Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

An Example of (Too Much) Hyper-Parameter Tuning in Suicide Ideation Detection

Annika Marie Schoene [1,2], **John Ortega**[1,2], **Silvio Amir**[1,2,3], **Kenneth Church**[1,2,3]

[1] Northeastern University
[2] Institute for Experiential AI
[3] Khoury College of Computer Sciences

Abstract

We investigate the performance of large language models (LLMs) fine-tuned with GFT on the TWISCO suicide ideation detection benchmark. Specifically, we aim to answer the following research questions: (i) how does the size and domain of LLM's pretraining data affect the performance; and (ii) what is the impact of matching the time period of the training data to that of the pretraining data using TimeLMs? To answer these questions, we conduct a large number of experiments comparing six widely used LLMs, plus 12 checkpoints of the same LLM corresponding to different time periods. We fine-tuned each of these models with 48 combinations of the learning rate, batch size and number of epochs. We find that (i) the best performing LLM can outperform the previously established baseline by 9 % and (ii) performance of TimeLMs varies substantially across different hyper-parameter configurations. However, only 10% of these configurations outperformed a strong baseline model. The time period of the pre-training data did not have much impact on the results, which was surprising given that the language of social media tends to change rapidly over time. Despite the observed improvements over the baseline, these results raise concerns about reproducibility — only a small fraction of the tested configurations outperformed a non-LLM baseline. It is common practice in the literature to run many experiments and report only the best, but this may be risky, especially in critical or sensitive tasks such as Suicide Ideation Detection.

Introduction

Reducing suicide rates is one of the key objectives of the UN's Sustainable Development Goals for health care[1]. An estimated 25–30% of people who die by suicide leave behind a suicide note; however, this figure can be as high as 50% depending on cultural, ethnic, or demographic differences (Shioiri et al. 2005). Previous work has shown that people increasingly turn to social media to express suicidal feelings or intent (Desmet and Hoste 2013; Sueki 2015). Therefore, methods for automated detection of suicide related discourse online can be important tools in harm reduction and prevention. However, there is dearth of publicly available datasets to support NLP models for this purpose (Schoene et al. 2022). Moreover, this is a complex and multifaceted endeavor that requires input from domain experts (e.g., in psychology, public health).

LLMs have shown remarkable performance on a variety of NLP tasks and benchmarks, including in low- and mid-resource settings (Ogueji, Zhu, and Lin 2021; Micheli, d'Hoffschmidt, and Fleuret 2020). The widespread availability of LLMs via model zoos such as Huggingface[2], has reduced the barriers in the access to state-of-the-art models for NLP researchers and practitioners. Off-the-shelf tools with model zoo integration, such as GFT (Church et al. 2022) promise to expand access to non-technical people. However, the performance of fine-tuned models can vary substantially as a function of the choice of hyper-parameters (Dodge et al. 2020; D'Amour et al. 2020; Amir, van de Meent, and Wallace 2021), and even the time period of pretraining data. For example, Loureiro et al. (2022) observed differences in the performance of LLMs trained on Twitter data from different time periods on a benchmark of social media analysis tasks. However, it is not clear whether similar differences are observed in the task of suicidal ideation detection.

In this paper, we examine the difficulty of choosing the appropriate LLM and set of hyper-parameters for online suicide discourse detection using a small dataset with fine-grained manual annotations due to (Schoene et al. 2022). We conduct a large number of experiments to determine which factors have the greatest impact in real-world deployments. Specifically, we analyze the effects of the LLM's pretraining data (in terms of source, domain, size and time-period) and the hyper-parameters used for fine-tuning (i.e. batch size, number of epochs and learning rate). Our experiments show that: (i) fine-tuned LLMs can outperform previously proposed models for the task of suicide ideation detection; and (ii) finding the optimal configuration is not trivial since only 10% of our runs outperformed the baseline (Figure 3). These results suggest that practical applications of fine-tuned NLP models still requires technical expertise and insight to understand the model's behaviors. More broadly, our results raise concerns about reproducibility, as it is common practice in the literature to run many experiments and report the best, but doing so may be risky, especially important given the sensitive nature of Suicide Ideation Detection.

[1] https://sdgs.un.org/goals

[2] http://huggingface.co/

Related Work

Suicide Ideation Detection on Social Media Our work is closely related to previous efforts in methods for suicide ideation detection and suicide-related content classification. Collecting annotated data for mental health related tasks is notoriously difficult and suicidal ideation is no exception. Work in this area usually relies on self-reports (Coppersmith et al. 2015), heuristics based on the presence of specific keywords (Du et al. 2018) and phrases (Burnap et al. 2017). While these methods allow for the collection of large datasets, they can also produce a large number of false positives, e.g. where suicide related phrases are used for humor (e.g. *the movie is so bad I want to kill myself*), to discuss awareness/prevention campaigns, or state facts about the issue. To overcome these limitations others have used human annotation to obtain more fine-grained labels, e.g. on risk-levels (O'dea et al. 2015), distinctions between worrying language and flippant references to suicide (Burnap et al. 2017), or content and affect of suicide related posts (Schoene et al. 2022). Several methods have been proposed to detect suicide intent and ideation, including feature based models with combinations of lexical features (Coppersmith et al. 2015), and psychological and affective features (Burnap et al. 2017). Others have explored deep learning models based on Convolutional Neural Networks (Du et al. 2018), Recurrent Neural Networks (Ji et al. 2018), and Graph Neural Networks (Mishra et al. 2019; Sawhney et al. 2021).

Fine-tuning pretrained Language Models Recent years have seen a rise in the development of pretrained large language models (LLMs) that can induce contextualized representations, such as BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020). These models can be fine-tuned with small task-specific datasets and achieve SOTA results on various NLP tasks while being less computationally less expensive. However, fine-tuning pretrained Transformer-based LLMs can be a brittle process (Phang, Févry, and Bowman 2018) that depends on the choice of adequate hyperparameters (Dodge et al. 2020), and is subject to optimization instability when used with small datasets (Mosbach, Andriushchenko, and Klakow 2020).

Language Change over time There is a rich literature in historical linguistics and the sociolinguistics of language evolution (Labov 2011; Milroy 1992). It has been observed that the performance of LLMs may degrade over time particularly in social media environments where there is a rapid change in linguistic patterns and themes of discussion (Jaidka, Chhaya, and Ungar 2018). There have been attempts to capture such changes in LLMs with dynamic contextualized representations (Hofmann, Pierrehumbert, and Schütze 2021), and by continuously updating models with new data (Loureiro et al. 2022).

Experimental Setting

Corpus We use the corpus introduced by Schoene et al. (2022), which contains 3,977 Tweets related to suicide with annotations by three experts who hold PhDs in psychology. The task is to predict content labels (see Table 1) that go beyond binary expressions of suicide. We investigate the difficulty of fine-tuning LLMs for fine-grained detection of suicide related content on social media for non-technical domain experts. Specifically, we aim to answer the two following research questions:

Content Label	Frequency
Facts about suicidality	131
Suicide discussed philosophically/religiously	309
Contacts for suicide-related help-seeking	51
News report, case studies or stories	291
Humorous use	165
Content not relevant	2,497
Expressing own suicidality	443
Expressing worries about suicidality of others	90
Total	3,977

Table 1: Description of TWISCO labels

- RQ1: *How does the size and domain of LLM's pretraining data affect the performance of fine-tuned classifiers?*
- RQ2: *What is the impact of matching the time period of the training data to that of the pretraining data of the LLM?*

Hyper-parameter fine-tuning To answer these questions, we conduct a large number of fine-tuning experiments on the TWISCO corpus. We compare the performance of several pretrained LLMs across a wide range of hyper-parameter configurations against a strong baseline model based on a Graph Convolutional Neural Network from the original TWISCO paper (Schoene et al. 2022). For each LLM, we evaluate 48 combinations:

- **Epochs:** 10, 25, 50 and 100
- **Batch size:** 1, 16, 32 and 64.
- **Learning Rate:** 0.001, 0.0001 and 0.00001.

We use GFT (Church and Kordoni 2022), a framework that enables easy and fast implementation of fine-tuned LLMs from Huggingface (Wolf et al. 2019), and conducted the experiments on 4 NVIDIA P100 GPUs. The corpus was split into 80% training, 10% validation and 10% test sets and we fix the random seed for reproducibility. To understand the impact of training data size, we have computed the approximate size of each LLM's training data as shown in Table 2, where we assume 100 bytes per sentence and 280 bytes for Tweets on average.

RQ1 We chose a variety of LLMs available on Huggingface based on (i) domain (e.g.: mental health), (ii) data source (e.g.: Twitter) and (iii) overall size of the training data.

- **RoBERTa**: a variant of the original BERT model that was trained for longer, without the next sentence prediction task, on 2 English corpora, Wikipedia and the BookCorpus (Liu et al. 2019) [3]

[3] https://huggingface.co/roberta-base

RQ1	Size (GB)	Epochs	BS	LR	F1
Baseline	0.39	50	128	0.001	0.62
MentalBERT	28.45	10	64	0.00001	0.62
RoBERTa	160	10	64	0.00001	0.69
MentalRoBERTa	163.82	20	64	0.00001	0.67
Twitter RoBERTa	**166.8**	**10**	**64**	**0.00001**	**0.71**
BioClinicalBERT	260.63	10	1	0.00001	0.48
XLM RoBERTa	2,500	25	16	0.00001	0.65
RQ2	**Size (GB)**	**Epochs**	**BS**	**LR**	**F1**
2019	25.27	10	16	0.0001	0.68
2020/03	26.44	25	32	0.0001	0.68
2020/06	27.62	100	16	0.00001	0.65
2020/09	28.80	10	32	0.0001	0.66
2020/12	29.97	100	16	0.00001	0.63
2021/03	31.15	25	64	0.0001	0.65
2021/06	32.32	100	16	0.00001	0.66
2021/09	**33.5**	**10**	**64**	**0.0001**	**0.69**
2021/12	34.68	10	32	0.0001	0.63
2022/03	35.85	100	32	0.0001	0.67
2022/06	37.03	100	16	0.0001	0.66
2022/09	47.28	10	64	0.0001	0.65

Table 2: Highest F1 scores for each LLM and the best hyper-parameter settings. The size column refers to the size of the pretraining data.

- **Twitter RoBERTa**: RoBERTa variant trained on around 58 million Tweets, developed for the Tweet Evaluation benchmark (Barbieri et al. 2020) [4]
- **XLM RoBERTa**: multilingual RoBERTa variant trained on 100 different languages (Conneau et al. 2020) [5]
- **MentalBERT and MentalRoBERTa**: BERT and RoBERTa variants trained on mental health related Reddit posts (Ji et al. 2022) [6][7]
- **BioClinicalBERT**: BERT based LLM trained on MIMIC, a database for electronic health records (Alsentzer et al. 2019) [8]

RQ2 We use 12 temporal checkpoints of the Time-aware tweet LLMs provided by Loureiro et al. (2022). The first LLM was trained on tweets between 2018 and 2019, and subsequent LLMs were retrained and updated in 3 month increments (December 2022 is not available at the time of writing).

Results

We conducted 288 and 576 experiments for RQ1 and RQ2, respectively, and present the main results in Table 2. For RQ1, Twitter RoBERTa improves performance over the baseline by 9%, showing that LLMs can significantly improve performance over a strong baseline. The results indicate that the size, source, and domain (e.g.: Twitter and

[4]https://huggingface.co/cardiffnlp/twitter-roberta-base
[5]https://huggingface.co/xlm-roberta-base
[6]https://huggingface.co/mental/mental-bert-base-uncased
[7]https://huggingface.co/mental/mental-roberta-base
[8]https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

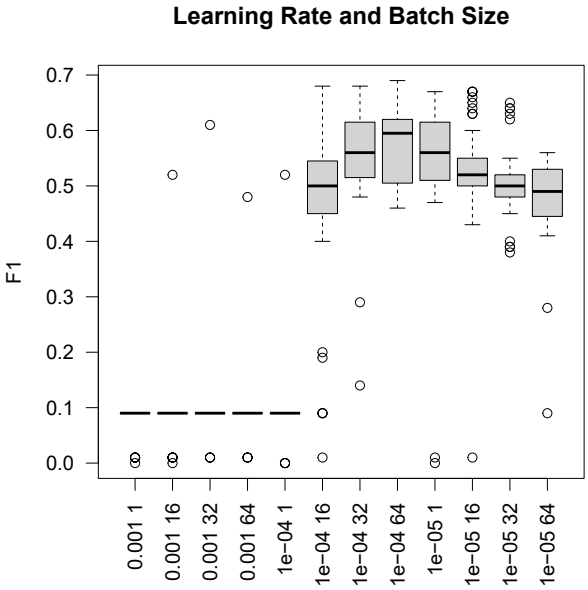

Figure 1: F1 results for RQ2 depend on learning rate and batch size. Results tend to be poor when learning rate is too large (0.001) or batch size is too small (1), though there are exceptions.

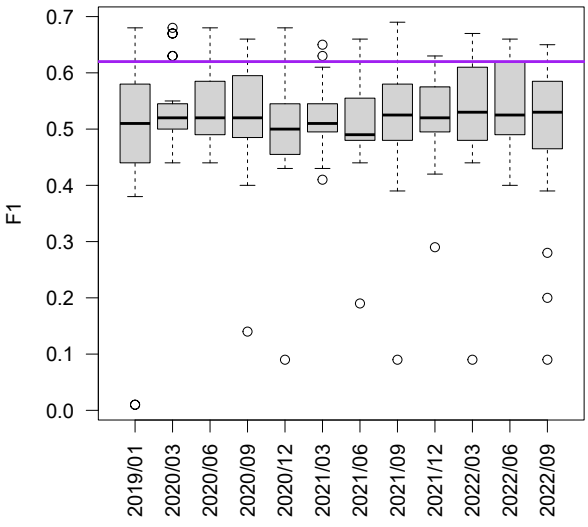

Figure 2: Results for TimeLLMs by time period (RQ2), excluding poor hyper-parameter settings (learning rate of 0.001 and batch size of 1). Only 10% of the experiments perform better than baseline (*purple line*).

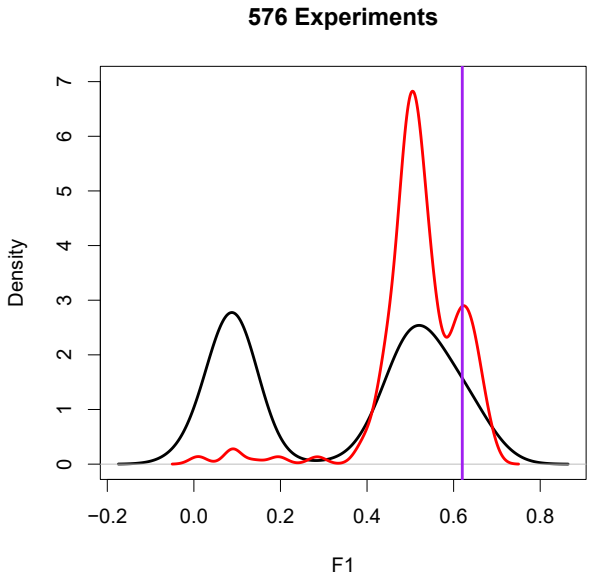

Figure 3: Density plots of F1 from all RQ2 experiments (*black curve*). The *red curve* excludes poor hyper-parameters settings (i.e., learning rate = 0.001 and batch size = 1). Only 10% of the experiments match or improve the TWISCO baseline (*purple line*).

mental health) of the training data seem to have more impact than size and domain alone. Experiments for RQ2 show that the performance also varies according to the time of the pretraining data. Even though there is no clear pattern, there is almost always a setting that beats the TWISCO baseline (0.62). That said, we are concerned about the standard practice of running a large number of experiments and selecting the best for publication. Figure 1, shows all combinations of learning rates and batch sizes for all the models. It is clear that the first 5 combinations almost always lead to poor results. While the remaining combinations are better, they do not consistently beat the baseline (F1 = 0.62). Figure 3 shows these results in a different way: of the 576 experiments (black line), only 10% are better than baseline (purple line). We can see that the average performance is bimodal, with the first mode associated with poor settings of learning rate and batch size. If exclude the poor settings (red line) most of the results remain below the baseline, indicating that excluding poor settings is not sufficient to ensure improvements over baseline, as shown in Figure 2.

Conclusions

In this work, we have shown that hyper-parameter fine-tuning is important to achieve competitive results in downstream tasks. The best combination improves over baseline by 9%. We find that some hyper-parameters are more important than others. Optimal settings of batch size and learning rate are much better than poor settings. Differences over time language models (TimeLMs) are smaller and less conclusive. That said, we are concerned about running too many experiments and wearing out the benchmarks. It is standard practice to run many experiments and report the best results. We find that while there are some settings of the hyper-parameters that beat the baseline, most settings do not. In this way, the standard practice can be misleading and only paint part of the picture. This is especially important to acknowledge given the domain they are used in, where suicide ideation and content detection can have serious real-world consequences. Whiles this work has given some insight into which hyper-parameters are important to achieving competitive results in suicide-related content detection, it has also highlighted the instability of F1 scores. We also acknowledge that our work is limited in that it does not explore how well our results would: (i) extrapolate to other suicide or mental health corpora, (ii) scale to larger datasets and (iii) vary with respect to choice of random seeds.

References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78.

Amir, S.; van de Meent, J.-W.; and Wallace, B. C. 2021. On the Impact of Random Seeds on the Fairness of Clinical Classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3808–3823.

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Burnap, P.; Colombo, G.; Amery, R.; Hodorog, A.; and Scourfield, J. 2017. Multi-class machine classification of suicide-related communication on Twitter. *Online social networks and media*, 2: 32–44.

Church, K.; Cai, X.; Ying, Y.; Chen, Z.; Xun, G.; and Bian, Y. 2022. Emerging trends: General fine-tuning (gft). *Natural Language Engineering*, 28(4): 519–535.

Church, K. W.; and Kordoni, V. 2022. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2): 249–269.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

Coppersmith, G.; Leary, R.; Whyne, E.; and Wood, T. 2015. Quantifying suicidal ideation via language usage on social media. In *Joint statistics meetings proceedings, statistical computing section, JSM*, volume 110.

Desmet, B.; and Hoste, V. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16): 6351–6358.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; and Smith, N. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Du, J.; Zhang, Y.; Luo, J.; Jia, Y.; Wei, Q.; Tao, C.; and Xu, H. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 18(2): 77–87.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*.

Hofmann, V.; Pierrehumbert, J.; and Schütze, H. 2021. Dynamic Contextualized Word Embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6970–6984.

Jaidka, K.; Chhaya, N.; and Ungar, L. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 195–200.

Ji, S.; Yu, C. P.; Fung, S.-f.; Pan, S.; and Long, G. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Labov, W. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loureiro, D.; Barbieri, F.; Neves, L.; Anke, L. E.; and Camacho-Collados, J. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260.

Micheli, V.; d'Hoffschmidt, M.; and Fleuret, F. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7853–7858.

Milroy, J. 1992. *Linguistic variation and change: On the historical sociolinguistics of English*. B. Blackwell.

Mishra, R.; Sinha, P. P.; Sawhney, R.; Mahata, D.; Mathur, P.; and Shah, R. R. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop*, 147–156.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

O'dea, B.; Wan, S.; Batterham, P. J.; Calear, A. L.; Paris, C.; and Christensen, H. 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2(2): 183–188.

Ogueji, K.; Zhu, Y.; and Lin, J. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 116–126. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Phang, J.; Févry, T.; and Bowman, S. R. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Sawhney, R.; Joshi, H.; Gandhi, S.; and Shah, R. R. 2021. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 22–30.

Schoene, A. M.; Bojanic, L.; Nghiem, M.-Q.; Hunt, I. M.; and Ananiadou, S. 2022. Classifying suicide-related content and emotions on Twitter using Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, (01): 1–12.

Shioiri, T.; Nishimura, A.; Akazawa, K.; Abe, R.; Nushida, H.; Ueno, Y.; KOJIKA-MARUYAMA, M.; and Someya, T. 2005. Incidence of note-leaving remains constant despite increasing suicide rates. *Psychiatry and Clinical Neurosciences*, 59(2): 226–228.

Sueki, H. 2015. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. *Journal of affective disorders*, 170: 155–160.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.