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

Anonymous AACL-IJCNLP submission

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 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 requires labeling by domain experts. Therefore, in this study, we used the Reddit Self-reported Depression Dataset, an uncurated text-based dataset with label noise, 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 the Empathy approach 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). Depressed patients show emotional, physical, and cognitive symptoms which impair their daily functioning (Haro et al., 2019; Rao et al., 1991), and clinical 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 caused significant distress around the globe and has caused serious damage to public mental health (Advice, 2020; Campion et al., 2020; Lu and Bouey, 2020; Pfefferbaum and North, 2020). The results indicated that people who do not have enough supplies to sustain 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). Many of these professions are experiencing larger job losses, reductions in hours, wages and labour force participation (Bank, 2020). People are relying on federal and provincial governments aid packages to alleviate their financial burdens (Conversation, 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). Because of continuing lockdown access to mental healthcare facilities also became limited (Moreno et al., 2020). Hence, early detection of depression symptoms is essential in these hard hit

times and to save individuals to take extreme measures such as self-harm (Losada et al., 2020).

With the rapidly increasing internet usage, people have shared their experiences and challenges with mental health disorders 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 in 1981 by Bucci and Freedman (1981) and by Weintraub (1981) and was then confirmed in a study at University of Texas that formerly- and currently-depressed students use the pronoun "I" and negative emotional 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-based classification algorithms, researchers succeeded (Nadeem, 2016; Paul et al., 2018; Benton et al., 2017; Coppersmith et al., 2015; Maupomé and Meurs, 2018; Resnik et al., 2015) in a higher performance improvement of new forms of potential health care solutions and methods.

BERT (Devlin et al., 2018) has become a standard and baseline language model that scientists have extensively implemented to achieve state-of-the-art performance in various language understanding tasks. Since being composed of Attention-based Transformer blocks and being pre-trained on large corpora (i.e. BookCorpus of 800M words, English Wikipedia of 2,500M words) (Zhu et al., 2015), BERT can capture rich linguistic knowledge and deep context in text; hence, BERT can serve as a backbone to be easily fine-tuned for downstream tasks with high performance. Recent research works (Trotzek et al., 2018) show prospective applications of Deep Neural Networks in detecting depression and severity of self-harm with high accuracy in self-reported postings on social media. Alambo et al. (2020) applied three variants of Bidirectional Encoder Representations from Transformers (BERT) on streaming news content related to COVID-19 and achieved an test accuracy of 89% for Depression-BERT and 78% for Drug Abuse-BERT.Because of proven performance of BERT, we chose BERT as our baseline method to classify self-reported depressing statements. Web-scraping with fixed labeling rules is a common approach for building a large-scale text dataset 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 depressing and non-depressing statements from 2 subreddits, /depression and /AskReddits. A drawback of this method is label noise that labels are mislabelled by non-experts or simple labeling rules. The simple labeling rules may mislabel a normal statement in the depressing group as depressed (Cornn, 2019). For example, in the sample post in the /depression subreddit, the fixed labeling rule may label both its negative title and the first comment with positive and supportive contents as depressed (Table 2). 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 and diagnosis of depression (Cornn, 2019). To exploit context 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. We then experimented 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 relies on finding heuristic points which are corrupted by label noise and filtering them out (Angelova et al., 2005; Brodley and Friedl, 1999). Current techniques focus on modifying learning algorithms and the architectures of neural networks 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. Deep neural network is trained on a set of noisy data in comparison to clean data. Researchers (Vahdat, 2017; Siddhant Garg and Thumbe, 2021) have used deep neural networks in NLP studies that are robust

Original text:	This is so frustrating. I'm sorry you're experiencing this. I know how you feel.
Tokens:	this is so frustrating . i ' m sorry you ' re experiencing this . i know how you feel .
Segment Embeddings:	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Negative Embeddings:	0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

Figure 1: Comparison between Segment Embeddings and Negative Embeddings

against label noise. We use an alternative approach to exploit context for robustness to label noise. Modifying deep neural networks with context modules have achieved state-of-the-art results for many image-based tasks (Elezi, 2020). For language-based tasks, recent studies addressed label noise by supplying additional contextual information to the attention models as 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 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,(Wu and Ong, 2020) and (Yang et al., 2019a) modified the attention mechanism to calculate attention weights based on contextual weights. Recently, (Dai et al., 2019) and (Yang et al., 2019b) introduced Transformer-XL and XL-Net which implemented autoregressive language understanding and outperformed standard attention networks in capturing contextual dependency. In the study, instead of proposing a new attention-based architecture, we presented two data augmentation techniques, Negative Embedding and Empathy to further fine tune attention networks to be noise tolerant in depression detection. Our solutions have 2 advantages: (1) they avoided increasing computational loads and (2) leveraged pretrained weights learned from large-scale text corpora.

2.1 Negative Embedding

In the novel Conditional Masked Language Model (MLM) pre-training task for Transformers (Wu et al., 2019), the segmentation embedding is replaced with 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 (Fig. 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 pre-defined in studies of (Hu and Liu, 2004) and (Liu et al., 2005).

2.2 Empathy

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, zest, 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: Empathy generated lexicons and concatenate them with original text

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 most commonly used libraries is LIWC (Linguistic Inquiry and Word Count) which counts words relevant to lexical categories such as sadness, health, and positive emotions (Tausczik and Pennebaker, 2010) . For example, positive emotion lexicon has relevant words such as happy, joy, fun, etc. In pub-

lished studies (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 text data.

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

3 Reddit Self-reported Depression Dataset (RSDD)

Currently, there is no publicly available, large-scale text dataset for the diagnosis of depression . Hence, we utilized Python Reddit API Wrapper to web-scrape posts from January 2018 to November 2020 in 2 the subreddits /AskReddit and /depression, which correspond to "nondepressed" and "depressed" classes respectively. For each post, its content and comments were web-scrapped and treated as separate samples. For each class, the first 100,000 samples were selected. The total number of text samples (see footnote) was 200,000 with a balanced class ratio of 1:1; and 160,00 and 40,000 text samples were split for class-balanced training and validation sets. Table 2 is one example of the dataset that belongs to the category "depression" 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. "Don't worry. You will be fine.") in the /depression subreddit that was classified as "depressed" during the automatic web-scraping process.

The inference that we can draw from the above Word-cloud is that words like Feel, depression , want , friend etc have more occurrence in depressed dataset in contrast to non-depressed dataset.

The inference that we can draw from the Word Cloud representations of Empath classes are that the depressed datasets tend to show more occurrence of emotional classes wrt the non-depressed classes. We can also see that pronouns also have a larger occurrence frequency in depressed classes.

/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.
/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 Reddit Self-reported Depression Dataset for /r/depression and /r/AskReddit

4 Experiments

In this study, BERT and DistilBERT serve as backbones for the classification module (Fig. 5). Both the architectures are multi-layer bidirectional Transformer encoders that use bidirectional self-attention. Transformer architecture (linear layer and layer normalisation) are highly optimized in modern linear algebra frameworks (Sanh et al., 2019). The variations on the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency (for a fixed parameters budget) than variations on other factors like the number of layers.The pretrained model is a BERT base (Devlin et al., 2018) that is composed of a global averaging layer with pool_size = 3 and strides = 3, and 2 hidden fully connected layers of 256 and 64 nodes, each followed by a rectified lin-

Figure 2: (A) Word cloud of the Reddit dataset for most frequently occurring words-depressed dataset (B) Word cloud of the Reddit dataset for most frequently occurring words-not depressed dataset.

Figure 3: (A) Word cloud of the Reddit dataset for most frequently occurring Empath classes-depressed dataset (B) Word cloud of the Reddit dataset for most frequently occurring Empath classes-not depressed dataset.

ear unit (ReLU) activation function. The specifics of the components are given in Table 3. In the DistilBERT models the number of layers are reduced. The parameters for pretrained DisitlBERT in this study are given in Table 3 (Sanh et al., 2019).

4.1 Hyper-parameter Setting and Fine-Tuning

For all experiments, the hyper-parameters are Adam (Kingma and Ba, 2014) for optimization, learning rate=0.0001, loss=Binary Cross-Entropy, batch_size=128, and max_token_length=256. Also, early stopping was implemented to stop training after 10 epochs of no accuracy improvement; and learning-rate scheduling was implemented to reduce the learning-rate by a factor of 0.1 after 10 epochs of no loss reduction. To leverage the pre-trained weights of BERT and DistilBERT, we initially unfroze the backbones and fine-tuned for 5 epochs; then, we unfroze them and fine-tuned in the next 50 epochs. All experiments were fine-tuned on a AMD Radeon VII 16Gb GPU.

4.2 Text Data Augmentation

We experimented with BERT and DistillBERT with 2 text data augmentations: Negative Embedding, and Empathy. 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 the WordPiece tokenization (Wu et al., 2016) to tokenize text samples and evaluated the newly-formed tokens with the predefined nega-

Components	BERT	DistilBERT
Transform Block (L)	12	6
Pool_size	3	3
Strides	3	3
Hidden size (H)	768	768
Fully Connected Layer	256	256
Self attention heads (A)	12	12

Table 3: Model parameters for BERT and DistilBERT

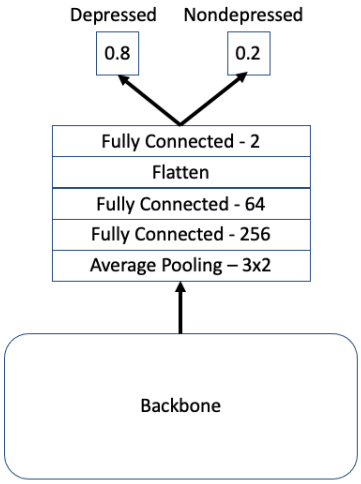

Figure 4: The shared (Distil) BERT classifier architecture

tive 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 origins text into new text samples (Figure 1) which are finally tokenized by the WordPiece tokenization (Wu et al., 2016).

5 Results and Analysis

We have examined the performance of BERT and Distil-BERT with or without the use of text data augmentation methods (Table 4). The results show that Negative Embedding leads to the greatest improvements of the classifiers' performance and therefore, outperforms Empathy and baseline BERT and Distil-BERT models in the discrimination between depressive and non-depressive statements in a dataset with label noise. The low precision and high recall of baseline models and Empathy (Table 4) suggest that label noise leads to more similarity in context of text among classes and causes true normal statements to be classified as depressive. While Negative Embedding achieved

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

Table 4: Training and validation evaluation metrics for different text augmentation methods on two classifier architectures. "Val." represents Validation, "NE" represents Negative Embedding

the best precision and recalls, it leads to the lower recalls. This observation shows that Negative Embedding increases the context difference by emphasizing the negative word usage in depressed statements. As a result, less true non-depressive statements are misclassified as depressive. Also, Negative Embedding may cause non-depressive statements in the depressing group to be classified as non-depressed.

For both BERT and DistilBERT, the use of Empathy led to low performance even compared with baseline models (Table 4). The concatenation of original text and lexicons generated by the Empath library attempts to add high-level context information to the original context. However, the Empath library may generate lexicons contradicted with the original context which leads to the final ambiguous context in the concatenated text. In Table 1, lexicons "optimism, joy, positive_emotion" are generated for the text containing negative sense. Also, 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.

Look on Table 4, performance of BERT and DistilBERT with Negative Embedding are comparable. Figures 5 and 6 show that BERT and DistilBERT with Negative Embedding comparably converge and converge faster than BERT and DistilBERT without text augmentation and with Empathy. These observations suggest that the model

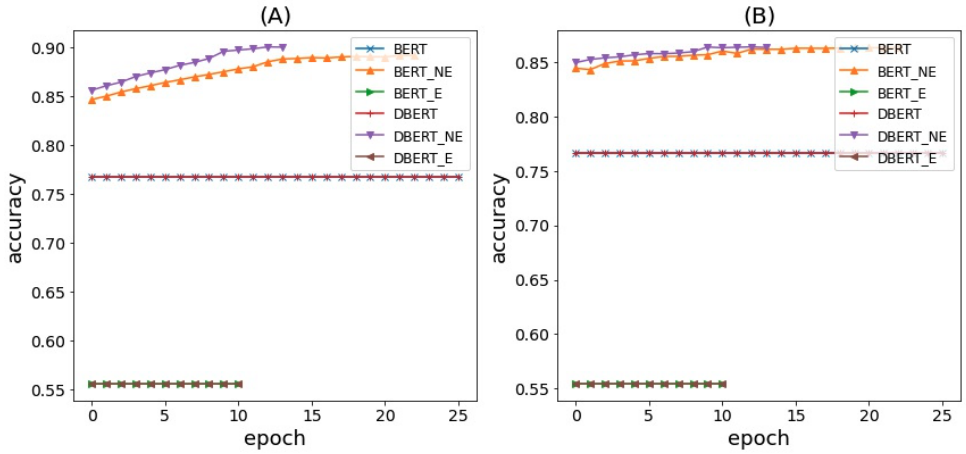

Figure 5: Accuracy evaluation by epoch when backbones are unfrozen. BERT+NE and DistilBERT+NE have the highest accuracies for training (A) and validation (B) sets. NE represents Negative Embedding, E represents Empathy, and DBERT represents DistilBERT.

distillation does not degrade the Negative Embedding performance. When fine-tuning, BERT and DistilBERT without text augmentation and with Empathy do not improve losses and accuracies (Figures 5 and 6).

Based on Figure 5 and 6 it is observed that BERT base and DistilBERT does not converge for a given number of epochs. This could be possible because of the *generalization gap* between train and test data. Hence need longer time to converge. Also the model training could have been benefited with larger batch size 4096 and by using different learning rate schemas (Krizhevsky et al., 2012). To overcome the limitation LAMB (LAYER-WISE ADAPTIVE MOMENTS OPTIMIZER) optimizer has been proposed by (You et al., 2019) that can scale the batch size of BERT pre-training to 64K without losing accuracy. For their training (You et al., 2019) used different sequence lengths of 128 and 512 respectively. Our future studies will include validating LAMB optimizer with labelled noise dataset.

The use of Negative Embedding and Empathy in BERT-based architectures on label noise is a novel application. 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 Patient Health Questionnaire (PHQ). They achieved a Concordance Correlation Coefficient (CCC) score equivalent to 0.497. Our models on Negative Embeddings achieved a higher F1 score of 87% score compared to (Dinkel et al., 2019). Dinkel et al. (2019) used Word2Vec and fastText embeddings on sparse dataset to achieve a F1 of 35% on average. Future studies will include severity analysis based on thresholds to classify the intensity of depression (none, mild, moderate, severe). Severity analysis can be approached by tagging severe/strong negative emotional words and basing on a threshold value based on clinical psychology (Karmen et al., 2015). Multimodal dataset that includes speech and text are essential in emotion recognition as is the case for depression detection [(Siriwardhana et al., 2020).Fusion architectures with BERT is used by Siriwardhana et al. (2020) to identify Speech Emotion Recognition. To improve our study further will include speech data for more contextual information (Baevski et al., 2019).

Apart from this we will also explore K-BERT which uses Knowledge-enabled Bidirectional Encoder Representation from Transformers. K-BERT is capable of loading any pre-trained BERT models as it is identical in parameters. In addition, K-BERT can easily inject domain knowledge into the models by equipped with a Knowledge Graph without pre-training.The idea behind this approach is that depression analysis is a very domain specific task and using a Knowledge graph to inject contextual information in the BERT model can significantly improve the model performance (Liu et al., 2020).

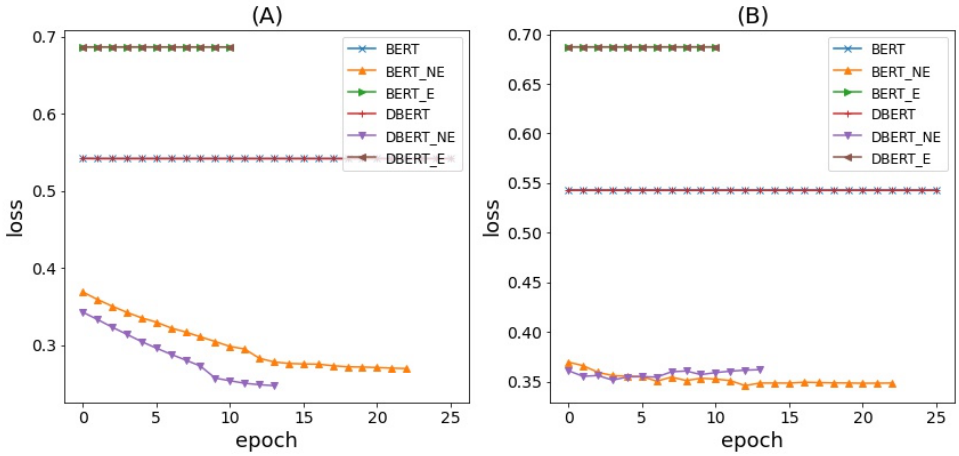

Figure 6: Learning curve by epoch when backbones are unfrozen. BERT+NE and DistilBERT+NE have the lowest losses for training (A) and validation (B) sets. NE represents Negative Embedding, E represents Empathy, and DBERT represents DistilBERT

6 Conclusion

In this paper, we investigated the negative impact of label noise in sentiment analysis for the diagnosis of depression and investigated whether text data augmentation methods exploit context for robustness to label noise. For this purpose, we introduce the RSDD dataset and propose 2 text augmentation techniques, Negative Embedding and Empathy. Our experimental results demonstrate that Negative Embedding may lead to improved performance when compared with baseline BERT and Distil-BERT models, however, using Empathy with these models can cause decrease in diagnostic accuracy . Taken together, when used with BERT and Distil-BERT models, Negative Embedding exploits contextual information and improves distinguishability between classes, showing high accuracy in the diagnosis of depression based on self-reported 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.

Goutham Ramakrishnan Siddhant Garg and Varun Thumbe. 2021. Towards robustness to label noise in text classification via noise modeling. *Computing Research Repository*, arXiv:2101.11214v1. Version 2.

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.

Arash Vahdat. 2017. Toward robustness against label noise in training deep discriminative neural networks. *arXiv preprint arXiv:1706.00038*.

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.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

LaTeX-specific details: Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudocode or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.