Towards Preemptive Detection of Depression and Anxiety in Twitter

David Owen **Jose Camacho-Collados** **Luis Espinosa-Anke**
School of Computer Science and Informatics
Cardiff University, United Kingdom
{owendw1,camachocolladosj,espinosa-ankel}@cardiff.ac.uk

Abstract

Depression and anxiety are psychiatric disorders that are observed in many areas of everyday life. For example, these disorders manifest themselves somewhat frequently in texts written by non-diagnosed users in social media. However, detecting users with these conditions is not a straightforward task as they may not explicitly talk about their mental state, and if they do, contextual cues such as immediacy must be taken into account. When available, linguistic flags pointing to probable anxiety or depression could be used by medical experts to write better guidelines and treatments. In this paper, we develop a dataset designed to foster research in depression and anxiety detection in Twitter, framing the detection task as a binary tweet classification problem. We then apply state-of-the-art classification models to this dataset, providing a competitive set of baselines alongside qualitative error analysis. Our results show that language models perform reasonably well, and better than more traditional baselines. Nonetheless, there is clear room for improvement, particularly with unbalanced training sets and in cases where seemingly obvious linguistic cues (keywords) are used counter-intuitively.

1 Introduction

Mental illnesses are psychiatric disorders that may cause sufferers significant distress and impair their ability to function in social and work activities (Bolton, 2008). The most prevalent mental illnesses are depression and anxiety, which are estimated to affect nearly one in ten people worldwide (676 million) according to a recent study (World Health Organization, 2016). While depression and anxiety are different disorders, they also share symptoms and, thus, clinicians often diagnose patients with both illnesses at consultation.[1] Identifying these conditions at early stages is relevant not only because of their inherent importance, but also because they are precursors to major related concerns in public health including self-harm (Centers for Disease Control and Prevention, 2015), making timely diagnosis and treatment even more essential. However, sufferers of depression and anxiety can find that it takes great courage and strength to seek professional treatment (Dennis C Miller, 2016). They may also be afraid to confide in their peers due to the stigma of mental illness (Wasserman et al., 2012).

With reluctance to seek professional treatment or rely on their peers, sufferers often turn to online resources for support. These include both specialised and general communities, with Twitter and Reddit (Yates et al., 2017) being paramount examples of the latter. Because of this, systems that can automatically detect and flag such cases at a large-scale are highly desirable (Guntuku et al., 2017). They may enable prompt analysis and treatment, which is crucial in the early development of such conditions. Moreover, the interpersonal and economic effects of these illnesses may be mitigated with prompt intervention (Lexis et al., 2011).

In this paper, we build a classification dataset[2] to assist in the detection of depression and anxiety in Twitter, and compare several text classification baselines. The results show that state-of-the-art language models (LMs henceforth) like BERT (Devlin et al., 2019) unsurprisingly outperform competing baselines. However, when the dataset shows an unbalanced distribution, linear models perform on par.

[1] https://www.bupa.co.uk/newsroom/ourviews/2017/10/anxiety-depression

[2] The datasets and code used in our experiments are available online at the following repository: https://bitbucket.org/nlpcardiff/preemptive-depression-anxiety-twitter.

Finally, alongside quantitative results, we also provide a qualitative analysis through which we aim to better understand the strengths and limitations of the models under study. Further, we identify the linguistic patterns alluding to the presence of depression and anxiety that elude all of the classifiers, and consider how we might improve performance against such patterns in the future.

2 Related Work

Baclic et al. (2020) surveyed the use of NLP in healthcare and identified the inherent opportunities and challenges. NLP permits speedy analysis of large volumes of unstructured text such as electronic patient records or social media posts, which can help support early healthcare interventions. For example, automated analysis of consumers' online reviews was used to predict the presence of depression in reviewers prior to its formal diagnosis (Harris et al., 2014). However, Baclic et al. noted that the effectiveness of NLP in the domain of mental health is constrained by a lack of high quality training data. Consequently, we chose to build our own labelled dataset.

In terms of detecting depression and anxiety in social media, the work of Yates et al. (2017) is perhaps the most related to ours. They showed that depression among Reddit participants can be detected by identifying certain lexical and psycholinguistic features in the contents of their Reddit postings. These features include indications of negative psychological processes such as anger and sadness, which may be denoted by the words "hate" and "grief" respectively (Pennebaker et al., 2015). The work of Yates et al. is therefore highly relevant to ours since it concerns automated detection of mental illness in written social media discourse. However, there are three key aspects that set our paper apart, namely: (1) We identify users of social media platforms who appear to have *not* yet been diagnosed with mental illness. We do this in the spirit of providing the opportunity for early healthcare intervention in these cases; (2) we consider automated classification using feature rich representations of users' online postings using state-of-the-art NLP methods such as word embeddings and pre-trained LMs; and (3) we consider concise written discourse in the form of tweets, rather than Reddit postings, which are generally more verbose.

3 Dataset Construction

In this section, we describe our process for building a dataset for detecting depression and anxiety in Twitter. We describe the tweet collection procedure (Section 3.1), annotation (Section 3.2), and provide information about the inter-annotator agreement (Section 3.3).

3.1 Tweet collection

First, we used Twitter's Stream API to compile a large corpus of tweets. All tweets were of English language and published between May 2018 and August 2019.[3] We only considered tweets containing at least three tokens and without URLs so as to avoid bot tweets and spam advertising. All personal information, including usernames (denoting the author or other users) and location were removed from the corpus - only textual information was retained. We did however retain emojis and emoticons. We surmised that they may, in part at least, be indicative of depression or anxiety.

The corpus was then filtered. We aimed to identify tweets whose authors may be suffering from depression or anxiety but may not yet have been diagnosed by a clinician. To achieve this, we sought tweets containing occurrences of *depress*, *anxie*, or *anxio*, but not *diagnos*[4] - an approach similar to that used by Bathina et al. (2020). This produced an initial set of 89,192 tweets. From these tweets we proceeded to annotate a random subset of 1,050 tweets to arrive at our dataset.

3.2 Annotation

Three human annotators were appointed. The prerequisites for these annotators were to be fluent in English and to have familiarity with Twitter. The 1,050 tweet dataset was divided into three distinct subsets of 300 and one distinct shared subset of 150. Each annotator received one of the former subsets in addition to the latter subset. They were tasked with labelling the 450 tweets that they had received.

One of two numerals was selected by the annotator with respect to each tweet:

[3] Twitter's automatic language labelling was used to identify English tweets.

[4] e.g. "My anxiety is terrible today"

1: The tweeter appears to be suffering from depression or anxiety.

0: The tweeter does not appear to be suffering from depression or anxiety.

Guidelines were compiled to aid the annotation exercise. Their purpose was to ensure a consistent approach amongst annotators and to resolve ambiguous cases. These guidelines are defined below along with examples and their suggested labels:

1. The tweeter states that they have depression or anxiety
   Example: *"I feel sick to my stomach, I hate having such bad anxiety"* - **1**

2. The tweeter states that they have had depression or anxiety in the past
   Example: *"Counselling fixed my depression"* - **0**

3. The tweeter is referring to a fellow tweeter who may have depression or anxiety
   Example: *"@user I wish you all the best in beating your anxiety"* - **0**

4. The tweeter is temporarily depressed or anxious due to a short or superfluous event
   Example: *"Nothing gives me anxiety more than the tills at Aldi"* - **0**

5. The tweet is ambiguous or does not provide definitive information
   Example: *"Depression is not taken serious enough"* - **1**

Guideline 5 recommends positive labelling in ambiguous cases. This is to help achieve high recall in terms of tweeters who appear to be suffering from depression or anxiety. Whilst this approach will inevitably retrieve negative instances, an eventual real-world application would require all retrieved instances to be verified manually by medical experts, and therefore high recall at the expense of lower precision is an acceptable tradeoff. In fact, it is recommended that results from automatic classifiers used in healthcare settings should be verified via an "expert-in-the-loop approach" (Holzinger, 2016).

The guidelines evolved following the annotators' first attempts at the exercise. A conflict resolution meeting revealed that while agreement was due to be acceptable, there were instances where annotators felt unable to assign either label. This gave rise to the addition of guideline 5, which allowed the annotators to complete the exercise with confidence.

3.3 Inter-Annotator Agreement

Once the annotation was completed, we calculated the Average Pairwise Percentage Agreement of the three annotators with respect to the 150 common tweets that they had received (Table 1).

Annotators 1 and 2	Annotators 1 and 3	Annotators 2 and 3	Average Pairwise Agreement	Fleiss' Kappa	Krippendorff's Alpha
78.67	80.67	80.67	80.00	0.60	0.60

Table 1: Pairwise Percentage Agreements and Inter-Annotator reliability.

An average pairwise agreement of 80% was recorded. To validate the quality of the exercise two further measures of inter-annotator reliability were selected: Fleiss' Kappa and Krippendorf's Alpha. They are apt for inter-annotator exercises involving more than two annotators (Zapf et al., 2016). Both measures returned scores indicating "substantial agreement" amongst the annotators (Xie et al., 2017).

Confidence in the annotation guidelines was therefore established versus the 150 tweets common to each annotator. Disagreements in the labels were decided by majority voting among the three annotators. For example, the tweet *"My seasonal depression automatically begun tonight at 12am"* was labelled **1** by two of the annotators and labelled **0** by the third annotator. However, majority voting meant that it was finally labelled **1**. The annotators then proceeded to label their distinct 300 tweet subsets independently.

4 Experimental evaluation

In the following we detail the experimental setting (Section 4.1) and then present the results (Section 4.2) and an analysis (Section 4.3).

4.1 Experimental setting

4.1.1 Data

We prepared our annotated dataset described in Section 3 for input to a series of supervised classifiers. The three distinct subsets of 300 annotated tweets were combined to form a training set of 900 tweets. The 150 tweets labelled by all annotators formed the test set. We named this dataset *DATD* (Depression and Anxiety in Twitter Dataset). The test set's ratio of positive instances to negative instances was exactly 1:1 following the annotation exercise. This contrasts with related published datasets upon which no annotation had been performed and all instances were deemed mental illness-related.[5]

Following a similar approach to Bathina et al. (2020), we also compiled a non-annotated set of 3,600 random tweets which did not contain any occurrence of *depress*, *anxie*, *anxio*, or *diagnos*. These were merged with the 900 tweet training set to form a larger training set of 4,500 tweets. The purpose of this large training set (DATD+Rand henceforth) was to recreate a more realistic (and noisy) setting where most training instances are negative. This meant that only 10.5% of the instances in this training set contained any of the keywords used to compile the positive examples. The 150 tweets labelled by all annotators formed the test set once again. The main characteristics of the two datasets are summarised in Table 2.

	Training		**Test**
	DATD	DATD+Rand	DATD
Positive Instances	473	473	75
Negative Instances	427	4,027	75
Total Instances	900	4,500	150

Table 2: Characteristics of the datasets used in the evaluation.

4.1.2 Comparison systems

We evaluated several binary classifiers on both the DATD and DATD+Rand datasets, guided by existing research concerning problems similar to the one at hand. To this end, we deemed a Support Vector Machine (SVM) and an LM to be suitable classifiers. SVMs have demonstrated effectiveness when used with Twitter datasets in healthcare contexts (Prieto et al., 2014; Han et al., 2020). For our experiments we used both a standard SVM classifier with TF-IDF features and a classifier based on the average of word embeddings within the tweet.

With regards to pre-trained LMs, we used BERT (Bidirectional Encoder Representations from Transformers) and ALBERT (A Lite BERT) (Lan et al., 2019). These LMs have been deployed effectively in NLP tasks, leading to state-of-the-art results in most standard benchmarks (Wang et al., 2019) including Twitter (Basile et al., 2019; Roitero et al., 2020). In particular, ALBERT has been shown to provide competitive results despite being relatively light-weight compared to other LMs.

Finally, for completeness we added a naïve baseline that predicts positive instances in all cases.

4.1.3 Training details

We used the scikit-learn SVM model (Pedregosa et al., 2011) as well as its TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer implementations.[6] The word embeddings generated for each tweet were drawn from vectors trained on Twitter data (Pennington et al., 2014, GloVe). These vectors had a dimensionality of 200, and so did the averaged embedding generated.

We performed tweet text preprocessing prior to their input to the SVM. In one series of SVM experiments all tweets underwent tokenization and lowercasing only, but in a second series all tweets also underwent tweet specific preprocessing[7] (SVM+preproc henceforth). The preprocessing entailed the removal of hashtags, user mentions, reserved words (such as "RT" and "FAV"), emojis, and smileys. This

[5] https://github.com/AshwanthRamji/Depression-Sentiment-Analysis-with-Twitter-Data/blob/master/tweetdata.txt

[6] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[7] https://pypi.org/project/tweet-preprocessor/

Classifier	Features	Accuracy	Precision	Recall	F1
SVM	TF-IDF	0.633	0.619	0.693	0.654
	Word Embs	0.727	0.693	0.813	0.748
	TF-IDF + Word Embs	0.733	0.711	0.787	0.747
SVM+preproc	TF-IDF	0.633	0.616	0.707	0.658
	Word Embs	0.713	0.695	0.760	0.726
	TF-IDF + Word Embs	0.727	0.698	0.800	0.745
BERT	LM	**0.749**	**0.713**	**0.856**	**0.774**
ALBERT	LM	0.675	0.651	0.779	0.705
Naïve baseline	-	0.500	0.500	1.000	0.670

Table 3: Results of the first experimental setup (DATD).

enabled us to see how the presence of these common tweet features affected classification performance. In both cases, the SVM used a linear kernel and default hyperparameters.

To deploy the LM classifiers we used the Simple Transformers[8] software library. It provides a convenient Application Programming Interface (API) to the Transformers Library, which itself provides access to BERT and ALBERT models, amongst others (Wolf et al., 2019). The BERT and ALBERT classifiers used were "bert-base-uncased" and "albert-base-v1",[9] respectively.

Unlike the SVM experiments it was not necessary to tokenize tweet texts prior to their input to the BERT or ALBERT classifiers; they perform their own tokenization. Tweet texts did undergo prior lowercasing however. The classifiers were instantiated with Simple Transformers' default hyperparameters.

4.2 Results

Experimental results are presented in Table 3 (DATD) and Table 4 (DATD+Rand).[10]

In the first setting, BERT achieves the best overall results, which is not unexpected. More importantly, the overall accuracy (i.e. 0.749) is close to the pairwise IAA (i.e. 0.800), which suggests that BERT is able to follow the guidelines provided in Section 3.2 to a reasonable extent. As for the linear models, the Twitter-specific preprocessing for the SVM does not lead to any improvements. In the following analysis section, we aim at shedding light on the types of error made by these models.

In the second setting where random tweets are added as negative instances in the training sets (Table 4), SVM with word embeddings features perform similarly to BERT, being in fact slightly better overall. This result is perhaps surprising, but may be due to the relative robustness of SVMs with respect to unbalanced training sets, which seem to have a greater effect on the LMs. Another explanation may be that by concatenating TF-IDF features and word embeddings the classifier is effectively leveraging both global and local dependencies, which have been shown to be crucial in tweet classification tasks such as emoji prediction (Barbieri et al., 2018) and stance detection (Mohammad et al., 2016).

More generally, the results in this setting are not hugely different from the first setting's. This is encouraging, as it suggests that supervised models can also perform in a more realistic setting where the negative instances are more prevalent than the positive ones.

4.3 Analysis

Perhaps the main highlights of our experiments are the results obtained by BERT and the concatenation of TF-IDF features and word embeddings in SVMs. BERT performs remarkably well despite not being trained on Twitter data. This could suggest that, although slang, jargon, misspellings, and emoji are typical in microposts, users suffering from mental illness are more articulate in their online writing than their mentally healthy counterparts. Thus their writing style is more likely to be picked up by an LM with restricted vocabularies.[11] Another surprising set of results concerns ALBERT, which was trained on the

[8] https://github.com/ThilinaRajapakse/simpletransformers

[9] https://huggingface.co/transformers/pretrained_models.html

[10] Results for the LMs BERT and ALBERT were reported for the average of five different runs.

[11] While this is overcome within BERT by the use of WordPiece (Wu et al., 2016), the quality of its internal representations degrade as the number of OOV words it has to deal with increases.

Classifier	Features	Accuracy	Precision	Recall	F1
SVM	TF-IDF	0.673	0.681	0.653	0.667
	Word Embs	0.740	0.737	0.747	0.742
	TF-IDF + Word Embs	**0.747**	**0.740**	0.760	**0.750**
SVM+preproc	TF-IDF	0.660	0.658	0.667	0.662
	Word Embs	0.720	0.704	0.760	0.731
	TF-IDF + Word Embs	0.740	0.725	0.773	0.748
BERT	LM	0.693	0.656	0.851	0.737
ALBERT	LM	0.648	0.609	**0.880**	0.715
Naïve baseline	-	0.500	0.500	1.000	0.670

Table 4: Results of the second experimental setup (DATD+Rand).

same corpus as BERT. We could expect that given the modest size of this dataset, an overparameterized model like BERT could fall short when compared to lighter versions, but this does not seem to be the case. In any case, ALBERT achieves the highest recall score on the DATD+Rand dataset.

Let us now highlight a selection of tweets in the DATD dataset and the performance of the classifier configurations against them. For example, the tweet *"got a yellow phone case hoping it will cure my depression"* was labelled **1**, a label only predicted by two of the eight configurations, namely BERT and SVM with TF-IDF + Word Embs features. This tweet is a good example of the overarching complexity of the problem, the presence of prosaic terms like "phone case", or positive words like "hope" or "cure" may have confused the simpler word-based models.

Another illustrative example is *"you know that i'm the best, is that why you depressed?"*, which was labelled **0** and was misclassified by all configurations. We hypothesize that this may be due to having three instances of two distinct pronouns ("I" and "you"), which are likely used often by depressed or anxious tweeters, although it was not the case in this particular example.

There are other interesting examples, including cases where only the LMs (BERT and ALBERT) made correct predictions. For example, *"Got my first call center job and my anxiety is through the roof"*, which was labelled **1**, and *"Hot shot screaming gives me anxiety watching these game lol mans be stressed but these games good as hell learning a lot."*, which was labelled **0**. Conversely, there are also cases where only SVMs made correct predictions. For example *"big ass spider in my room and it disappeared so I'll just have anxiety for the rest of the night I'm in here"*, which was labelled **0**. While it is not clear whether there is a systematic pattern to draw conclusions from, it does seem that when only the LMs succeed there is some degree of world or semantic understanding required to capture the condition of the tweeter (for example, the fact that you are probably not actually anxious by watching a game).

5 Conclusion and Future Work

In this paper, we have presented an experimental evaluation for detecting depression and anxiety in social media. We have developed a dataset, DATD, for predicting depression and anxiety in Twitter. Using this dataset we have run a comparative analysis of pre-trained LMs and traditional linear models. Not surprisingly, LMs performed relatively well on this task with a balanced set, but they do not outperform lighter-weight methods when the training data is unbalanced. Given the relatively small size of the dataset, we have also performed a qualitative analysis to identify areas for improvement.

Since these automatic models are intended for use by medical experts future work could involve collaboration with them. Moreover, classifier performance must be measured and certified versus large, heterogeneous datasets before adoption in healthcare is likely to be considered (Kelly et al., 2019). Collaboration may serve mental health experts to better understand social media at a large-scale, and to develop better guidelines and treatments.

Finally, we are also planning to extend this work to develop a dataset with finer grained distinctions, similar to Bathina et al. (2020) for Cognitive Distortion Schemas.

References

Oliver Baclic, Matthew Tunis, Kelsey Young, Coraline Doan, Howard Swerdfeger, and Justin Schonfeld. 2020. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*, 46(6):161–168.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Krishna C. Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2020. Depressed individuals express more distorted thinking on social media. *arXiv preprint arXiv:2002.02800*.

Derek Bolton. 2008. *What is mental disorder? : an essay in philosophy, science, and values*. International perspectives in philosophy and psychiatry. Oxford University Press, Oxford ; New York.

Centers for Disease Control and Prevention. 2015. Suicide: Facts at a glance [fact sheet].

Dennis C Miller. 2016. Mental health awareness month: Take the first step towards a mentally healthy workplace.

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.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Kai-Xu Han, Wei Chien, Chien-Ching Chiu, and Yu-Ting Cheng. 2020. Application of support vector machine (svm) in the sentiment analysis of twitter dataset. *Applied Sciences*, 10(3):1125.

Jenine K Harris, Raed Mansour, Bechara Choucair, Joe Olson, Cory Nissen, and Jay Bhatt. 2014. Health department use of social media to identify foodborne illness—chicago, illinois, 2013–2014. *MMWR. Morbidity and mortality weekly report*, 63(32):681.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Monique AS Lexis, Nicole WH Jansen, Marcus JH Huibers, Ludovic GPM Van Amelsvoort, Ate Berkouwer, Gladys Tjin A Ton, Piet A Van Den Brandt, and IJmert Kant. 2011. Prevention of long-term sickness absence and major depression in high-risk employees: a randomised controlled trial. *Occupational and Environmental Medicine*, 68(6):400–407.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

James W. Pennebaker, Ryan L. Boyd, Kayla N Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. In *Psychometrics manual for text analysis program LIWC2015*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Víctor M. Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. 2014. Twitter: a good place to detect health conditions. *PloS one*, 9(1):e86191.

Kevin Roitero, VDMSM Cristian Bozzato, and G Serra. 2020. Twitter goes to the doctor: Detecting medical tweets using machine learning and bert. In *Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Camilla Wasserman, Christina W Hoven, Danuta Wasserman, Vladimir Carli, Marco Sarchiapone, Susana Al-Halabi, Alan Apter, Judit Balazs, Julio Bobes, Doina Cosman, et al. 2012. Suicide prevention for youth-a mental health awareness program: lessons learned from the saving and empowering young lives in europe (seyle) intervention study. *BMC public health*, 12(1):776.

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.

World Health Organization. 2016. *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zheng Xie, Chaitanya Gadepalli, and Barry MG Cheetham. 2017. Reformulation and generalisation of the cohen and fleiss kappas. *LIFE: International Journal of Health and Life-Sciences*, 3(3).

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):93.