fact, evidence suggests that IF are more appropriate explanations than saliency maps for non trivial NLP tasks such as language understanding [18]. Risk assessment of mental health issues from social media content is such a complex classification task as the labels usually require additional information about the authors. The work presented in this paper seeks to study the use of IF in mental health assessment by NLP. Specifically, its goals are to:

RQ1 Examine the impact of crowdsourced annotations on model predictions.
RQ2 Assess whether crowdsourced annotations are globally more influential than expert annotations.

The paper is structured as follows. Section 2 defines influence functions and Section 3 discusses the methodology used to conduct our experiments. Finally, Section 4 discusses the results, and Section 5 concludes this paper with potential future work.

## 2. Influence Functions

Influence Functions (IF) – a classic notion from robust statistics – monitor the changes occurring after small modifications to the problem formulation [19]. This analysis rely on the hypothesis that slight perturbations should cause at most small variations on the results of a model or a statistical test [20]. The most popular of said perturbations regards the distribution of the observations.

Given a machine learning model, IF aim to answer the following question: *what if we had a different training set?* By infinitesimally up-weighting an observation $z$, IF allow practitioners to assess the influence of this data point on a given prediction: a large loss variation indicates that $z$ is *influential*. From an explainability standpoint, the sign of the loss variation indicates whether $z$ is helpful or not for the prediction. Moreover, one can also probe local robustness by monitoring parameter change [21]. Other alternatives, such as assessing the stability of the predictions between both sets of parameter values, can also be considered. Thus, IF can improve algorithmic transparency as defined by Lipton [17] and can indicate the presence of predictive uncertainty [22] More formally, let $\mathcal{D} = \{(\mathbf{x}_k, y_k) : 1 \leq k \leq n\}$ be the training data of a classic supervised task. The learned parameters $\hat{\theta}$ are obtained by resolving optimization problem $\arg\min_\theta \sum_k \mathcal{L}(z_k, \theta)$, where $\mathcal{L}$ is the loss function combined with regularization factors if applicable. To assess the influence of $z_i \in \mathcal{D}$, we consider the parameter change occurred when up-weighting it by $\epsilon$, *i.e.* $\hat{\theta}_{\epsilon,i} := \arg\min_\theta \sum_k \mathcal{L}(z_k, \theta) + \epsilon\, \mathcal{L}(z_i, \theta)$. The parameter change is approximately

$$\hat{\theta}_{\epsilon,i} - \hat{\theta} \approx \epsilon \left.\frac{\mathrm{d}\,\hat{\theta}_{\epsilon,i}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -\epsilon H_{\hat{\theta}}^{-1} \nabla_\theta \mathcal{L}(z_i, \hat{\theta}) \tag{2.1}$$

where $H_{\hat{\theta}}^{-1} := \frac{1}{n} \sum_k \nabla_\theta^2 \mathcal{L}(\hat{\theta}, z_k)$. Calculation details of the right term in (2.1) are provided in Section 5.2 of [19]. Some properties of this estimator are discussed in [23]. To assess whether an observation $z_i$ is helpful or not to predict another observation $\tilde{z}$, we refer to the variation of the loss function

$$\mathcal{I}(\tilde{z}, i) := \left.\frac{\mathrm{d}\,\mathcal{L}(\tilde{z}, \hat{\theta}_{\epsilon,i})}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -\nabla_\theta \mathcal{L}(\tilde{z}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \mathcal{L}(z_i, \hat{\theta}) \tag{2.2}$$

obtained with the chain rule [21]. For (2.1) and (2.2) to hold, the loss function must be twice differentiable and convex, assumptions that most machine learning models do not uphold. Nonetheless, this approximation can remain fairly accurate for Transformer models [13, 18] and smaller neural architectures [21, 24].

## 3. Methodology and Resources

**Influence Functions.** IF are particularly useful for error analysis as they bring forth "insights about how models rely on and extrapolate from the training data" [21]. Hence, we use IF to examine prediction errors and assess whether there is a correlation with the annotator of those influential (helpful or harmful) examples with respect to (2.2). Additionally, IF can give insights regarding latent structures from the training set [13]. To examine RQ2, the average influence as well as globally influential examples are considered. A data point is *globally influential* when the absolute value of its influence ranks among the highest 25% for at least half the test examples. As the dominant class comprises 44% of the test set (see Table 1), this guarantees that these examples are influential for multiple risk levels. FastIF [13] is used to compute influence functions. A damping parameter of 5E-3 is added to the diagonal of $H_{\hat{\theta}}$ to ensure that all the eigenvalues are positive.

**Data.** The experiments conducted concern suicidal risk assessment using data from CLPsych [25, 26]. The data consists of posts written on the Reddit forum r/SuicideWatch, which aims to support peers struggling with suicidal ideation (This corresponds to CLPsych 2019 Task A v2.). The suicidal risk estimated for each author is placed on a four-point scale – None, Low, Moderate, and Severe. The task is modified slightly for the purposes of this investigation: each post is treated as a single observation, using the label of its author as its own. The task then becomes one of document classification, simplifying the architecture. While some authors span several posts, 73% count exactly one post. Further, given that annotation was performed on the basis of these documents, this simplification remains sound.

Lastly, the original dataset is divided between observations annotated by crowdsource workers and experts. As such one additional modification was made to it: whereas expert-annotated examples were used only in testing by [25], expert-annotated examples were used in training in the present work so as to be assess their influence. Specifically, expert annotations are spread into the training, validation and test sets at rates of 60%, 20% and 20%, respecting label proportions. Table 1 presents the distribution of risk levels among different annotation sources. For further details regarding the annotation guidelines and process, see [25]. Word counts for each set are presented in Table 5 (see Appendix A).

**Model.** The classifier is an adaptation of the RoBERTa model [27]. In order to keep the model small, the parameters of all but the topmost Self-Attention layers are left fixed. To mitigate class imbalance, observations are weighted in inverse proportion to the weight of their class. The model is trained over 30 epochs with mini-batches of 32 observations, using the Adam optimizer ($\beta_1 = .9, \beta_2 = .999$) [28] with a learning rate of 5E-5 and a weight decay of 1E-5. At the end of each epoch, the model is evaluated in order to select the best-performing point. This evaluation is done on a randomly selected validation set with equal proportions of each of the four classes. The selection is based on the macro-aggregated f1-score.

## 4. Results

Models are evaluated using macro-averaged precision, recall and f1-score, counteracting label imbalance. Classification results are presented in Table 2. Furthermore, to assess

| | training set | | validation set | | test set | |
|---|---|---|---|---|---|---|
| Risk level | crowdsourced | expert | crowdsourced | expert | crowdsourced | expert |
| None | 130 | 28 | 32 | 8 | 34 | 9 |
| Low | 48 | 47 | 11 | 15 | 13 | 15 |
| Moderate | 124 | 99 | 30 | 31 | 41 | 32 |
| Severe | 436 | 57 | 108 | 18 | 98 | 18 |

*Table 1.* Distribution of risk levels of each post according to their annotator

| Risk level | precision | recall | f1-score |
|---|---|---|---|
| None | .59 | .77 | .67 |
| Low | .32 | .43 | .36 |
| Moderate | .30 | .18 | .22 |
| Severe | .51 | .53 | .52 |
| Macro-avg | .43 | .48 | .45 |

*Table 2.* Test results on CLPsych data combining expert and crowdsourced annotations for training and test sets

| | crowdsourced | | | | expert | | | |
|---|---|---|---|---|---|---|---|---|
| | None | Low | Moderate | Severe | None | Low | Moderate | Severe |
| None | 25 | 1 | 2 | 6 | 8 | 0 | 1 | 0 |
| Low | 1 | 6 | 1 | 5 | 3 | 6 | 2 | 4 |
| Moderate | 4 | 9 | 9 | 19 | 0 | 2 | 4 | 26 |
| Severe | 15 | 12 | 18 | 53 | 0 | 2 | 7 | 9 |

*Table 3.* Confusion matrix of test set predictions with respect to annotation source

| | RoBERTa | | | CLaC [29] |
|---|---|---|---|---|
| Risk level | precision | recall | f1-score | f1-score |
| None | .86 | .75 | .80 | .74 |
| Low | .38 | .38 | .38 | .24 |
| Moderate | .37 | .46 | .41 | .40 |
| Severe | .63 | .60 | .61 | .54 |
| Macro-avg | .56 | .55 | .55 | .48 |

*Table 4.* Results obtained training and testing only on crowdsourced annotations against best results at CLPsych2019

the relevance of the classification approach, a second model was trained with the same configurations using the training and test sets from [26]. At test time, the highest-risk label predicted for a document becomes the predicted label for its author. Results are presented in Table 4.

**RQ1.** As our results show, the Moderate risk category appears to be poorly captured by our model. This label is seldom predicted, and most Moderate risk observations are assigned Severe risk instead (see Table 3). Furthermore, throughout misclassified Moderate risk examples, the majority of highly helpful examples are from the Severe class. Of those, 88% were labeled by crowdsourced annotators. This could indicate the presence of mislabeled data: Moderate risk being classified as Severe by crowdsourced annotators. This is reflected by the higher frequency of Severe risk in crowdsourced annotation, as compared to expert annotation (see Table 1), and is consistent with previous findings [25].

While error rates are similar between annotators, the model is more likely to underestimate the risk level of crowdsourced annotated data. In particular, 15% of Severe risk crowdsourced-annotated data are classified as No risk while none of the expert-annotated ones are. Further, the prediction of Moderate risk data follows a similar pattern. This is problematic for triage-based applications, given that a high recall for high risk documents is of utmost importance.

**RQ2.** Among training examples found to be helpful to classification, expert-annotated ones had higher influence on average than examples issuing from crowdsourced annotation (0.673 vs 0.561). In contrast, harmful examples from each annotation source had comparable influence (0.441 vs 0.443). This further suggests that expert annotations are of greater quality as fine-tuning on them would likely improve the model performances [13]. Nonetheless, the gap between annotation types is modest, which indicates that crowdsourced annotations have

value. Additionally, most globally helpful examples are labeled as Severe risk, 14% of which were expert-annotated. Thus, crowdsourced annotation actively contributes to improve our model. In contrast, the examples most harmful to classification are labeled as Moderate risk, with the balance falling slightly to crowdsourced annotation.

## 5. Conclusion and Future work

Our experiments demonstrate the potential of IF in analyzing the effects of annotation quality on model predictions in suicidal risk assessment. Given, the importance of establishing the soundness of annotation in mental health applications, this area warrants further study. Error correction could be applied in order to improve on our results, particularly for low- and moderate- risk examples. Additional improvements could be made by considering pretrained language models trained on domain-specific data or using domain adaptation.
**Reproducibility.** The source code is licensed under the GNU GPLv3 and the data are provided on demand by the CLPsych organizers.

## Acknowledgements

FR would like to thank Frédéric Branchaud-Charron for insightful discussions, and we also thank Dr. Leila Kosseim for her support. This research was enabled by Calcul Québec resources. MJM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada [NSERC Grant number 06487-2017] and the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFE-2018-00484].

## Appendix A. Data statistics

| | training set | | validation set | | test set | |
|---|---|---|---|---|---|---|
| Risk level | mean | std | mean | std | mean | std |
| None | 146 | 172 | 160 | 217 | 154 | 183 |
| Low | 268 | 306 | 361 | 415 | 211 | 184 |
| Moderate | 235 | 255 | 272 | 273 | 210 | 190 |
| Severe | 213 | 242 | 174 | 203 | 251 | 211 |

*Table 5.* Mean and standard deviation of word counts

## References

[1] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al. "Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* (2018).

[2] S. Saxena. "Excess mortality among people with mental disorders: a public health priority". In: *The Lancet Public Health* (2018).

[3] M. Schotanus-Dijkstra, C. H. C. Drossaert, M. E. Pieterse, B. Boon, J. A. Walburg, and E. T. Bohlmeijer. "An early intervention to promote well-being and flourishing and reduce anxiety and depression: A randomized controlled trial". In: *Internet Interventions* (2017).

[4] P. D. McGorry and C. Mei. "Early intervention in youth mental health: progress and future directions". In: *Evidence-Based Mental Health* (2018).

[5] H.-C. Shing, P. Resnik, and D. W. Oard. "A Prioritization Model for Suicidality Risk Assessment". In: *Annual Meeting of the Association for Computational Linguistics*. 2020.

[6] S. Chancellor, T. Mitra, and M. De Choudhury. "Recovery amid pro-anorexia: Analysis of recovery in social media". In: *Conference on Human Factors in Computing Systems*. 2016.

[7] G. Coppersmith, M. Dredze, and C. Harman. "Quantifying mental health signals in Twitter". In: *Workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 2014.

[8] D. E. Losada, F. Crestani, and J. Parapar. "eRisk 2020: Self-harm and depression challenges". In: *European Conference on Information Retrieval*. Springer. 2020, pp. 557–563.

[9] S. Chancellor and M. De Choudhury. "Methods in predictive techniques for mental health status on social media: a critical review". In: *npj Digital Medicine* (2020).

[10] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury. "Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals". In: *CHI Conference on Human Factors in Computing Systems*. 2019.

[11] U. Bhatt, M. Andrus, A. Weller, and A. Xiang. "Machine Learning Explainability for External Stakeholders". In: *arXiv preprint arXiv:2007.05408* (2020).

[12] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley. "Explainable Machine Learning in Deployment". In: *Conference on Fairness, Accountability, and Transparency*. 2020.

[13] H. Guo, N. F. Rajani, P. Hase, M. Bansal, and C. Xiong. "FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging". In: *arXiv preprint arXiv:2012.15781* (2020).

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?" Explaining the Predictions of Any Classifier". In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

[15] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* (2017).

[16] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *Workshop at International Conference on Learning Representations*. 2014.

[17] Z. C. Lipton. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *ACM Queue* (2018).

[18] X. Han, B. C. Wallace, and Y. Tsvetkov. "Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions". In: *Annual Meeting of the Association for Computational Linguistics*. 2020.

[19] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. 1982.

[20] P. J. Huber. "Robust Statistics". In: *Wiley Series in Probability and Mathematical Statistics* (1981).

[21] P. W. Koh and P. Liang. "Understanding Black-box Predictions via Influence Functions". In: *International Conference on Machine Learning*. 2017.

[22] P. Schulam and S. Saria. "Can You Trust This Prediction? Auditing Pointwise Reliability After Learning". In: *International Conference on Artificial Intelligence and Statistics*. 2019.

[23] L. T. Fernholz. *von Mises Calculus For Statistical Functionals*. Lecture Notes in Statistics. 1983.

[24] S. Basu, P. Pope, and S. Feizi. "Influence Functions in Deep Learning Are Fragile". In: *International Conference on Learning Representations*. 2020.

[25] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik. "Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings". In: *Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 2018.

[26] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead. "CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts". In: *Workshop on Computational Linguistics and Clinical Psychology*. 2019.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv:1907.11692* (2019).

[28] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980* (2014).

[29] E. Mohammadi, H. Amini, and L. Kosseim. "CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts". In: *Workshop on Computational Linguistics and Clinical Psychology*. 2019.