Depression Detection: Text Augmentation for Robustness to Label Noise in Self-reports

Dat Quoc Ngo [1,2,*] Aninda Bhattacharjee [2,*] Tannistha Maiti [2] Tarry Singh [2]
Jie Mei [2,3]
[1] University of Texas at Dallas , United States
dqn170000@utdallas.edu
[2] Deepkapha.ai, Netherlands
{tarry.singh, aninda, tannistha.maiti}@deepkapha.ai
[3] Unviersity of. Quebec at Trois-Rivieres, Canada
jie.mei@uqtr.ca
[*] Authors equally contributed

January, 2021

Abstract

With a high prevalence in both high and low-middle-income countries, depression is regarded as one of the most common mental disorders around the globe, placing heavy burdens at a societal level. Depression severely impairs the daily functioning and quality of life of individuals of different ages, and may eventually lead to self-harm and suicide. In recent years, advancements have emerged in the fields of deep learning and natural language understanding, leading to improved detection and assessment of depression using methods including convolutional neural networks (CNNs) and bidirectional encoder representation from transformers (BERT). Nevertheless, previous work focused on data acquired through brain functional magnetic resonance imaging (fMRI), clinical screening or interviews, thus required labeling by domain experts. Therefore, in this study, we used the Reddit Self-reported Depression Diagnosis dataset, an uncurated text-based dataset, to enable detection of depression using easily accessible data. To reduce the negative impact of label noise on the performance of transformers-based classification, we proposed two data augmentation approaches, i.e., Negative Embedding and Empathy for BERT and DistilBERT, to exploit the usage of pronouns and affective, depression-related words in the dataset. As a result, the use of Negative Embedding improves the accuracy of the model by 31% compared with a baseline BERT and a DistilBERT, whereas Empathy underperforms baseline methods by 21%. Taken together, we argue that the detection of depression can be performed with high accuracy on datasets with label noise using various augmentation approaches and BERT.

1 Introduction

Depression is a mental health disorder that affects more than 264 million people worldwide (James et al., 2018). Patients with depression show emotional, physical, and cognitive alterations which impair their daily functioning (Haro et al., 2019; Rao et al., 1991), and symptoms vary widely from trouble concentrating, remembering details (Kreutzer et al., 2001), making decisions to feelings of guilt, worthlessness, and helplessness, even pessimism and hopelessness. Depression can lead to suicidal thoughts and attempts (Pedersen, 2008; Toolan, 1962): According to the WHO, approximately 800,000 people die due to suicide every year. Suicide is the second leading causes of death in adolescents and young adults (Werbart Törnblom et al., 2020).

COVID-19 has a significant impact on psychological distress in health professionals and led to a public mental health crisis (Advice, 2020; Campion et al., 2020; Lu and Bouey, 2020; Pfefferbaum and North, 2020). Studies indicated that people who do not have access to sufficient supplies during the lockdown were most affected, and family affluence was found to be negatively correlated with stress, anxiety, and depression (Bandyopadhyay and Dutta, 2020; Rehman et al., 2020). COVID-19 has also negatively impacted the labour market outcomes for various professions (Conversation, 2020; Fana et al., 2020), and many of these professions are experiencing job losses, reductions in hours, wages and labour force participation (Bank, 2020). Among different professions, students and healthcare professionals were found to experience stress, anxiety, and depression more than others (Rehman et al., 2020; Alambo et al., 2020; Vizheh et al., 2020; Nelson and Kaminsky, 2020). As access to mental healthcare facilities also became limited (Moreno et al., 2020), detection of depression is of great importance during the pandemic to prevent suicide and suicide attempts (Losada et al., 2020).

With the rapidly increasing internet use, people have more opportunities to share their stories, personal challenges and mental health problems through online platforms such as Reddit or Twitter (Naslund et al., 2020; Burdisso et al., 2019). Analysis of text data provides valuable insights into the understanding and early detection of depression: For example, the more frequent use of first-person singular pronouns by depressed patients was first observed by Bucci and Freedman (1981) and Weintraub (1981), and it was then confirmed in a study that formerly- and currently-depressed subjects use the pronoun "I" and negative affective words more frequently than healthy controls (Rude et al., 2004). This observation infers that text is a possible indicator of an individual's psychological status. Using different natural language processing (NLP) techniques and machine learning algorithms, researchers have proposed novel technical approaches to the diagnosis of mental disorders (Nadeem, 2016; Paul et al., 2018; Benton et al., 2017; Coppersmith et al., 2015; Maupomé and Meurs, 2018; Resnik et al., 2015) .

In recent years, bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) has become a widely used language model that researchers extensively implemented to achieve state-of-the-art performance in various language understanding tasks. As BERT is composed of attention-based transformer blocks and pre-trained on large corpora (i.e. BookCorpus of 800M words, English Wikipedia of 2,500M words) (Zhu et al., 2015), it can capture a variety of linguistic features and contexts. Therefore, BERT can serve as a backbone to be fine-tuned for downstream tasks for higher performances. Recent research (Trotzek et al., 2018) demonstrated applications of deep neural networks in the detection and severity assessment of depression with high accuracy using social media postings. Alambo et al. (2020) applied three variants of BERT to streaming news content related to COVID-19 to assess the spatio-temporal progression of depression and drug abuse. Given, the proven performance of BERT, we used it as our baseline method in the classification of depressive and non-depressive statements.

Web-scraping with fixed labeling rules is a common approach for building large-scale text datasets for the diagnosis of depression (Cornn, 2019; Shen et al., 2017b; AlSagri and Ykhlef, 2020). Similarly, we built our Reddit Self-reported Depression Diagnosis (RSDD) dataset by web-scraping depressive and non-depressive statements from 2 subreddits, */depression* and */AskReddits*. A drawback of this method is label noise, that is, mislabeling by non-experts or oversimplified labeling criteria. Oversimplified labeling criteria may lead to mislabeling of a non-depressive statement in the */depression* subreddit as depressed (Cornn, 2019). For example, in Table 2, the sample post in the */depression* subreddit, can have both its negative title and the first comment with positive and supportive contents labeled as depressed . Due to the pattern-memorization effects, label noise may significantly compromise the performance of deep learning models in classification tasks (Zhu and Wu, 2004; Flatow and Penner, 2017), particularly in the detection of depression (Cornn, 2019). To exploit contexts for robustness to label noise in the detection of depression, we proposed two data augmentation methods, i.e., Negative Embedding and Empathy. We used the RSDD dataset to evaluate and demonstrate the performance of the two proposed methods in improving the diagnostic accuracy for the diagnosis of depression, and experimented the two augmentation methods with both BERT and DistilBERT (Sanh et al., 2019) to understand impacts of model distillation on the performance of the two augmentation methods.

2 Related Work

There are many approaches to learning under label noise. Traditionally, data cleaning has been applied which relied on finding heuristic points that

Original text: This is so frustrating. I'm sorry you're experiencing this. I know how you feel.

Tokens:	[CLS]	this	is	so	frustrat	##ing	.	[SEP]	i	'm	sorry	you	're	experienc	##ing	this	.	[SEP]	i	know	how	you	feel	.	[SEP]
Segment Embeddings:	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
Negative Embeddings:	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1: A comparison between Segment Embeddings and Negative Embeddings. In Segment Embeddings, numbers represent sentence indices. In Negative Embeddings, 1 represents negative tokens and 0 represents non-negative tokens.

were corrupted by label noise and filtering them out (Angelova et al., 2005; Brodley and Friedl, 1999). Current techniques focus on improving learning algorithms and modifying neural network architectures for estimating the true labels based on noisy labels. For example, using bootstrapping to combine multiple weak models trained on k folds of data into a strong model to learn under label noise (Algan and Ulusoy, 2020).

Large datasets in NLP suffer from noisy labels, due to erroneous automatic and human annotation procedures. Modifying deep neural networks with context modules have achieved state-of-the-art results for many image-based tasks (Elezi, 2020) with label noise. For language-based tasks, recent studies addressed the issue of label noise by supplying additional contextual information to the attention models. The attention mechanism individually computes attention weights of each token over the bag-of-word tokens (Vaswani et al., 2017). As a result, attention models such as generative pre-trained transformer (GPT) (Radford et al., 2019) and BERT (Devlin et al., 2018) neglect the contextual information in the calculation of dependencies between tokens (Yang et al., 2019a). To address this limitation, some studies modified the attention mechanism to calculate attention weights based on contextual weights (Wu and Ong, 2020; Yang et al., 2019a). Recently, Transformer-XL and XL-Net were introduced, which implemented the autoregressive pre-training for language understanding and outperformed standard attention networks in capturing contextual dependency (Dai et al., 2019; Yang et al., 2019b). In this study, we present two data augmentation techniques, i.e. Negative Embedding and Empathy to further fine tune attention networks to be noise tolerant for the detection of depression using text data. Our solutions have two advantages: First, they avoid increasing computational loads. Secondly, they leverage pre-trained weights learned from large-scale text corpora (i.e. BookCorpus of 800M words, English Wikipedia of 2,500M words).

2.1 Negative Embedding

In the conditional masked language model (MLM) pre-training task for BERT (Wu et al., 2019), the segmentation embedding was replaced by label embedding to control word predictions on conditions of labels while preserving context. Inspired by this work, we replaced segmentation embeddings with negative embeddings in order to emphasize depressive contexts on conditions of the existence of negative tokens and to estimate true labels from noisy labels. The negative embedding labels binary classes (1 and 0) for negative and non-negative tokens respectively (Figure 1). The objective is to compute the probability of depression $p(\cdot|S \setminus \Sigma n_i)$ given the negative token n_i, the sequence S and the context $S \setminus \Sigma n_i$. The negative tokens are common negative tokens in the sentiment analysis task and have been pre-defined in previous studies (Hu and Liu, 2004; Liu et al., 2005).

2.2 Empathy

An alternative approach to exploit contexts is to generate high-level lexicons which represent the overall emotional context. Researchers have relied on such high-level lexicons to identify signs of depression in social media posts and to understand the overall meaning of texts at scale. One of the most commonly used libraries is Linguistic Inquiry and Word Count (LIWC) which counts words relevant to lexical categories such as sadness, health, and positive emotions (Tausczik and Pennebaker, 2010). For example, positive lexicons include words such as happy, joy, fun, etc. In published work (Shen et al., 2017a), LIWC was used to generate lexicons as high-level text features for logistic regression models to classify depression in social media posts. LIWC has a fixed list of 40 lexical categories that limits its ability to capture signs of depression in

Original text: Wow. I understand that the rules are the rules, you just painted ”everyone” who offers that as either a psycho or a predator. I must say I am feeling like one now because ...
Lexicons: hate, nervous, suffering, art, optimism, fear, zerst, speaking, sympathy, sadness, joy, lust, shame, pain,negative_emotion, contentment, positive_emotion, depression, pronoun, ...
Post-processed text: Wow. I understand that the rules are the rules, you just pained ”everyone” who offers that as ... hate, nervous, suffering, art, optimism, fear, zest, speaking, sympathy, sadness, joy, lust, shame, pain, ...

Table 1: Example of *Empathy* generating lexicons and concatenating generated lexicons with original text.

/depression
Title: I am so tired of people taking me for granted. I give them too much of energy. I am sick of everything. my life, my family, my friends. ***Comments:*** - I'm sorry. I'm really hoping the best for you. - I know how you feel. I feel exactly the same right now. I wish I could give this post a thousand rewards.
/AskReddit
Title: What's something that impresses most people that doesn't impress you? ***Comments:*** - Limousines. As a kid, I used to think that was the sign that you made it. Now I realize you just need $95 - If you've get more than 5 people getting a limo or party bus is miles cheaper than getting multiple Ubers. Plus you can drink in them.

Table 2: Samples of the Reddit Self-reported Depression Diagnosis (RSDD) dataset for */depression* and */AskReddit*. The first comment in */depression* is a non-depressive sentence. This is an example of label noise.

text data.

Unlike LIWC, the Empath library is designed using deep learning techniques and crowdsourcing that allow it to incorporate new lexical categories (Fast et al., 2016). In the present study, the proposed data augmentation method Empathy utilizes the Empath library and initially updates the library with 2 lexicons, "pronoun" and "depression", which consider relevant words as possible indicators of depression. This process is theoretically aligned with previous findings that depressed patients use first-person singular pronouns and depression-related words more frequently than healthy controls (Rude, Gortner, and Pennebaker, 2004). Each text sample is evaluated by the Empath library to generate high-level lexicons which are then linearly concatenated with the text sample into a new text sample (Table 1). The generated text sample consists of both original contexts and high-level, extracted emotional contexts.

3 Reddit Self-reported Depression Diagnosis (RSDD) Dataset

Currently, there is no publicly available, large-scale text dataset for the diagnosis of depression. Hence, we utilized the Python Reddit API Wrapper to web-scrape posts from January 2018 to November 2020 in the 2 subreddits */depression* and */AskReddit*, which correspond to "depressed" and "non-depressed" classes, respectively. For each post, its title and comments were web-scraped, anonymized, and treated as separate samples. The total number of text samples[1] was 229,729 with 102,102 depressive text samples and 127,627 nondepressive text samples. For each class, text samples were split by a ratio of 80:20 for training and validation sets. Table 2 shows one example of the dataset that belongs to the category ”depressed” and its comments. An obvious challenge in this web-scraped dataset is the label noise: There was a high number of positive and supportive statements (e.g. ”I 'm sorry. I'm really hopping the best for you” in Table 2) in the */depression* subreddit that was labelled as "depressed" during the automatic web-scraping process.

Based on the word cloud shown in Figure 2, words such as feel, depression, want, and friend

[1]Dataset is included for submission to ACL-IJCNLP 2021.

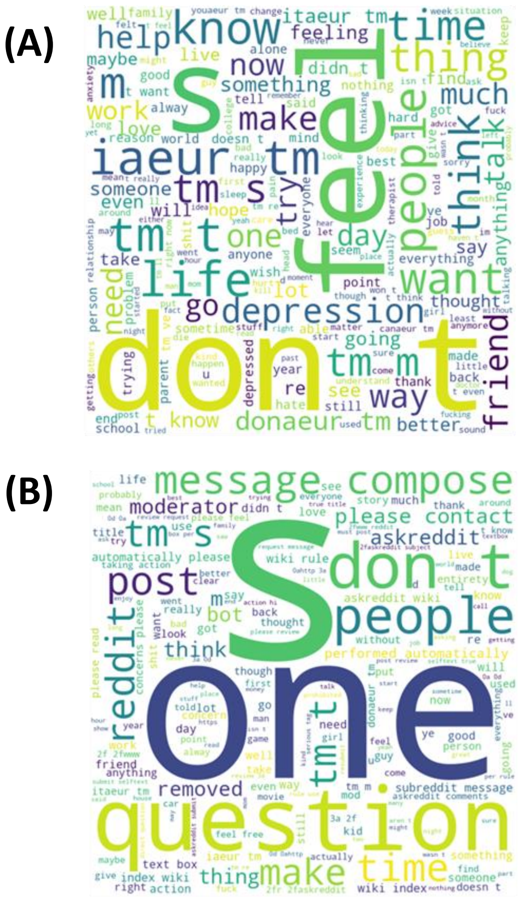

Figure 2: Word cloud demonstrating the frequency distribution of words in (A) depressive samples of the RSDD dataset, (B) non-depressive samples of the RSDD dataset.

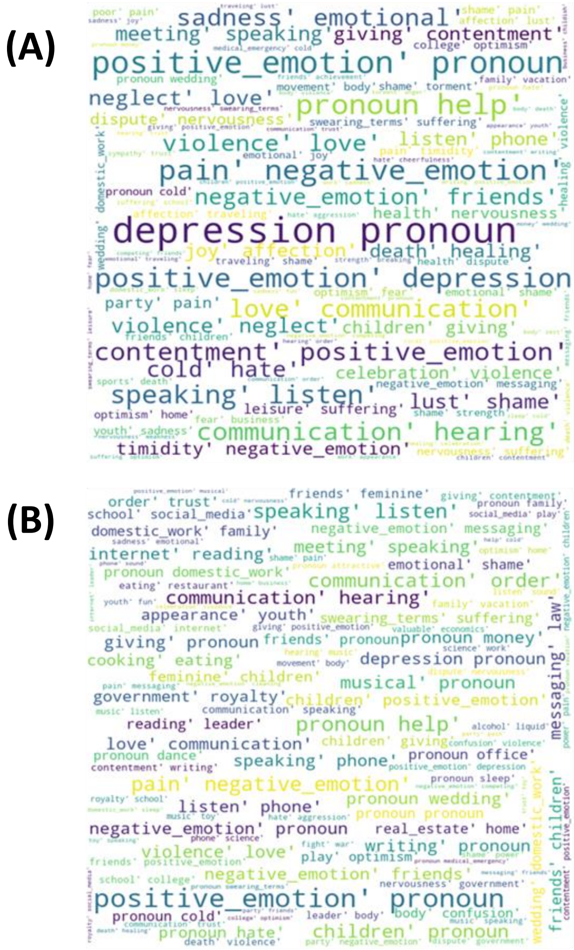

Figure 3: Word cloud demonstrating the frequency distribution of words in (A) lexicons generated by the *Empath* library from the depressive samples of the RSDD dataset, (B) lexicons generated by the *Empath* library from the non-depressive samples of the RSDD dataset.

are seen more frequently in the depressive samples. The word cloud of the lexicons generated by *Empath library*, as shown in Figure 3, demonstrates that the depressive samples include more affective words. Pronouns are also commonly observed in depressive samples.

4 Experiments

In this study, BERT and DistilBERT are used as backbones for the classification module. Both architectures are multi-layer bidirectional transformer encoders that use multi-head-attention. Transformer architectures (linear layer and layer normalization) are highly optimized in modern linear algebra frameworks (Sanh et al., 2019). Variations of the last dimension of the tensor (i.e., hidden size dimension) have a small impact on computation efficiency for a fixed parameters budget than variations of other factors such as the number of layers. BERT and DistilBERT backbones are the pre-trained BERT-Base-Uncased and DistilBERT-Base-Uncased which parameters are adapted from Devlin et al. (2018) and Sanh et al. (2019) (see Table 3). The classification module is composed of a global averaging layer with a pool size set to 3 and a stride of 3, then 2 hidden fully connected layers of 256 and 64 units, each followed by a rectified linear unit (ReLU) activation function.

Components	BERT	DistilBERT
Transform Block (L)	12	6
Hidden Size (H)	768	768
Self attention heads (A)	12	12
Max Sequence Length	256	256

Table 3: Model parameters used in this study for BERT and DistilBERT.

4.1 Hyper-parameter Setting and Fine-Tuning

For all experiments, models were optimized with a binary cross-entropy loss function using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001, and was trained with a batch size of 128. The maximum token length was set to 256. Early stopping was used to stop training after 10 epochs of no improvement in accuracy, and learning-rate scheduling was implemented to reduce the learning rate by a factor of 0.1 after 10 epochs of no reduction in loss. To leverage the pre-trained weights of BERT and DistilBERT, we initially fine tuned the models for 5 epoch without updating backbones. Then, we fine tuned the models for next 50 epochs with updating backbones. All experiments were performed on a AMD Radeon VII 16Gb GPU.

4.2 Text Data Augmentation

We tested BERT and DistilBERT with the two proposed text data augmentation techniques. For Negative Embedding, we updated the BERT and DistilBERT's vocabulary with the predefined negative words in order to avoid unknown padding (Devlin et al., 2018; Sanh et al., 2019). We applied WordPiece tokenization (Wu et al., 2016) to tokenize text samples and evaluated the newly-formed tokens with the predefined negative words to generate binary-valued negative embeddings (Figure 1). For Empathy, we firstly added 2 depression-related lexicons ("pronoun" and "depression") to the Empath library and applied the library to analyze text and generate emotional lexicons. The lexicons were then concatenated with original texts into new text samples (Table 1), which were finally tokenized by the WordPiece tokenization (Wu et al., 2016). The codes used in this study have been made publicly available[2].

[2]GitHub repository: https://github.com/deepkapha/depressio

Model	Train/ Val Loss	Train/ Val Precision	Train/ Val Recall	Train/ Val Accuracy
BERT	0.54 0.54	0.77 0.77	1.00 1.00	0.77 0.77
BERT+NE	**0.27** **0.35**	**0.92** **0.89**	0.88 0.85	**0.89** **0.86**
BERT +Empathy	0.69 0.69	0.56 0.56	1.00 1.00	0.56/ 0.55
DistilBERT	0.54 0.54	0.77 0.77	1.00 1.00	0.77 0.77
DistilBERT +NE	**0.25** **0.36**	**0.93** **0.90**	0.89 0.85	**0.90** **0.86**
DistilBERT +Empathy	0.69 0.69	0.56 0.55	1.00 1.00	0.56 0.56

Table 4: Training and validation, evaluation metrics for two text augmentation methods and no-augmentation on BERT and DistilBERT architectures. Val: validation set, NE: Negative Embedding.

5 Results and Analysis

We have examined the performance of BERT and DistilBERT with or without the use of text data augmentation methods. Results demonstrate that Negative Embedding leads to great improvements in model performance and therefore, outperforms Empathy and baseline BERT and DistilBERT models in the discrimination between depressive and non-depressive statements in a dataset with label noise (Table 4). The low precision and high recall of baseline models and Empathy suggest that label noise leads to more similarity in textual context among classes and causes the true non-depressive statements to be classified as depressive. While Negative Embedding achieved the highest precision, it led to lower recall. This shows that Negative Embedding increases the contextual differences by emphasizing negative words used in depressive statements. As a result, fewer true non-depressive statements were misclassified as depressive. Another expected result of Negative Embedding was the lower recall that some non-depressive statements in the depressive group might be classified as non-depressive.

For both BERT and DistilBERT, the use of Empathy led to low performance even compared with baseline models (Table 4). The concatenation of original texts and lexicons generated by the Empath library attempts to add high-level contextual information to the original context. However, the Empath library may generate lexicons that contradict the original context which leads to greater ambiguity in the concatenated text. In Table 1, lexicons "optimism, joy, positive_emotion" are generated for texts that are overall negative. Apart from that, the depressive and non-depressive statements processed by Empathy may share many similar lexicons generated by the Empath library due to its fixed list of lexicons. This could worsen the high contextual similarity between depressive and non-depressive classes caused by label noise.

According to Table 4, performances of BERT and DistilBERT with Negative Embedding are comparable. Figures 4 and 5 show that BERT and DistilBERT with Negative Embedding converge comparably, and converge faster than BERT and DistilBERT without text augmentation and BERT and DistilBERT with Empathy. These observations suggest that model distillation does not compromise the performance of models with Negative Embedding. During fine tuning, loss and accuracy of BERT and DistilBERT without text augmentation and BERT and DistilBERT with Empathy do not improve (Figures 4 and 5).

Based on Figures 4 and 5, BERT and DistilBERT without text augmentation does not converge for a given number of epochs, and one possible reason is the generalization gap between training and test data. The model training could have been benefited from a larger batch size of 4096 and the use of different learning rate schemas (Krizhevsky et al., 2012). To address this issue, the layer-wise adaptive moments optimizer (LAMB) has been proposed by (You et al., 2019) which can scale the batch size of BERT pre-training to 64K without losing accuracy. In this study, (You et al., 2019) used different sequence lengths of 128 and 512. In our future studies, we will include validation of the LAMB optimizer.

The use of Negative Embedding and Empathy in BERT-based architectures on label noise is a novel approach. Earlier studies by (Rodrigues Makiuchi et al., 2019) have used textual embeddings, to extract BERT textual features and employ a convolutional neural network (CNN) followed by a LSTM layer. Rodrigues Makiuchi et al. (2019) trained their model on a relatively clean, labelled dataset of the Patient Health Questionnaire (PHQ). They achieved a concordance correlation coefficient (CCC) score of 0.497. Our models on Negative Embeddings achieved a higher F1 score (approximately 87%) compared to (Dinkel et al., 2019), where Word2Vec and fastText embeddings were used on a sparse dataset and a F1 score of 35% on average was observed.

Future studies will investigate threshold-based severity grading of depression, which could be approached by tagging severe/strong negative affective words and application of thresholds and clinically established diagnosis criteria (Karmen et al., 2015). Multimodal datasets that include speech and text data are essential in emotion recognition for the detection of depression (Siriwardhana et al., 2020). Fusion architectures with BERT have been used by Siriwardhana et al. (2020) to perform emotion recognition using speech data, and to improve our study further, we will include speech data for incorporating more contextual information (Baevski et al., 2019).

Apart from the above, we will also explore knowledge-enabled bidirectional encoder representation from transformers (K-BERT) (Liu et al., 2020). K-BERT is capable of loading any pretrained BERT models as they are identical in parameters. In addition, K-BERT can easily inject domain knowledge into the models by using a knowledge graph without pre-training. The idea behind this approach is that the detection and assessment of depression is a very domain specific task and using a knowledge graph to inject contextual information into the BERT model may significantly improve model performance.

6 Conclusion

In this study, we investigated the negative impact of label noise on sentiment analysis for the detection of depression and investigated whether text data augmentation methods exploit contexts for robustness to label noise. For this purpose, we created and introduced the RSDD dataset and proposed 2 text augmentation techniques, i.e. Negative Embedding and Empathy. Our experimental results demonstrate that Negative Embedding leads to improved performance when compared with baseline BERT and Distil-BERT models, however, the use of Empathy with these models can cause decrease in detection accuracy. Taken together, when used with BERT and DistilBERT models, Negative Embedding exploits contextual information and im-

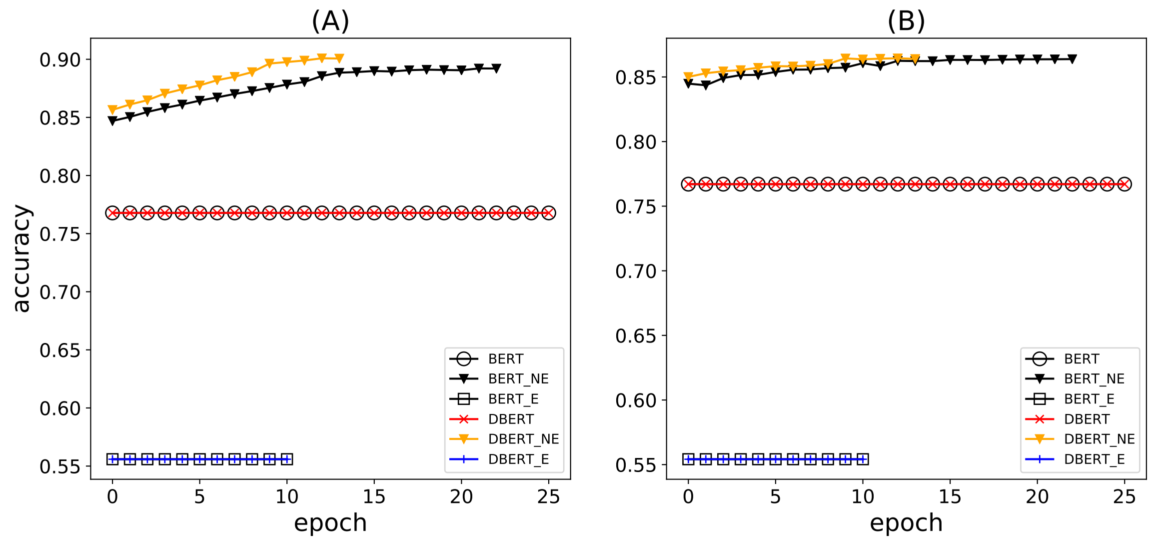

Figure 4: Accuracy during training when backbones are unfrozen. BERT with Negative Embedding and DistilBERT with Negative Embedding have the highest accuracy on (A) training and (B) validation sets. NE: Negative Embedding, E: Empathy, DBERT: DistilBERT.

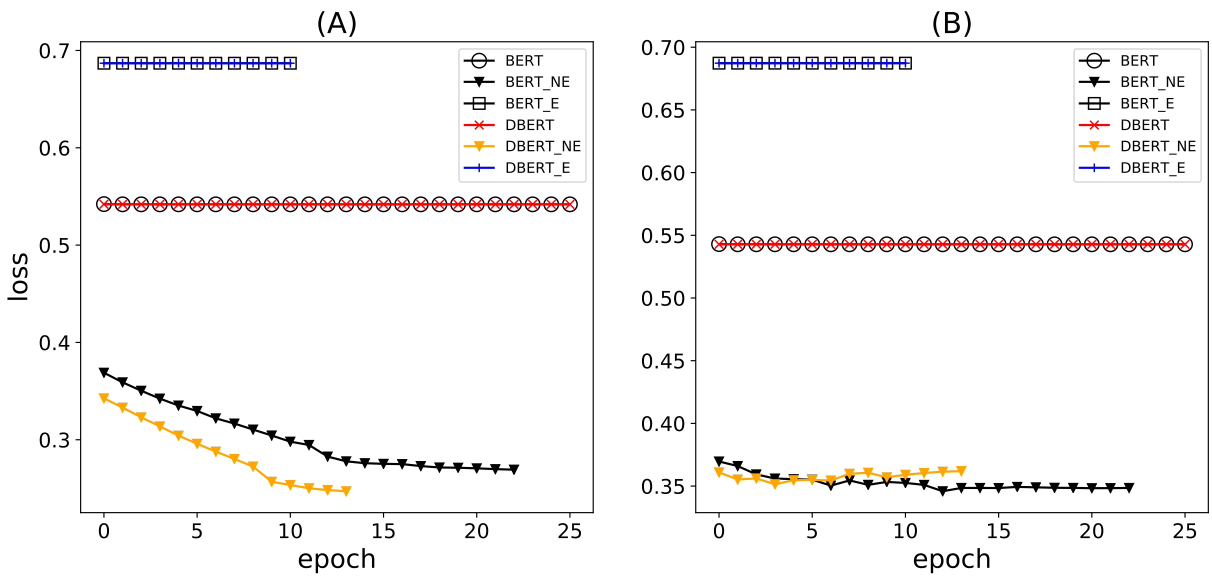

Figure 5: Loss during training when backbones are unfrozen. BERT with Negative Embedding and DistilBERT with Negative Embedding have the lowest loss on (A) training and (B) validation sets. NE: Negative Embedding, E: Empathy, DBERT: DistilBERT.

proves distinguishability between non-depressive and depressive classes, leading to high accuracy in the detection of depression based on text data.

References

CAMH Policy Advice. 2020. Mental health in canada: Covid-19 and beyond.

Amanuel Alambo, Manas Gaur, and Krishnaprasad Thirunarayan. 2020. Depressive, drug abusive, or informative: Knowledge-aware study of news exposure during covid-19 outbreak. *arXiv preprint arXiv:2007.15209*.

Görkem Algan and Ilkay Ulusoy. 2020. Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*.

Hatoon S AlSagri and Mourad Ykhlef. 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8):1825–1832.

Anelia Angelova, Yaser Abu-Mostafam, and Pietro Perona. 2005. Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 494–501. IEEE.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

Samir Bandyopadhyay and Shawni Dutta. 2020. Analysis of stress, anxiety and depression of children during covid-19.

World Bank. 2020. The impact of covid-19 on labor market outcomes: Lessons from past economic crises.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167.

Wilma Bucci and Norbert Freedman. 1981. The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Jonathan Campion, Afzal Javed, Norman Sartorius, and Michael Marmot. 2020. Addressing the public mental health challenge of covid-19. *The Lancet Psychiatry*, 7(8):657–659.

The Conversation. 2020. Here's how the coronavirus is affecting canada's labour market.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Kali Cornn. 2019. Identifying depression on social media.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Heinrich Dinkel, Mengyue Wu, and Kai Yu. 2019. Text-based depression detection on sparse data. *arXiv e-prints*, pages arXiv–1904.

Ismail Elezi. 2020. Exploiting contextual information with deep neural networks. *arXiv preprint arXiv:2006.11706*.

Marta Fana, Sergio Torrejón Pérez, and Enrique Fernández-Macías. 2020. Employment impact of covid-19 crisis: from short term effects to long terms prospects. *Journal of Industrial and Business Economics*, 47(3):391–410.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

David Flatow and Daniel Penner. 2017. On the robustness of convnets to training on noisy labels.

Josep Maria Haro, Lene Hammer-Helmich, Delphine Saragoussi, Anders Ettrup, and Klaus Groes Larsen. 2019. Patient-reported depression severity and cognitive symptoms as determinants of functioning in patients with major depressive disorder: a secondary analysis of the 2-year prospective perform study. *Neuropsychiatric Disease and Treatment*, 15:2313.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. 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. *The Lancet*, 392(10159):1789–1858.

Christian Karmen, Robert C Hsiung, and Thomas Wetter. 2015. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jeffrey S Kreutzer, Ronald T Seel, and Eugene Gourley. 2001. The prevalence and symptom rates of depression after traumatic brain injury: a comprehensive examination. *Brain injury*, 15(7):563–576.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. erisk 2020: Self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 557–563. Springer.

Dong Lu and Jennifer Bouey. 2020. Public mental health crisis during covid-19 pandemic, china. *Emerging Infectious Diseases*, 26(7).

Diego Maupomé and Marie-Jean Meurs. 2018. Using topic extraction on social media content for the early detection of depression. *CLEF (Working Notes)*, 2125.

Carmen Moreno, Til Wykes, Silvana Galderisi, Merete Nordentoft, Nicolas Crossley, Nev Jones, Mary Cannon, Christoph U Correll, Louise Byrne, Sarah Carr, et al. 2020. How mental health care should change as a consequence of the covid-19 pandemic. *The Lancet Psychiatry*.

Moin Nadeem. 2016. Identifying depression on twitter. *arXiv preprint arXiv:1607.07384*.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5(3):245–257.

Bryn Nelson and David B Kaminsky. 2020. Covid-19's crushing mental health toll on health care workers: Beyond its devastating physical effects, the pandemic has unleashed a mental health crisis marked by anxiety, depression, posttraumatic stress disorder, and even suicide. here, in part 1 of a 2-part series, we examine the growing effort to identify and alleviate the fallout for health care workers. *Cancer cytopathology*, 128(9):597–598.

Sayanta Paul, Sree Kalyani Jandhyala, and Tanmay Basu. 2018. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *CLEF (Working notes)*.

W Pedersen. 2008. Does cannabis use lead to depression and suicidal behaviours? a population-based longitudinal study. *Acta Psychiatrica Scandinavica*, 118(5):395–403.

Betty Pfefferbaum and Carol S North. 2020. Mental health and the covid-19 pandemic. *New England Journal of Medicine*, 383(6):510–512.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen M Rao, Gary J Leo, L Ellington, T Nauertz, L Bernardin, and F Unverzagt. 1991. Cognitive dysfunction in multiple sclerosis.: Ii. impact on employment and social functioning. *Neurology*, 41(5):692–696.

Usama Rehman, Mohammad G Shahnawaz, Neda H Khan, Korsi D Kharshiing, Masrat Khursheed, Kaveri Gupta, Drishti Kashyap, and Ritika Uniyal. 2020. Depression, anxiety and stress among indians in times of covid-19 lockdown. *Community mental health journal*, pages 1–7.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017a. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3838–3844.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Hu Tianrui, Chu Tat-Seng, and Wenwu Zhu. 2017b. Depression detection via harvesting social media: A multimodal dictionary learning solution. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3838–3844.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

James M Toolan. 1962. Suicide and suicidal attempts in children and adolescents. *American journal of psychiatry*, 118(8):719–724.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Maryam Vizheh, Mostafa Qorbani, Seyed Masoud Arzaghi, Salut Muhidin, Zohreh Javanmard, and Marzieh Esmaeili. 2020. The mental health of healthcare workers in the covid-19 pandemic:

A systematic review. *Journal of Diabetes & Metabolic Disorders*, pages 1–12.

Walter Weintraub. 1981. *Verbal behavior: Adaptation and psychopathology*. Springer Publishing Company New York.

Annelie Werbart Törnblom, Kimmo Sorjonen, Bo Runeson, and Per-Anders Rydelius. 2020. Who is at risk of dying young from suicide and sudden violent death? common and specific risk factors among children, adolescents, and young adults. *Suicide and Life-Threatening Behavior*, 50(4):757–777.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

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*.

Zhengxuan Wu and Desmond C Ong. 2020. Context-guided bert for targeted aspect-based sentiment analysis. *arXiv preprint arXiv:2010.07523*.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019a. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 387–394.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Reducing bert pre-training time from 3 days to 76 minutes. *arXiv preprint arXiv:1904.00962*.

Xingquan Zhu and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.