Data Augmentation for Mental Health Classification on Social Media

Gunjan Ansari
JSS Academy of Technical Education, Noida, India
gunjanansari@jssaten.ac.in

Muskan Garg
Thapar Institute of Engineering & Technology Patiala, Punjab, India
muskanphd@gmail.com

Chandni Saxena
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
csaxena@cse.cuhk.edu.hk

Abstract

The mental disorder of online users is determined using social media posts. The major challenge in this domain is to avail the ethical clearance for using the user-generated text on social media platforms. Academic researchers identified the problem of insufficient and unlabeled data for mental health classification. To handle this issue, we have studied the effect of data augmentation techniques on domain-specific user-generated text for mental health classification. Among the existing well-established data augmentation techniques, we have identified Easy Data Augmentation (EDA), conditional BERT, and Back-Translation (BT) as the potential techniques for generating additional text to improve the performance of classifiers. Further, three different classifiers- Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) are employed for analyzing the impact of data augmentation on two publicly available social media datasets. The experimental results show significant improvements in classifiers' performance when trained on the augmented data.

1 Introduction

Recent studies over mental health classification (Salari et al., 2020; Garg, 2021; Biester et al., 2021) convey that amid COVID-19 pandemic, the number of stress, anxiety and depression related mental disorders have increased. As per the recent survey, the rate of increase of mental disorders is more than those of physical health impacts on the Chinese population (Huang and Zhao, 2020). In this context, the early detection of psychological disorders is very important for good governance. It is observed that more than 80% of the people who commit suicide, disclose their intention to do so on social media (Sawhney et al., 2021). Clinical depression is the result of frequent tensions and stress. Further, prevailing clinical depression for a longer time period results in suicidal tendencies.

The information mining from social media helps in identifying stressful and casual conversations (Thelwall, 2017; Turcan and McKeown, 2019; Turcan et al., 2021). Many Machine Learning (ML) algorithms are developed in literature using both automatic and handcrafted features for classifying Microblog. The problem of data sparsity is under-explored for mental health studies on social media due to the sensitivity of data (Wongkoblap et al., 2017). Multiple ethical clearances are required for new developments in mental health classification. To deal with this issue of data sparsity, we have used data augmentation techniques to multiply the training data (Turcan and McKeown, 2019; Haque et al., 2021). The increase in training data may help to improve the hyper-parameter learning of textual features and thereby, reducing overfitting. Data Augmentation is the method of increasing the data diversity without collecting more data (Feng et al., 2021). The idea behind the use of Data Augmentation (DA) techniques is to understand the improvements in training classifiers for mental health detection on social media.

In this manuscript, the mental health classification is performed for two datasets to test the scalability of data augmentation approaches for mental healthcare domain. The classification of casual and stressful conversations (Turcan and McKeown, 2019), and classifying depression and suicidal posts (Haque et al., 2021) on social media. We select a rule based approach which preserves the original label and diversifies the text. To the best of our knowledge, this is the first attempt of stuffing additional data for mental health classification and there is no such study in the existing literature. The key contributions of this work are as follows:

- To determine the feasibility and the impor-

Proceedings of the 18th International Conference on Natural Language Processing, pages 152–161
Silchar, India. December 16 - 19, 2021.

tance of data augmentation in the domain-specific study of mental health classification to solve the problem of data sparsity.

- The empirical study for different classification algorithms show significantly improved F-measure.

Ethical Clearance: We use limited, sparse and publicly available dataset for this study and so, no ethical approval is required from the Institutional Review Board (IRB) or elsewhere.

We organize rest of the manuscript in different sections. Section 2 describes the historical perspective of data augmentation and mental health classification on social media. We discuss the data augmentation methods and the architecture for experimental setups in Section 3. Section 4 elucidates the experimental results and evaluation over the proposed architecture of experimental setup which shows the significance and feasibility of data augmentation over mental health classification problems. Finally, Section 5 gives the conclusion and future scope of this work.

2 Related Work

Mental health classification can be quite challenging without the availability of sufficient data. Although the users' posts can be extracted from the social media platforms such as Reddit, Twitter and Facebook, annotating these posts is quite expensive. To address this issue, researchers have proposed different data augmentation techniques suitable for Natural Language Processing (NLP) which varies from simple rule-based methods to more complex generative approaches (Feng et al., 2021). The data augmentation tasks is categorized into conditional and unconditional augmentation task (Shorten et al., 2021).

2.1 Evolution of textual Data Augmentation

The unconditional data augmentation models like *Generative adversarial networks* (Goodfellow et al., 2014) and *Variational autoencoders* (Kingma and Welling, 2014) generates the random texts irrespective of the context. We do not use unconditional data augmentation for this task as it is required to preserve the context of the information as per the label. The conditional masking of a few tokens in the original sentence was observed to boost the classification performance in NLP tasks (Li et al., 2020; Wu et al., 2021). Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the pre-trained language models, are proposed with the objective to capture the left and right context in the sentence to generate the masked tokens. The pre-trained autoencoder model conditional BERT (Wu et al., 2019; Kumar et al., 2021) is used as a well-established technique for generating label compatible augmented data from the original data.

One of the simplest rule-based data augmentation techniques is proposed as Easy Data Augmentation (EDA) (Wei and Zou, 2019). The authors proposed four random operations such as *random insertion, random deletion, random swapping and random replacement* on the given text for generating new sentences. The experimental results give better performance on five benchmark text classification tasks (Wei and Zou, 2019), as the true labels of the generated text were conserved during the process of *data augmentation*. A graph based data augmentation is proposed for sentences using balance theory and transitivity to infer the pairs generated by augmentation of sentences (Chen et al., 2020). The sentence-based data augmentation is not suitable for the problem of mental health classification on Reddit data as the posts contain large paragraphs.

Back Translation (BT) or Round-trip translation is another augmentation technique which is used as a pipeline for text generation (Sennrich et al., 2015). The BT approach converts the A language of text to B language of text and then back to A language of the same text. This back-translation (Corbeil and Ghadivel, 2020) of data helps in diversifying the data by preserving its contextual information. Although, the interpolation techniques are proposed for data augmentation (Zhang et al., 2017), it is minimally used for textual data in existing literature (Guo et al., 2020).

In our work, we have studied the effect of all three different augmentation techniques- EDA, Conditional BERT and Back-translation to increase the size of training data for the task of mental health classification.

2.2 Mental Health Classification: Historical Perspective

The existing literature on mental health detection and analysis of social media data (Garg, 2021) shows the problem of automatic labeling as *noisy*

labels. To handle this, either the label correction of noisy labels is required as shown in SDCNL (Haque et al., 2021) for manual labeling, or data augmentation (Chen et al., 2021). Since many existing datasets for mental health detection like RSDD, SMHD (Harrigian et al., 2020), CLPsych (Preoţiuc-Pietro et al., 2015) needs ethical clearance and are available only on request, we intend to pick small dataset with limited set of instances which are available in the public domain.

The Dreaddit dataset is manually labelled as stressful and casual conversation (Turcan and McKeown, 2019). In SDCNL dataset (Haque et al., 2021), the posts related to clinical depression and suicidal tendencies use similar words. Thus, we hypothesize that experimental results with data augmentation for classifying depression and suicidal risk may not generate well diversified data. In this manuscript, we use three data augmentation methods to text and validate the performance of the classifiers over both Dreaddit and SDCNL dataset.

3 Background: Data Augmentation Methods

Data augmentation (Feng et al., 2021) is a recent technique used for NLP to handle the problem of data sparsity by increasing the size of the training data without explicitly collecting the data. In this Section, we describe three potential textual data augmentation techniques, problem formulation, and architecture of the experimental setup.

3.1 Textual Data Augmentation

Out of many data augmentation tasks for NLP classification, very few are related to this problem domain of mental healthcare. This limitation is due to the presence of ill-formed (user-generated) text and the need to preserve the contextual information as per the label of the instances. To handle this issue, we use three different approaches. The first approach is based on NLP-based Augmentation technique (Wei and Zou, 2019), the second is based on conditional pre-trained language models such as BERT (Kumar et al., 2021) and the third approach is based on back translation (Ng et al., 2019). We briefly explain these methods in this section.

3.1.1 Easy Data Augmentation

In the previous work (Wei and Zou, 2019), NLP-based operations have been shown to achieve good results on text classification tasks. This method of data augmentation helps in diversifying the training samples while maintaining the class label associated with the post of a user at sentence level. The following four operations have been used in this work for augmenting the data:

- Synonym Replacement. Randomly n-words are chosen other than stop words from each sentence and replaced by one of its synonyms.
- Random Insertion. In this operation, a random synonym of a random word is inserted into a random position of a sentence for n number of times.
- Random Swap. Two words are randomly chosen in a sentence and swapped.
- Random Deletion. A word is deleted from a sentence with probability p.

3.1.2 Pre-Trained Language Models

Recently, deep bi-directional models have been used for generating textual data (Kobayashi, 2018; Song et al., 2019; Dong et al., 2017). These models are pre-trained with unlabelled text which can be fine tuned in autoencoder (Devlin et al., 2019), auto-regressive (Radford et al., 2019), or seq2seq (Lewis et al., 2019) settings. In *autoencoder settings*, a few tokens are randomly masked and the model is trained to predict alternative tokens. In *auto-regressive settings*, the model predicts the succeeding word according to the context. In *seq2seq settings*, the model is fine tuned on denoising autoencoder tasks. These transformers use associated class labels to generate the augmented text which helps in preserving its label. In this work, we adopt a framework[1] defined by (Kumar et al., 2021) and fine tune pre-trained BERT in auto-regressive settings.

3.1.3 Back Translation

Back translation (BT) is the data augmentation technique used for diversifying the information by changing the language of textual data to some language A and changing it back to its original language. In this experimental framework, we have used German as an intermediate language A. We use BT for the Microblogs by first converting it into German language using Neural Machine Translation (Ng et al., 2019) and then converting it back to

[1]https://github.com/amazon-research/transformers-data-augmentation

the English language. It is interesting to note that ill-formed and user-generated information is converted to the standard English language using BT and thus, spelling mistakes are reduced. Although the content is changed, contextual information is preserved.

3.2 Problem Formulation

Given a dataset D consisting of n-training samples where each sample is a text sequence x consisting of m-words and each sequence is associated with a label y. The objective is to generate an augmented data D_{syn} of n-synthetic samples using EDA, BERT and Back Translation.

3.2.1 AugEDA: Data augmentation using Easy Data Augmentation

In our work, 30% words of i^{th} training sample are randomly chosen for applying any one of the four EDA operation-Synonym Replacement, Random Insertion, Random Swap and Random Deletion (Wei and Zou, 2019). In synonym replacement, the chosen word is substituted by any one of the randomly selected synonym of this word from WordNet(Miller, 1995). In random insertion, j random positions are chosen for inserting random synonym of randomly chosen word out of m-words. In random swap, two words are randomly chosen from m-words and swapped with each other. A word is deleted with 10% probability in random deletion operation. The new sentence generated after applying any one of the lexical substitution method is added to the synthetic dataset D_{syn}. The process is repeated for n-training samples to create an augmented dataset of size n.

3.2.2 AugBERT: Data augmentation using BERT

We use the conditional BERT language model to generate the augmented data. We consider the label y and sequence $S = S_1, S_2...S_N$ of n-tokens to calculate the probability $p(t_i) = (.|y, S)$ of masked token t_i unlike masked language models that use only sequence S for predicting the probability of masked tokens. As defined by (Kumar et al., 2021), the conditional BERT model prepends associated label y to each sequence S in dataset D without adding it to the vocabulary of the model. For fine tuning of the model, some tokens of the sequence are randomly masked and the objective is to predict the original token according to the context of the sequence.

3.2.3 AugBT: Data augmentation using Back-Translation

To generate new textual data using Back-Translation, each of i^{th} training sample x_i is converted into a sentence y_i written in German language and then y_i is converted back to a sentence z_i in English. The generated sentence z_i is added to the augmented dataset D_{syn}. This process is repeated for n training samples to create an augmented dataset of n samples.

3.3 Architecture: Experimental Setup

The architecture of the experimental setup for augmenting domain-specific data of mental health classification from social media posts is shown in Figure 1. The Microblogs are given as an input for classifying the mental health of the users. The idea behind this approach is to generate some sequence of sentences and augment some more data for better training of classifiers. Thus, the number of instances are increased by using different data augmentation techniques.

The results are implemented for two publicly available mental health datasets, namely, Dreaddit and SDCNL. The dataset is divided into training and testing data. The training data is given as an input to the data augmentation methodologies, namely, EDA (Wei and Zou, 2019), Autoencoder conditional BERT (Wu et al., 2019) and Back-Translation (Ng et al., 2019). These three approaches are well established approaches for data augmentation in classification of the textual data. The original training data is almost doubled in the process of the data augmentation. The original and augmented data are fed to different machine learning classifiers for results and analysis.

4 Experimental Results and Evaluation

In this section, we discuss the datasets and the experimental results. We further analyze results for data diversity and statistical significance of the classifiers over augmented data as compared to the original data.

4.1 Dataset

The idea behind this study is to improve the training parameters of the classifier by removing the limitation of data sparsity. The two sparse datasets which are used for domain-specific data augmentation are Dreaddit [2] (Turcan and McKeown, 2019)

[2]http: //www.cs.columbia.edu/˜eturcan/data/ dreaddit.zip.

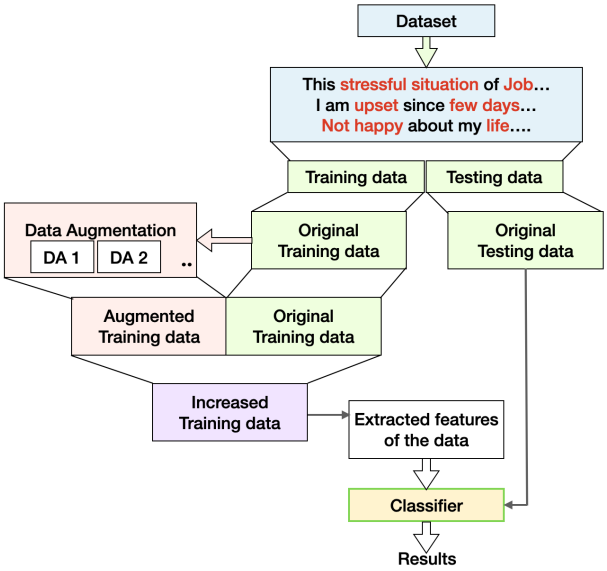

Figure 1: The Architecture of Experimental Setup for Data Augmentation

and SDCNL[3] (Haque et al., 2021) from existing literature are explained in this Section.

4.1.1 Dreaddit dataset

The Dreaddit dataset(Turcan and McKeown, 2019) consists of lengthy posts in five different categories and is used for classifying stressful posts from casual conversations. The categories of subreddits selected by authors having stressful conversations are interpersonal conflicts, mental illness (anxiety and PTSD), financial and social.

Dataset	Stress	Non-Stress
Training data	1488	1350
Testing data	369	346

Table 1: Dreaddit Dataset Statistics

Out of total 187444 posts scraped from these five categories, the authors have manually labelled 3553 Reddit posts. While selecting the posts for annotation, the authors selected those segments whose average token length was greater than 100. The average tokens per post in this dataset is 420 tokens. This statistics of the Dreaddit dataset is shown in Table 1.

4.1.2 SDCNL dataset

The SDCNL dataset(Haque et al., 2021) is scrapped from Reddit social media platform from two subreddits: *r/SuicideWatch* and *r/Depression* to carry out the study for classifying posts into depression specific or suicide specific. This dataset contains 1895 posts containing 1517 training samples and 379 testing samples. The dataset contains title, self-text and megatext of the reddit tweets along with other fields.

Dataset	Depression	Suicide
Training data	729	788
Testing data	186	193

Table 2: SDCNL Dataset Statistics

In this dataset, 729 out of 1517 instances are labelled as depression specific posts as shown in Table 2. The dataset is manually labelled to reduce noisy automated labels. The idea behind using this data is that we hypothesise that this dataset is even more complex than the Dreaddit dataset due to the presence of similar domain-specific words in posts.

4.2 Experimental Setup

The original and the augmented dataset used for experimentation is quite noisy as the posts used in this data is user-generated natural language text expressing the feelings of the writer. The pre-processing steps are applied using the NLTK library[4] of Python (Bird, 2006). The data is transformed before applying the supervised learning models employed in this work. The posts are long paragraphs, so in the first step the data is tokenized into sentences and then sentences are further tokenized into words. After removal of stop-words, punctuations,unknown characters from the extracted tokens, we use stemming and lemmatization to extract the root words.

After pre-processing of the data, it is transformed to a feature vector using Term Frequency- Inverse Document Frequency (TF-IDF), Word2Vec (Goldberg and Levy, 2014) and Doc2Vec (Lau and Baldwin, 2016). Word2Vec embedding and Doc2Vec embedding provides dense vector representation of data while capturing its context. In this research work, the Gensim library[5] is used to learn word embeddings from the training corpus using skip-gram algorithm. A vector of 300 dimensions is chosen and default settings of Word2Vec and Doc2Vec models are used for experiments and evaluation.

The learning based classifiers which are used for this research work are the Logistic Regression

[3]https://github.com/ayaanzhaque/SDCNL

[4]https://www.nltk.org/

[5]https://pypi.org/project/gensim/

(LR), the Support Vector Machine(SVM), and the Random Forest (RF) with the default settings of scikit-learn[6] (sklearn) library of Python. The hardware configuration of the system which is used to perform this study is 2.6 GHz 6-core Intel Core i7, Turbo Boost up to 4.5 GHz, with 12 MB shared L3 cache.

4.3 Experimental Results

We reference (Kumar et al., 2021) for implementation [7] and use AugBERT, AugEDA, and AugBT on two datasets- Dreaddit and SDCNL. The dataset is divided into 75% training and 25% testing set and the value of Precision (P), Recall(R) and F1 score (F1) are computed on the testing samples to evaluate the performance of the classifiers with and without domain -specific data augmentation for mental health classification. Table 3 and Table 4 presents the results achieved for original and augmented data for Dreaddit and SDCNL using three different classifiers, namely, Logistic regression (LR), Support Vector Machine (SVM) and Random Forest (RF), respectively.

4.3.1 Experimental Results for Dreaddit

As observed from Table 3, the F1 score showed an average improvement of around 1.4% achieved by all models with AugBERT as compared to the original training dataset. It is also found that the AugEDA gives maximum improvement of around 4% when Word2Vec and Doc2Vec embeddings were employed with LR. Also, there is negligible improvement in the results with AugBT.

4.3.2 Experimental Results for SDCNL

In this Section, the results of the experimental study are presented for the SDCNL dataset. As observed from Table 4, the average improvement of around 2.3% is observed for all the models as per F1 score with AugBERT. The AugEDA shows maximum improvement of more than 5% when Word2Vec and Doc2Vec embeddings were employed with RF. The results also indicate a minor improvement of around $1-2\%$ when classifiers employed Doc2Vec and TF-IDF embeddings for representing augmented data using Back Translation.

Due to increase in the size of augmented data, the input vector representations using TF-IDF requires higher computational time as compared to other embeddings. Thus, a few results are shown empty in Table 3 and Table 4. In healthcare, more *precise* results are expected than *recall* which means that the content which is identified as stressful must be correct and matters more than diagnosing the total number of correct instances. Thus, precision must improve more than recall values. We have considered these nuances to examine the results of classifiers and found that Logistic Regression gives improved results with the Doc2Vec encoding scheme.

4.4 Data Diversity of Augmented Data

The diversity of the generated data by different augmentation techniques are measured by the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002). The BLUE score ranges between 0 and 1. The lower the value, the better is the diversity in the data. Thus, the BLEU score is computed by comparing n-grams of both original and generated text where $n = 2$.

As observed from Table 5, the BLEU score for augmented data varies from 82% - 99%. The training samples are multiplied by 1.75 to 2.0 times for data augmentation approaches. The data for AugBERT is more diversified and thus, the results are significantly improved for AugBERT rather than AugEDA and AugBT as evident from Table 3 and Table 4. The experimental results show that the samples are upto 18% more diverse than those of original training samples for AugBERT over the Dreaddit dataset. However, the least data diversity is observed for AugEDA and AugBT over the SDCNL dataset.

4.5 Statistical Significance

In this Section, to understand the importance of generating more instances in training data is performed using three different data augmentation techniques. The statistical student's t-test was used to test the significance of the improvement in classifier using augmented data with $p-value$ as 0.05, 0.10, and 0.15. The resulting value for t-test in Dreaddit and SDCNL over AugBERT is obtained as 0.00033 and 0.09241 which shows the overall significant improvements with 5% and 10% significant levels, respectively. The results are improved in 83%, and 66% in the cases of different encoding vectors and classifiers which are used as learning based algorithms for AugBERT and AugEDA data augmentation techniques, respectively.

[6] https://scikit-learn.org/stable/

[7] https://github.com/varunkumar-dev/TransformersDataAugmentation

Methods used	Original			AugBERT			AugEDA			AugBT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RF+Word2Vec+TFIDF	0.68	0.84	0.75	0.69	0.84	**0.76**	0.68	0.81	0.74	0.67	0.84	0.74
SVM+Word2Vec+TFIDF	0.69	0.75	0.72	0.69	0.79	**0.74**$^{+}$	0.69	0.79	**0.74**$^{+}$	0.74	0.66	0.69
LR+Word2Vec+TFIDF	0.71	0.78	0.74	0.71	0.79	**0.75**	0.72	0.79	**0.75**	0.71	0.77	0.74
RF+Doc2Vec	0.65	0.78	0.71	0.62	0.83	0.71	0.65	0.80	**0.72**	0.63	0.82	**0.72**
SVM+Doc2Vec	0.74	0.76	0.75	0.73	0.78	**0.76**	0.74	0.73	0.73	0.73	0.75	0.74
LR+Doc2Vec	0.73	0.75	0.74	0.73	0.78	**0.76**$^{+}$	0.72	0.73	0.72	0.72	0.76	0.74
RF+Word2Vec+Doc2Vec	0.85	0.67	0.75	0.67	0.86	0.75	0.68	0.86	**0.76**	0.66	0.84	0.74
SVM+Word2Vec+Doc2Vvec	0.75	0.68	0.71	0.70	0.74	**0.72**	0.73	0.71	**0.72**	0.71	0.74	**0.72**
LR+Word2Vec+Doc2Vec	0.76	0.69	0.72	0.71	0.76	**0.73**	0.75	0.78	**0.76**$^{+}$	0.71	0.76	**0.73**
RF+TFIDF	0.70	0.73	0.71	0.69	0.79	**0.74**	0.67	0.84	**0.74**$^{+}$	-	-	-
SVM+TFIDF	0.79	0.68	0.73	0.74	0.78	**0.76**$^{+}$	0.70	0.73	0.72	-	-	-
LR+TFIDF	0.68	0.82	0.74	0.70	0.80	**0.75**	0.76	0.75	**0.75**	-	-	-

Table 3: Classification Results on Dreaddit Dataset: Precision(P), Recall(R), F-measure(F1) score on the Original and Augmented Datasets using BERT, EDA and BackTranslation. Text in bold shows the maximum F1 score achieved by the model. '-' indicates no results. '+' indicates significantly different results using statistical t-test.

Methods used	Original			AugBERT			AugEDA			AugBT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RF+Word2Vec+TFIDF	0.68	0.67	0.67	0.69	0.69	**0.69**$^{+}$	0.63	0.66	0.65	0.65	0.68	0.67
SVM+Word2Vec+TFIDF	0.69	0.67	0.68	0.65	0.66	0.66	0.67	0.70	0.69	0.63	0.67	0.65
LR+Word2Vec+TFIDF	0.63	0.70	0.66	0.67	0.73	**0.70**$^{+}$	0.65	0.69	**0.67**	0.61	0.67	0.64
RF+Doc2Vec	0.65	0.57	0.61	0.63	0.51	0.56	0.64	0.55	0.60	0.59	0.57	0.58
SVM+Doc2Vec	0.65	0.66	0.65	0.66	0.73	**0.70**$^{+}$	0.66	0.70	**0.68**$^{+}$	0.65	0.69	**0.67**$^{+}$
LR+Doc2Vec	0.65	0.67	0.66	0.68	0.76	**0.71**$^{+}$	0.68	0.71	**0.69**$^{+}$	0.66	0.69	**0.68**$^{+}$
RF+Word2Vec+Doc2Vec	0.63	0.64	0.63	0.63	0.58	0.60	0.67	0.69	**0.68**$^{+}$	0.66	0.67	**0.67**$^{+}$
SVM+Word2Vec+Doc2Vec	0.65	0.67	0.66	0.64	0.70	**0.67**	0.60	0.66	0.66	0.62	0.64	0.63
LR+Word2Vec+Doc2Vec	0.64	0.64	0.64	0.64	0.73	**0.68**$^{+}$	0.59	0.65	0.62	0.61	0.65	0.63
RF+TFIDF	0.61	0.85	0.71	0.63	0.81	0.71	-	-	-	-	-	-
SVM+TFIDF	0.71	0.75	0.73	0.67	0.85	**0.75**$^{+}$	0.76	0.73	**0.75**$^{+}$	0.71	0.77	**0.74**
LR+TFIDF	0.70	0.77	0.73	0.68	0.85	**0.76**$^{+}$	0.76	0.75	**0.75**$^{+}$	0.71	0.77	**0.74**

Table 4: Classification Results on SDCNL Dataset: Precision(P), Recall(R), F-measure(F1) score on Original and Augmented Datasets using BERT, EDA and Back Translation. Text in bold shows the maximum F1 score achieved by the model. '-' indicates no results.'+' indicates significantly different results using statistical t-test.

	Dreaddit	SDCNL
AugEDA	0.97	0.99
AugBERT	0.82	0.97
AugBT	0.88	0.99

Table 5: Data Diversity using BLEU Score

Dreaddit	AugBERT	AugEDA	AugBT
t-test	-4.69041	1.07605	0.75593
p-value	0.00033	0.15247	0.23568

Table 6: Statistical Significance of overall results with Original Data

4.5.1 Statistical Significance for Dreaddit

It is evident from Table 6 that AugBERT and AugEDA show significantly improved results and there is no effect of AugBT over domain-specific data augmentation for mental health.

On drilling down the results, it is observed that the AugBERT based augmented results for SVM classifier are significantly better than the other classification techniques. Some more significant improvements with the use of LR classifier is observed as shown in Table 3 with as high as 5% for AugEDA. The variation of improvement in results ranges upto 4.1%, 5.5% and 1.3% for AugBERT, AugEDA and AugBT, respectively.

4.5.2 Statistical Significance for SDCNL

The significant improvements over SDCNL dataset is observed on the basis of $p - value$ as 0.05, 0.10 and 0.15 as shown in Table 7. The results have shown that the AugBERT and AugEDA gives better results for 10% variation in results and validates the hypothesis that the augmented data gives significant improvements over the original dataset.

SDCNL	AugBERT	AugEDA	AugBT
t-test	-1.42426	-1.6361	0.25118
p-value	0.09241	0.06644	0.40338

Table 7: Statistical Significance of overall results with Original Data

Similar to the Dreaddit observations, the significant improvements with LR classifier are observed for classifying mental health into clinical depression and suicidal tendencies. On the contrary, SVM with Doc2Vec shows much better results with AugBERT, AugEDA and AugBT.

5 Conclusion

In this work, we use the data augmentation approach for mental health classification on two different social media datasets. The experimental results using Logistic Regression classifier and Doc2Vec embedding shows significant improvements in F1 score and Precision with AugBERT. To tackle the problem of data sparsity and support the automation of the 3-Step theory over social media data (Klonsky and May, 2015), the data augmentation over mental healthcare may give remarkable results. In future, we are planning to use other domain-specific libraries and neural machine translation for explainable and conditional data augmentation.

References

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2021. Understanding the impact of covid-19 on online mental health forums. *ACM Transactions on Management Information Systems (TMIS)*, 12(4):1–28.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Hannah Chen, Yangfeng Ji, and David Evans. 2020. Finding friends and flipping frenemies: Automatic paraphrase dataset augmentation using graph theory. *arXiv preprint arXiv:2011.01856*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *arXiv preprint arXiv:2106.07499*.

Jean-Philippe Corbeil and Hadi Abdi Ghadivel. 2020. Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. *arXiv preprint arXiv:2009.12452*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Muskan Garg. 2021. Quantifying the suicidal tendency on social media: A survey. *arXiv preprint arXiv:2110.03663*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Demi Guo, Yoon Kim, and Alexander M Rush. 2020. Sequence-level mixed sample data augmentation. *arXiv preprint arXiv:2011.09039*.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. *arXiv preprint arXiv:2102.09427*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Proceedings of the 2020 conference on empirical methods in natural language processing: findings*, pages 3774–3788.

Yeen Huang and Ning Zhao. 2020. Generalized anxiety disorder, depressive symptoms and sleep quality during covid-19 outbreak in china: a web-based cross-sectional survey. *Psychiatry research*, 288:112954.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes.

E David Klonsky and Alexis M May. 2015. The three-step theory (3st): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy*, 8(2):114–129.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2021. Data augmentation using pre-trained transformer models.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. *arXiv preprint arXiv:2004.14769*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Daniel Preoţiuc-Pietro, Maarten Sap, H Andrew Schwartz, and Lyle Ungar. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 40–45.

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

Nader Salari, Amin Hosseinian-Far, Rostam Jalali, Aliakbar Vaisi-Raygani, Shna Rasoulpoor, Masoud Mohammadi, Shabnam Rasoulpoor, and Behnam Khaledi-Paveh. 2020. Prevalence of stress, anxiety, depression among the general population during the covid-19 pandemic: a systematic review and meta-analysis. *Globalization and health*, 16(1):1–11.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Mike Thelwall. 2017. Tensistrength: Stress and relaxation magnitude detection for social media texts. *Information Processing & Management*, 53(1):106–121.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *LOUHI@EMNLP*.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6):e228.

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.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021. Conditional adversarial networks for multi-domain text classification. *arXiv preprint arXiv:2102.10176*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

A Appendix

Samples of original and augmented data

Original data	Augmented data (AugEDA)	Augmented data(AugBERT)	Augmented data(AugBT)
He said he could run some more tests, but he didn't think it would help.	he talk about my said run tests but he didnt think would help	he said he tried run some more medicine, but he weren't think it would help.	he said he could run some clinical tests, but he didn't think it would okay.
Is always adamant about keeping contact with the people she cheated with.	is always headstrong about contact with the people me cheated with	then rather adamant about keeping contact featuring the people she cheated with	is always adamant by keeping track featuring strange people she cheated with
It seemed like a circulation problem, and I panicked and of course ended up in the ER again.	it seemed as a circulation problem i panicked and of course finish up in er again	many said like a relationship collapsed, and i, and same course ended up entering the er again	it seemed had mostly circulation problem, and i panicked and of course ended up in er again.