Part I: Original BTP report

Transfer Learning for the Detection of Depression and Suicidal Ideation Using BERT

A B. Tech. Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Ahlaam Rafiq
(190121061)

under the guidance of

Professor Girish Sampath Setlur

to the

DEPARTMENT OF PHYSICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Transfer Learning for the Detection of Depression and Suicidal Ideation Using BERT**" is a bonafide work of **Ahlaam Rafiq** (**Roll No. 190121061**), carried out in the Department of Physics, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Professor Girish Sampath Setlur**

Professor,

Department of Physics,

Indian Institute of Technology Guwahati,

Nov, 2022

Guwahati.

Assam.

Acknowledgements

I am grateful to Professor Girish S. Setlur for providing me with the opportunity to work in an area I am passionate about, and for his constant guidance throughout the course of the project. I am also grateful to his research team for providing me with very helpful feedback. I would also like to thank the Department of Physics, IIT Guwahati, for providing me with all the necessary support and resources.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Depression is a common mood disorder affecting millions of us everyday - a number that has only shot up during the pandemic. Not all depressed persons are suicidal; however, research indicates that most (possibly at least 90% of) people who have died by suicide have suffered from mental disorders.[53, 6, 11], Further, the risk of suicide is elevated for individuals with several mental disorders, such as depression, schizophrenia, bipolar disorder, and alcoholism.[34] It is, therefore, imperative to be able to identify both depression and suicidal ideation in posts by both individuals with and without mental disorders; in order to ensure at-risk individuals receive timely care. Manual inspection of posts - which can be overwhelmingly large in number, especially when made on social media forums - can be both painstaking and fatally slow in emergency situations.

This project researches the performance of the state-of-the-art natural language processing model, BERT, in classifying "depressive" ideation from "suicidal" ideation in Reddit posts. We find that our basic BERT base model (without any optimizations) outperforms the classical Naïve Bayes and SVM classifiers across all metrics, with a best F1-score of 0.93 (macro-averaged over both classes). Using a relatively small primary dataset (under 2,000

sequences combined), we also find that BERT performs reasonably well in a low resource setting. This study aims to ultimately optimize our models to improve performance scores, and transfer this learning to be able to detect depression and suicidal ideation in posts made by users with various mental illnesses.

1.1 An introduction to mental health and its importance

Defining mental health is not an easy task. An introduction can perhaps be borrowed from the World Health Organization, which defines mental health to be "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community." [49]

The past few years have seen a worldwide recognition of the importance of mental health - be it for overall physical health, community well-being, or economic stability - resulting in a much-needed (albeit relatively small) shift of resources to focus on mental healthcare. There is still, however, a long way to go. Mental health may not be quantifiable, but its numbers are.

1.1.1 Case study: Mental health figures, India

India's increased budget allocation to mental health (2.18%)[**?**] this year serves as a reminder of its poor track record. A few stark figures to consider:

- In 2019, one in every seven Indians was living with a mental health disorder[13]
- In 2017, the contribution of mental health disorders to the national disease burden nearly doubled from 1990.
- The WHO estimates that India will lose over one trillion USD due to the same between 2012-2030,

- The suicide mortality rate (per 100,000 population) in India is higher (16.3) than the global average (10.5) or even the average for Southeast Asia (13.4). [48]
- India is severely understaffed when it comes to mental health workers, with numbers of psychiatrists, nurses, psychologists, social works, therapists, etc. falling far below global and regional averages.[48]

,

It is evident that India's mental health infrastructure is poor, and generally lags behind the rest of South East Asia and the world. A 2019 study [26] estimates that India was then short of 27,000 doctors - requiring 2,700 new psychiatrists (keeping population growth and attrition rates at 0%) annually to fill the gap in the next 10 years - and notes that there are only 700 psychiatrists trained every year in post graduate seats. This gap is also not uniform across the country, and while there has been a an increase in psychiatric facilities in a few states, some states have seen a stagnation or even decline.

The ongoing COVID-19 pandemic has also worsened the current mental health crisis: the Indian Psychiatry Society reports a 20% increase in mental illness cases since the pandemic [45]. In addition, mental health disorders and seeking treatment for the same continue to be shrouded in stigma. A systematic review on the stigma associated with mental health problems among young people in India [24] also finds that one third displays poor knowledge of mental health problems and negative attitudes towards people with mental health problems and one in five has actual/intended stigmatizing behavior.

There is still, thus, a lot of ground to cover. And while there is no tailor-made solution, there have been attempts to streamline the process by involving technology, Numbers provide support: while Indians might lack mental health facilities, their access to technology and the internet is impressive. To give a rough idea about access, an estimated 844.84 million Indians were smartphone users and 107.81 million households had internet access at home in 2021 [56, 31]. As of 2020, India was the world's second-largest internet population at

over 749 million users in 2020, of which 744 million users accessed the internet via their mobile phones [55]. Technology, thus, seems to be primed for utilization to meet the mental health needs of people. A 2015 World Health Organization (WHO) survey of 15,000 mobile health apps revealed that 29% of them emphasized on mental health diagnosis, treatment, or support. [5]

But for a field as intricate as mental health, we need to refine the search for greater effectiveness. This is where artificial intelligence (AI) and its applications can play a major role.

1.2 Artificial intelligence in mental health

AI systems aim to think humanly and act humanly with the ultimate goal of obtaining rational outcomes. [32] It's a field that has seen massive growth in research and development in the past few decades. While not a perfect science by any means, its data-driven and knowledge-based methods - making an extensive use of various kinds of knowledge specific to the domain - have been responsible for significant progress in multiple fields, including expert systems, natural language processing, speech recognition, computer vision, and robotics.[15]

Machine learning is a branch of AI, focusing on using data and algorithms to imitate the way humans learn. Our project focuses on natural language processing (NLP), a field that - standing at a junction of linguistics, computer science, and AI - studies the interactions between computers and human language. NLP remains one of the biggest, and most common, practical applications of machine learning. From an economic point of view, AI either decreases the costs of prediction or improves the quality of predictions available at the same cost. [2]

Its application to mental health is multifold, with research supporting its use as an independent tool or clinical aid in therapy, training, screening, self-management, counseling, and

diagnosing. AI applications in healthcare have previously been found to improve life quality for citizens and efficiency of governance [36], and there's a lot of potential for research in applying it to mental health for reducing the burden on healthcare providers and ensuring equitable access.

1.2.1 Data types and sources

Machine learning is built on data collection and analysis. Choosing, and sourcing, data specific to the mental health domain in question thus opens up challenges in itself.

The "data" in themselves are pretty varied - machine learning enjoys the flexibility of working of data of different modalities. AI-based technologies in psychiatry rely on the identification of specific patterns within highly heterogenous multimodal sets, including: various psychometric scales or mood rating scales, brain imaging data, genomics, blood biomarkers, data based on novel monitoring systems (eg. smartphones), data scraped from social media platforms, speech and language data, facial data, dynamics of the oculometric system, attention assessment based on eye-gaze data, and various features based on the analysis of the peripheral physiological signals (eg. respiratory sinus arrhythmia, startle reactivity).

In this project, I focus on textual data scraped from Reddit and apply natural language processing methods to detect mental health problems (depression and suicidal ideation) in a variety of tasks. We choose text from a non-clinical source (here, online forums) due to its ready availability and extraction. Non-clinical textual data (social media, online forums, instant messaging, etc.) have been, in particular, a hotbed of research for their potential in mapping mental health.

1.3 Machine learning for text classification

Text classification is a classical NLP (natural language processing, a field of machine learning) problem that aims to assign labels to text objects. Background and related work has been

explored in detail in the next chapter.

This project is further strongly centred around transfer learning, which uses the knowledge learned while solving one task and applies to solve another related task.

- Transfer learning is critical to the implementation of BERT (Biderectional Encodeer Representations from Transformers) - a state-of-the-art language model for NLP used throughout the project, and
- We aim to transfer the classification learned in RQ1 and RQ3 to solve RQ4 (see section 1.4).

1.4 Research goals

Briefly, this project aims to solve the following broad research questions:

- RQ1: Use an optimized BERT-based model to classify text in a low-resource setting with depressive sentiment from text with suicidal ideation (two labels: "depressed" or "suicidal").
- RQ2: Compare the performance of BERT-based models with baseline classical machine learning methods, including Naive Bayes and SVM classifiers.
- RQ3: Extend the classification task to include a third, control "neither suicidal nor depressed" label, again using an optimized BERT-based model.
- RQ4: Use transfer learning to classify mental health disorder-specific textual data into the aforementioned three labels.

Chapter 2

Background and related work

This section contains a brief overview of the background of the project and related work. We cover (a) machine learning for text classification, (b), the application of text classification for detection of mental health disorders, (c) transfer learning and its success in low-resource settings, and (d) an additional literature review of the different data sources used for the application of machine learning for mental health solutions.

2.1 Machine learning for text classification

Text classification is a popular machine learning (or more specifically, natural language processing) task that entails categorizing text into organized groups that are defined using labels. "Text" can include sentences, paragraphs, documents, etc. Typical text classification problems include sentiment analysis, natural language inference, question answering, etc.

While text classification can be performed through manual annotation, the increasing scale of data calls for automatic labeling. This can be performed by either rule-based methods (which classify text based on a set of predefined, domain-specific rules) or by machine-learning based methods (which are data- and observation-driven, and try to capture the

inherent relationship between text and label).

Machine learning methods for text classification can largely be divided into two types:

1. Classical models, such as Naïve Bayes, random forests, support vector machines (SVMs) and gradient boosting trees. These models first extract features from the text - common features include bag of words (BoW) and term frequency-inverse document frequency (tfidf) - and then feed these features into a classifier. Classical models, although widely used, have the same limitation as rule based methods: substantial domain knowledge is required to choose appropriate features, thus making it hard to apply learned models for cross-domain tasks. We currently use classical Naïve Bayes and SVM classifiers as our baseline for our classification task.

2. Neural network approaches, which replace extracted features with an embedding model that uses machine learning to map text into a continuous, low-dimensional feature vector. There are several deep learning models with varying architectures built for text classification, including RNN-based models, CNN-based models, graph neural networks, and Transformers. Transformers form the basis for Devlin et al's BERT (Bidirectional Encoder Representations from Transformers) [18], which as of 2021 is the state-of-the-art embedding model and has outperformed classical methods in multiple NLP problems. Based on BERT's excellent performance on other text classification tasks, we hypothesize that BERT can successfully separate suicidal ideation from non-suicidal depression and detect the two in test data, and do this more accurately than classical methods. BERT has been explained in greater depth in the following chapter.

2.2 Text classification for mental health

We reviewed 60 papers on the applications of machine learning in mental healthcare. We found a variety of data types and sources used across the research, discussed in brief in section 2.4. Here we focus on textual data.

Sources of textual content may include transcriptions of clinical interviews or sessions and non-clinical text (such as social media, online forums, instant messaging). Language and voice can be incredibly useful as clinical text data. For example, natural language processing analysis is used [12] to discriminate speech in psychosis from normal speech. This involves preprocessing of transcripts, latent semantic analysis, part-of-speech tagging analysis, and then ML classification and validation. Similarly, speech features extracted from phone conversations can be used for the classification of bipolar disorder episodes. [4] Other examples of clinical sources are clinical assessments using questionnaires, self-assessments, and clinical records. Electronic Medical Records or EMRs (containing medical history from individual clinical practices) and Electronic Health Records (EHRs) (containing comprehensive long-term history collecting multiple EMRs) feature in several studies and are of special importance - In France, for example, emergency department EMRs have been used for automated surveillance of suicide attempts [42]; a UCLA study uses EHRs to differentiate risk of suicide attempt and self-harm after general, recurrent medical hospitalization of women with mental illness [19]; and the UK Clinical Record Interactive Search system has provided de-identified information sourced from EHRs for identifying suicidal ideation and attempts [22].

Non-clinical data finds a major source in social media, which has particularly been a hotbed of research for its potential in mapping mental health. A very brief search shows us that Facebook posts can predict depression in medical records [20], analyzing tweets can help estimate the effects of exercise on mental health [35], and a user's Instagram profile can detect major depressive disorder [52]. Most studies use purely textual data, others may also include features regarding the user's profile (such as accounts followed, number of posts, etc.) and their activity. For a low-resource study like this one, sourcing data from social media is attractive because it is potentially free, relatively less time-consuming, and can be abundantly sourced anywhere. Limitations include the introduction of bias due to participants' voluntary responses, the lack of clinical validation, and the impossibility of offering

personalized treatment.

Weighing the benefits and limitations, however, we find social media - specifically, Reddit forum posts - to be the most appropriate sources for our use case. Reddit is a popular online network or forum of communities - i.e., "subreddits", that are built around a single focal subject - where registered users can post multimodal content and interact with other users and their posts. As of 2022, it has 330 million monthly active users. [30] Several studies we reviewed have used subreddit data for text classification in the mental health domain. Reddit offers its users anonymity and security, hence allowing users with various mental illnesses to discuss their experiences without the fear of being stigmatized. [10] Unlike other popular, anonymous social media forums (such as Twitter), it does not cap post length, thus allowing us to build a dataset with a substantial amount of text.

Jiang et al. used a large-scale dataset of Reddit posts containing data from users with eight disorders, and a control user group [61]. They built strong classifiers using deep, contextualized word representations and concluded that these vastly outperformed previously applied statistical models with simple linguistic features. One such model uses BERT encoding and an attention-based classifier. The authors compared F1- and accuracy scores with a non-contextual baseline text analysis program classifier that maps words to psychologically motivated labels. Results across all labels/disorders vary between 0.67-0.8 for the non-contextual classifier and 0.8-0.9 for BERT, which is a significant improvement. BERT was also used by Nisa et al. for the early detection of self-harm by Reddit users [60], with improved recall over baseline logistic regression. El-Ramly et al. implemented CairoDep [21], built using BERT Transformers, to detect depression in Arabic posts. CairoDep achieves high accuracy, precision, recall, and F1-scores (above 96%) as compared to lexicon-based or shallow machine-learning models (80% or below).

A preceding work that is highly relevant to our study is Haque et al.'s *Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction* [9], which uses

(among others) BERT-based models to classify depression from suicidal ideation, as we do. Their model is further optimized using label correction and a control dataset (optimization remains a future area for us to explore).

Our review indicates that learning context greatly improves a classifier's performance on a mental health dataset. Further, BERT is among the state-of-the-art models available for the task. Lastly, Reddit is a competent source of data for our research goals.

Chapter 3

An overview of BERT

3.1 Self-supervised learning: a brief introduction

We begin by introducing self-supervised learning (SSL). SSL acts as an intermediate between supervised and unsupervised learning. It is like humans in the sense that it depends on previously acquired background knowledge of how the world works to make decisions. As researchers at Meta AI put it: "... self-supervised learning is one of the most promising ways to build such background knowledge and approximate a form of common sense in AI systems".[3]

SSL methods are built around two primary steps in neural network training: first, "pretraining" (neural network training is started with a pre-trained model), followed by task-specific "fine-tuning". Because pre-training needs massive amounts of annotated data (such as learning a language for natural language processing tasks), a pretrained model for the required task may not exist. SSL solves this problem by looking for "labels" that are naturally part of the input data rather than requiring separate external labels. The general technique is to predict an unobserved property of the input from an observed part – for example, masked language modelling in natural language processing (described in detail further) or predicting

hidden future/past frames from the current frame in a video. Without relying on labels, thus, SSL can use a variety of supervisory signals across large datasets. So, essentially, it is still "supervised" in a way – just not by humans.

The need for SSL in NLP arises from the fact that it is very difficult for an AI system to grasp the intricacies of text and language from available supervised training data. Consider a binary classification problem. To achieve high accuracy, we would like to train our model on a large train dataset – and this would need to be labelled by hand. Because of human capacity, there is a limit to how large this dataset could be.

However large this figure may appear, it is too small for a model to effectively learn the subtleties of language. This is where our model can benefit from self-supervised, transfer learning. What happens in SSL-based NLP models is this: the models are pre-trained on massive text corpuses with millions of datapoints (for example, BERT has been pre-trained on the English Wikipedia corpus with 2,500 million words and the BooksCorpus with 800 million words) – and these corpuses are unlabelled data! Given some specific tasks to look for, the model learns the underlying structure of text on its own. Pre-trained models are available off-the-shelf for most tasks. Researchers can take these models and "fine-tune" them on smaller, task-specific labelled datasets. This is supervised.

Today, SSL-based models are the state-of-the-art when it comes to NLP problems. One such incredibly popular model is the Bidirectional Encoder Representations from Transformers model, or BERT, explained in detail below.

3.2 BERT: Bidirectional Encoder Representations from Transformers

Language model pre-training was found to be effective for improving NLP tasks, both sentence- and token- level, and two strategies existed to apply pre-trained language repre-

sentations to downstream tasks [18] (i.e., supervised learning tasks that utilize a pre-trained model):

- *Feature-based*, with task-specific architecture and pre-trained representations as additional features, and
- *Fine-tuning-based*, where all pre-trained parameters and minimal task-specific parameters were fine-tuned.

Prior to BERT, NLP models were largely unidirectional (that is, tokens could only attend to previous tokens in the self-attention layers of the Transformer), and thus suboptimal. While other models (for example, bidirectional-CNNs) were introduced to take into account bidirectional context, BERT – introduced in 2018 by researchers at Google – was the first self-supervised deeply bidirectional system for pre-training NLP [18].

BERT pre-trains deep bidirectional representations from unlabelled text by jointly conditioning on both left, right context in all layers. It can be fine-tuned with just one extra output layer and can be used to model for tasks without major task-specific architecture modifications. Applications of BERT are far-ranging: it can perform on sentence-level (inference, paraphrasing, etc.) and token-level (named entity recognition, question-answering, etc.) tasks.

There are two stages in the BERT SSL framework:

1. **Pre-training**: The model is trained on unlabelled data over different pre-training tasks. This is expensive, but is only a one-time procedure.
2. **Fine-tuning**: The model is initialized with pre-trained parameters. All parameters are fine-tuned using labelled data from downstream tasks – and each downstream task has separate fine-tuned models despite being initialized with the same pre-trained parameters. This is the inexpensive stage.

BERT's mantra, therefore, is building a "unified architecture across different tasks".

3.2.1 BERT and transfer learning

There is an additional benefit to the general pre-training/ specific fine-tuning process. Due to a large number of parameters, training BERT from scratch leads to overfitting. Fine-tuning, on the other hand, takes the pre-trained model as a starting point and further trains on a relatively small dataset.

BERT is pre-trained using 2 unsupervised tasks [18]:

1. *Masked language modelling (MLM)*

   In the initial BERT implementation, the training generator chooses 15% token positions at random. The procedure is as follows: if the i^{th} token is chosen,

   - replace the i^{th} token with [MASK] 80% of the time,
   - replace the i^{th} token with a random token 10% of the time, and
   - leave unchanged 10% of the time.

   Then, a final hidden vector is chosen to predict the original token with cross-entropy loss. This involves predicting only the masked words, not reconstructing the entire input. Further, the underlying Transformer keeps a distributional contextual representation of every input token. It does not know which words it will be asked to predict or which have been replaced randomly. The more recent BERT implementation uses whole word masking instead of randomly selecting WordPiece tokens to mask, as the previous implementation did. Whole word masking works as follows: first, it masks all tokens corresponding to a word at once. The overall masking rate remains the same. Next, each masked WordPiece token is predicted independently.

2. *Next sentence prediction (NSP)*

   Pretraining for binarized NSP is useful to understand sentence relationships. When choosing sentences (i.e., two spans of text from the corpus) "A" and "B", for each

pre-training example:

- 50% of the time, B is the actual next sentence (label *isNext*)
- 50% of the time, B is a random sentence (label *isNotNext*)

The sample is such that the combined length less that or equal to 512 tokens.

Finally, the net training loss is the sum of the mean MLM likelihood and the mean NSP likelihood.

When it comes to fine-tuning, there are three ways to fine-tune [18]:

- Train the entire architecture on the dataset and feed the output to the softmax layer. Error is back-propagated through entire architecture, and pre-trained weights are updated based on the new dataset.
- Train some layers while freezing others – that is, train partially, retraining only higher layers.
- Freeze the entire architecture. That is, freeze all the layers, attach a few neural network layers and train this new model.

The underlying procedure for fine-tuning, however, remains the same. First, for each task, task-specific inputs and outputs are plugged into BERT. Next, all parameters are fine-tuned end-to-end. At the input, the sentences A and B from pre-training are equivalent to: sentence pairs (paraphrasing), question passage pairs (question answering), hypothesis premise (entailment), and text-label pair (text classification/ sequence tagging). At the output, token representations are fed into the output layer for token level tasks. A special representation token is fed into the output layer for classification.

For further optimization, the optimal hyperparameter values (hyperparameters include learning rate, number of training epochs, batch size, etc.) can be tuned and are task-specific.

Ablation studies on BERT reveal that [18]: Removing NSP as a pre-training task hurts

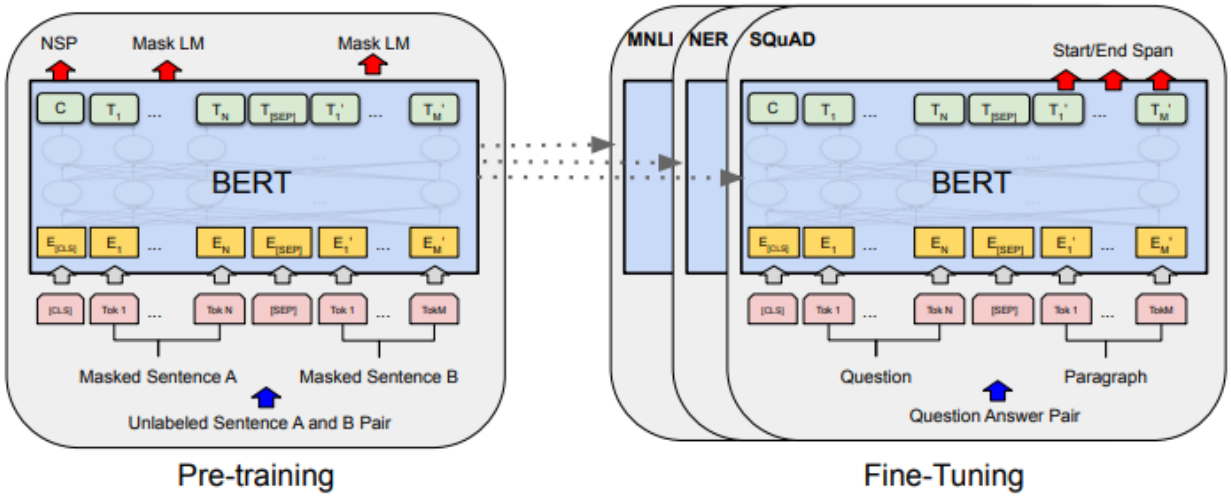


Fig. 3.1 Overall pre-training and fine-tuning procedures for BERT [18]. The same designs are utilised for pre-training and fine-tuning, with the exception of output layers. Models are initialised for various down-stream activities using the same pre-trained model parameters. All parameters are adjusted during fine-tuning. Every input example now has the special symbol [CLS] before it, and [SEP] is a special token for separation.

performance significantly on multiple benchmarks. BERT needs a large amount of pre-training to achieve high fine-tuning accuracy. Fine-tuning is robust to different masking strategies. The purpose of masking is to reduce mismatch between pre-training and fine-tuning. A larger model size leads to strict improvement in accuracy.

The BERT hypothesis [18], in conclusion, is as follows: *when the model is fine-tuned directly on downstream tasks, it uses only a very small number of randomly initialized additional parameters. Thus, task-specific models can benefit from larger, more expressive pre-trained representations even when the downstream task data is very small.*

There are multiple off-the-shelf BERT models available for use. We use the BERT-base (uncased) variant, which consists of 110M parameters and has been trained on the English Wikipedia and BookCorpus - that is, a total of 3.3 billion words.

Many improvements have been made to BERT over the years. Variants include:

- Small BERTS, which have the same general architecture as BERT but fewer or smaller

Transformer blocks,

- RoBERTa [38], a more robust implementation introduced as an "improved recipe" for training BERT models in response to BERT being found to be significantly under-trained,
- ALBERT [62], "A Lite BERT", that reduces model size without affecting computation time by sharing parameters between layers, and
- Electra [33], which has a setup resembling a generative adversarial network (GAN). Electra and BERT share the same architecture, but Electra is pre-trained as a discriminator.

To simplify our initial research, we continue to use the original BERT models over their optimized variants.

Chapter 4

Research experiments and results

Our initial research goal, completed in this study, is to train a BERT-based classifier to separate text showing suicidal ideation from those showing depression.

4.1 Dataset and data preprocessing

We choose Reddit as our data source for the following reasons:

1. Reddit is anonymous. Research indicates that this gives posters a sense of security and anonymity to discuss their experiences and struggles without the fear of being stigmatized or discriminated against.

2. Unlike Twitter and other platforms, Reddit has no limit on the size of a text post, thus providing a larger amount of *contextual* data available to our model.

3. As of 2021, Reddit has a 2.8 million subreddits [37] (smaller communities where people with shared interests interact). We found over a hundred subreddits devoted to specific mental health disorders or mental health in general. For example, in "r/depression" users talk about their struggles with depression; "r/BipolarReddit" is specific to bipolar disorder, "r/anxietyhelp" is for users with anxiety, "r/add" is for users with ADD, etc.

We find this especially useful for our future research work, where mental illness-specific data will be required.

4. Data posted publicly on Reddit is freely accessible and can be downloaded via the Reddit API.

4.1.1 Subreddit overview

For the current study, we downloaded data from the following subreddits:

1. *r/depression*: This community offers a peer-support space for users who struggle with, or know someone who struggles with, a depressive disorder. At the time of writing this paper, it has over 914,000 members.

2. *r/SuicideWatch*: This community offers a peer-support space for at-risk users struggling with suicidal thoughts. At the time of writing this paper, it has over 388,000 members

Using the Reddit Pushshift API, we analysed the top keywords of each subreddit. For r/depression, top keywords include: *self-harming*, *self-hatred*, *dejected*, *quietness*, *numbs*, *friendless*, and *later*. For r/SuicideWatch, top keywords include: *fantasise*, *offing*, *mourned*, *dead-end*, *unlivable*, *outlived*, *hurted*, and *clear-cut*.

While not immediately obvious, it is clear that there is *some* distinction between posts by users with depression and suicidal, at-risk users. Training a machine to automatically detect the difference can be of critical help in emergency settings, and can help provide emergency care to patients based on the severity of their textual responses.

4.1.2 Dataset

We scraped Reddit using the author's personal account as user agent. Most recent data for each subreddit was downloaded in 50 batches. Due to the scale the study and our goal to confirm the effectiveness of BERT in low-resource settings, we kept our dataset small.

- 1230 posts were scraped from *r/depression*, of which 979 were unique and retained.
- 1251 posts were scraped from r/SuicideWatch, of which 974 were unique and retained.

For our study, we only require the title of the post, the body of the post, and the subreddit name. All other information - such as number of comments, post awards, URL, date, post author - is discarded, effectively ensuring our data is de-identified. We concatenate the title of the post to its body and use this as our "text". The subreddit name is used to define the labels - if a post belongs to r/depression, it is automatically labeled as having *depressive* ideation and labeled 0; if a post belongs to r/SuicideWatch, it is automatically labeled as having *suicidal* ideation and labeled 1.

To reduce noise in the datasets, we perform basic data preprocessing by removing non UTF-8 characters and extra white spaces. Further typical NLP preprocessing (such as tokenization or stemming) is not necessary and taken care of by BERT's superior encoding mechanism. However, for comparison with classical machine learning models, we perform additional preprocessing steps: (1) all text is lowered, (2) word tokenization is performed, (3) stop words are removed, (4) word lemmatization is performed. We use the NLTK library for performing tokenization and lemmatization, and the NLTK list of English stopwords as our stopwords.

4.2 Training and evaluating the classifiers

We implement two classical machine learning models (section) and one deep learning model (section). After performing the required data preprocessing steps for each model, we split the data into train, test, and validation sets following a 5-fold cross validation method. The split is in a 70:15:15 ratio. Trained classifiers are tested on the test set and we assess the performance of each using 4 *performance metrics*: accuracy, recall (for both classes), precision (for both classes), and F1 (for both classes).

4.2.1 Classical machine learning models

In this study, we consider only two classical models due to limited resources. For implementation, off-the-shelf classifiers available in the *sklearn* library are used.

1. *Support Vector Machine (SVM)*, a linear model that classifies the dataset as "depressive" (0) or "suicidal" (1) by drawing a line (in higher dimensions, this would be a hyperplane) separating the data into either class. We use a linear kernel. [27]

2. *Naïve Bayes (NB)*, a simple probabilistic classifier that applies Bayes' theorem - i.e., it assumes that each feature contributes independently to class assignment or that that presence/ absence of a particular feature is independent of the presence/ absence of another feature. We use sklearn's NB classifier for multinomial models, that is suitable for classification with discrete features (in our case, word counts). [51]

We use fractional word counts (tf-idf/ "term frequency-inverse document frequency") as discrete features for classification.

4.2.2 BERT-based model

We use the uncased BERT base model, available off-the-shelf as *bert-base-uncased* [28] from Hugging Face, as our primary model. While a cased model is also available (with a higher number of parameters), we chose the uncased model because information about casing is not relevant to our study. It is, further, possible to pretrain the model from scratch to be domain-specific, but we choose not to as: (1) the available BERT models show excellent results across domains, and (2) we wish to reduce the risk of overfitting [59]. Classification is performed using a PyTorch backend with the Hugging Face *Transformers* library and the *AutoModelForSequenceClassification* class. *transformers.AutoModelForSequenceClassification* is a generic model class instantiated from the pretrained BERT model (instantiation also loads pretrained model weights) that carries a sequence classification head.

Our input dataset has the following class distribution: 979 sequences in class 0, and 974

sequences in class 1. Since the dataset is almost evenly balanced, we did not see a need for additional class balancing or evaluating more appropriate metrics.

We follow a 5-fold cross-validation technique to improve our metric accuracy - that is, we train five separate times, with the model seeing a different train/ evaluation/ test dataset each time. For each fold, the dataset is split into train, evaluation, and test datasets in a 70:15:15 ratio (stratified along labels, randomly resampled for each fold).

For training, evaluating, and testing, we use the *Transformers* library's *Trainer* class [29]. We choose the following hyperparameters (default - future work includes performing hyperparameter optimization):

- Evaluation strategy: "epoch". That is, the performance of the model is evaluated at the end of each epoch using the specified *evaluation metric*. Note that the *evaluation metric* is not the *performance metric* - the former is only used to update the model weights at the end of each epoch, the latter is used to check the performance of the final model on the test dataset.
- Learning rate: $2e^{-5}$. This is a common initial learning rate, also recommended by Devlin et al. [cite]
- Training and evaluation batch sizes: 4.
- Number of training epochs: 10
- Evaluation metrics: accuracy, precision (macro-average), recall (macro-average), F1-score (macro-average). We use macro-averages for overall evaluation.

4.2.3 Performance metrics

We evaluate performances across models using the following four metrics (here TP, TN, FP, FN represent numbers of true positives, true negatives, false positives and false negatives respectively):

1. **Precision**, which evaluates the ratio of test posts that have been correctly assigned to their target class. That is, $precision = \frac{TP}{TP+FP}$. Precision can take values between 0 and 1. 0 indicates a poor model and 1 indicates a perfect model. We calculate precision separately for each class and also a macro-average over both classes to represent overall performance.

2. **Recall**, which evaluates the ratio of test posts that actually belong to the predicted class (i.e. target class is the predicted class) to all posts predicted as belonging to that class. That is, $recall = \frac{TP}{TP+FN}$. Recall can take values between 0 and 1. 0 indicates a poor model and 1 indicates a perfect model. We calculate recall separately for each class and also a macro-average over both classes to represent overall performance.

3. **F1-score**, which calculates the harmonic mean of precision and recall, useful for combining precision and recall into a single measure (when used in isolation, precision and recall and not particularly informative). $F_1 = 2 \times \frac{precison \times recall}{precision+recall}$. F1 can take values between 0 and 1. 0 indicates an imperfect model and 1 indicates a perfect model with perfect precision and perfect recall. We calculate F1-scores separately for each class and also a macro-average over both classes to represent overall performance.

4. **Accuracy**, which calculates the ratio between the number of correct predictions to the total number of predictions. That is, $accuracy = \frac{TP+TN}{TP+TN+FP+FS}$. Accuracy can take values between 0 and 1. 0 indicates an imperfect model and 1 indicates a good model. It is a useful metric when both classes are balanced, such as in our case.

4.3 Results

The subsequent figures 4.1-4.4 plot the scores for each performance metric across all models. *Green* bars are used to represent BERT base, and *red* bars are used to represent classical models. In the figures,

- BERT/A represents BERT base evaluated by accuracy,
- BERT/F1 represents BERT base evaluated by F1-score,
- BERT/R represents BERT base evaluated by recall,
- BERT/P represents BERT base evaluated by precision,
- NB represents the Naive Bayes classifier, and
- SVM represents the Support Vector Machine classifier.

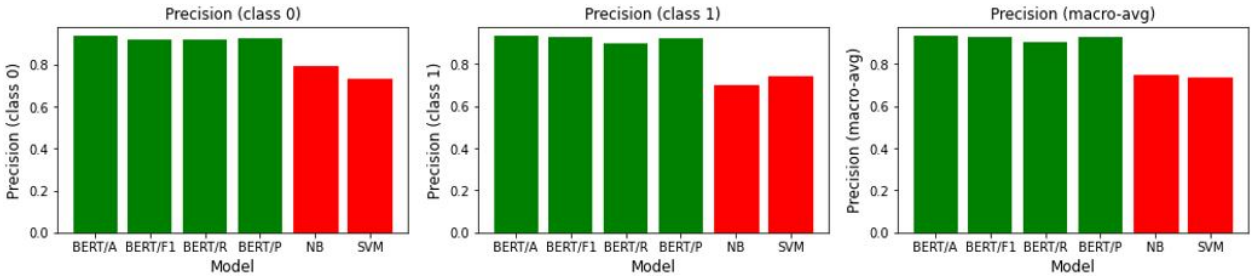


Fig. 4.1 Performance metric scores when model is evaluated on precision

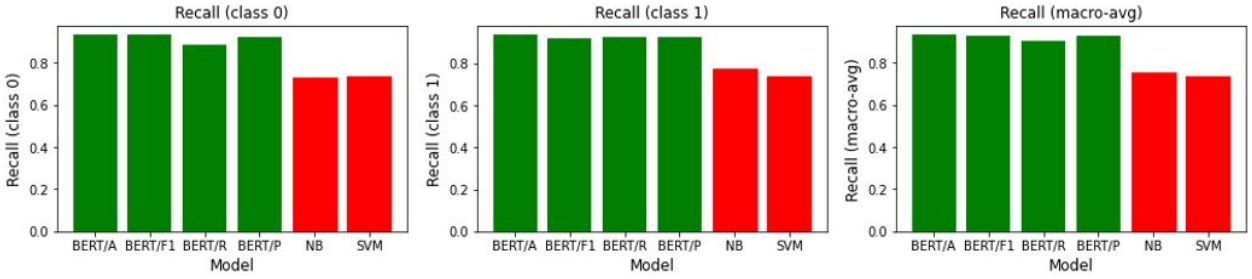


Fig. 4.2 Performance metric scores when model is evaluated on recall

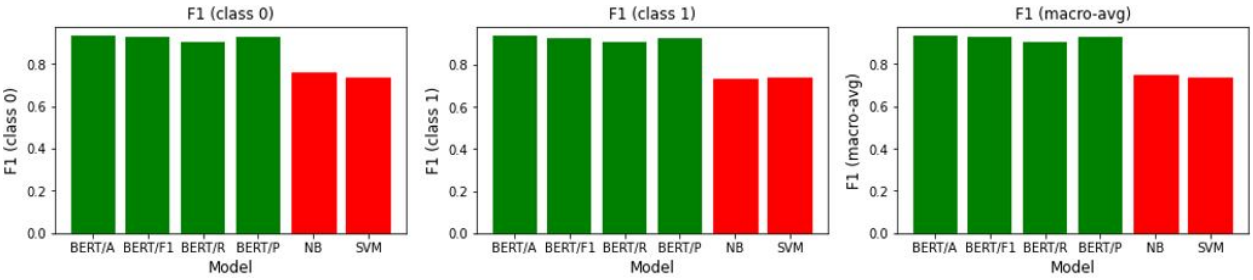


Fig. 4.3 Performance metric scores when model is evaluated on F1-score

The exact scores are given in table 4.1. "MA" represents the macro-average score over both classes.

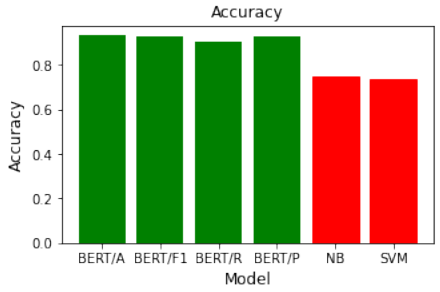


Fig. 4.4 Performance metric scores when model is evaluated on accuracy

4.4 Discussion

On observation, we conclude that BERT is the best classifier for separating posts showing depression from posts showing suicidal ideation. All our four experiments, which involve training the model by evaluating along different metrics, return performance scores for each metric (both macro-averages and scores for each class) above 0.90. The best scores across the board are obtained by evaluating BERT base along accuracy - all metrics (both macro-averages and scores for each class) attain their highest values (all ¿0.93). While classical models perform reasonably well - since no special measures have been taken to optimize them - they perform significantly worse that BERT in terms of all scores.

Model	**Evaluation metric**	**Precision**			**Recall**			**F1-score**			**Accuracy**
		0	**1**	**MA**	**0**	**1**	**MA**	**0**	**1**	**MA**	
BERT base	Accuracy	0.934	0.932	0.932	0.932	0.934	0.932	0.93	0.934	0.932	0.932
BERT base	F1	0.920	0.930	0.926	0.932	0.918	0.926	0.926	0.922	0.926	0.926
BERT base	Recall	0.916	0.896	0.906	0.888	0.922	0.904	0.900	0.908	0.904	0.904
BERT base	Precision	0.924	0.922	0.926	0.924	0.922	0.926	0.926	0.924	0.926	0.926
Naive Bayes	-	0.794	0.702	0.748	0.728	0.774	0.752	0.758	0.734	0.748	0.748
SVM	-	0.734	0.740	0.738	0.738	0.736	0.738	0.736	0.738	0.738	0.738

Table 4.1 All scores. *0* represents class 0, *depressive*. *1* represents class 1, *suicidal*. *MA* represents the macro-average score over both classes

Chapter 5

Conclusion and Future Work

In this initial stage of the project, we have successfully answered the first two of our research goals. We have shown that, even without optimization and in a low-resource setting, a BERT base model performs exceedingly well in classifying "depressive" posts from "suicidal posts". Further, we have shown that this model outperforms classical Naöve Bayes and SVM classifiers across four metrics: precision, recall, F1-score and accuracy for both "depressive"/0 and "suicidal"/1 classes - and their macro-averages, wherever applicable.

We aim to complete the following experiments or answer the following questions in our future work:

1. Add a control dataset (for instance, data from r/CasualConversation, a subreddit for users to have a friendly, casual conversation on any topic). How do results with a control dataset compare to results now? Further, how does performance compare with Haque et al's results?

2. Increase dataset size. There are two possibilities that can be explored and compared: (a) simply scraping more data, and (b) data augmentation methods, such as random swapping, random deletion, prepending, etc. How does performance differ in a higher-

resource setting?

3. In real life, it is likely that input data will be highly skewed. How capable is a model trained and evaluated on balanced datasets of performing well on an imbalanced test dataset? Does performance improve by introducing a corresponding skewness in the test and evaluation datasets? Class balancing methods such are random resampling may also be considered.

4. Perform hyperparameter optimization to improve overall scores.

5. Our final and primary research goal is to apply this learning to other mental-illness-specific data. For instance, our model should be able to identify "depressive" and "suicidal" ideation in posts by individuals with eating disorders. We believe this can find incredible use in emergency settings.

It's important to highlight, once more, how mental health is absolutely integral to society. It has taken a global crisis to come to terms with the fact that ignoring mental health, and surrounding it with stigma, is no longer an option. It doesn't just concern those with mental health disorders – it concerns all of us.

So far, overlooking mental health has had a huge toll on nearly every aspect of human life and wellbeing, from affecting families to severely costing the national economy. For society as a whole, mental, physical, and social health remain closely interwoven and equally vital. It's imperative, therefore, that mental health be accorded the same importance as, for example, physical health might. The accessibility and adaptability of AI – for all its challenges – does seem to be a promising solution.

Chapter 6

Appendices

6.1 Appendix 1: A brief note on Transformers

Transformers [7] are a network architecture based solely on attention mechanisms, hence doing away with recurrence and convolutions entirely. Because they are more parallelizable, require significantly less time to train, and draw global dependencies between input and output, Transformers are finding increasing use across machine learning problems.

Like LSTMs, Transformers are an architecture to transform one sequence to another, using two components: an encoder and a decoder. Unlike previous Seq2Seq models, however, it does not use any RNNs (GRU, LSTM, etc.). Due to this, there is no need to specify relative positions: positional encodings of different words are added to the embedded representation (this is an n-dimensional vector) of each word.

Model architecture comprises of an encoder and a decoder, both of which are stacks of modules mainly of multi-headed attention and feed-forward layers. The encoder takes the input sequence and maps it to a higher dimensional space. This n-dimensional vector is then fed into the decoder, which outputs a sequence.

While attention mechanisms were introduced prior to Transformers, multi-headed attention

(or self-attention) expands the model's ability to focus on different positions in the text. The attention mechanism looks at the input sequence and decides at each step which other parts of the sequence are important. For each input read in by the encoder, it takes into account several other inputs simultaneously and decides the importance of each by assigning weights.

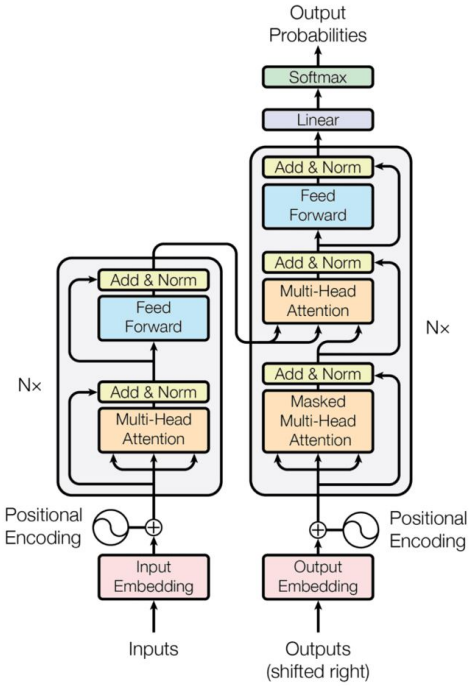


Fig. 6.1 Transformer model architecture

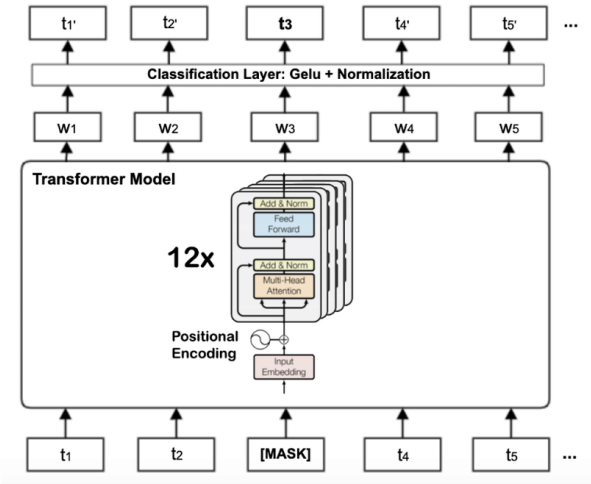


Fig. 6.2 BERT base model architecture, with 12 Transformer layers

Advantages of Transformers over LSTMs include [14]:

- The absence of locality bias due to multi-headed attention.
- A single multiplication per layer is performed - this is more efficient, as effective batch size is now the number of words, not the number of sequences.

The model used in our project, BERT-base, consists of 12 Transformer layers. Its architecture is show in the figure below.

6.2 Appendix 2: A brief note on other mental health data for AI research

The following table attemps to show India's stark mental health figures, as released by the WHO's 2017 Mental Health Atlas [48] (released every three years; India did not participate in the more recent 2020 Atlas). For comparison, we have included figures from the South East Asia region (SEAR), the European Region (EUR), Germany (a world leader in terms of mental health care), and the global average (GLB). Fields marked with "-" are indicative of unavailable data.

Figures	India	GLB*	SEAR*	EUR*	Germany*
Total mental health expenditure per person	4 INR/ ∼0.06 USD	2.5 USD	0.1 USD	21.7 USD	350.58 EUR/ ∼388 USD
Suicide mortality rate**	16.3	10.5	13.4	12.9	13.6
Disability-adjusted life years** (DALYs)	2,433.41	-	-	-	3603.56
Total number of mental health professionals (govt. and non-govt.)	25,312	-	-	-	118,367
Total mental health workers**	1.93	-	-	-	144.87
Number of psychiatrists**	0.29	1.3	0.4	9.9	13.20
Number of child psychiatrists**	0.00	<0.1	-	-	2.76
Total number of child psychiatrists	49	-	-	-	2.76
Number of other specialist doctors**	0.15	-	-	-	3.5
Number of mental health nurses**	0.80	3.5	0.80	23.2	-
Number of psychologists**	0.07	0.9	0.1	4.6	49.55
Number of social workers**	0.06	0.9	0.2	0.8	-
Number of occupational therapists**	0.03	-	-	-	56.43
Number of speech therapists**	0.17	-	-	-	19.41
Number of other paid mental health workers**	0.36	0.5	0.4	11.2	-

** Relative figures, wherever applicable*
** Rate per 100,000 population*

Table 6.1 Figures from the 2017 WHO Mental Health Atlas

Current research and data sources

In our literature review, 35 studies used interviews (and/or interview transcripts) and inventories/questionnaires/assessments (both self-report and clinician-administered) based on psychometric or mood rating scales. These are primarily used for 4 purposes: (1) for screening during participant recruitment, (2) for classification of participants into study subgroups, (3) for initial diagnosis, classification or scoring and follow-ups over the course of the study, and (4) for use as features for the machine learning model.

For implementation in India, the WHO's recommended process of translation and adaptation of instruments [47] may serve as a guide. Previous work on cross-cultural adaptation includes

the translation and validation of several SRIs into multiple Indian languages [58, 41, 50, 57, 17, 1, 25, 16].

EMRs and EHRs, mentioned previously in our discussion of textual sources, are extremely useful in addition to being rich in data for research and development. They ensure anytime/anywhere accessibility of patient records, improve the quality of records and are cost-effective, help track patients' clinical records and improve patient compliance, can be transferred easily within and across healthcare facilities, are easy to update, and facilitate improved healthcare decisions and provide evidence-based care [54]. For research specifically, they serve multiple uses. They can help in the recruitment, identification and screening of the study population and its medical history, and most importantly, they provide a large clinical database with multiple features that can be used for training, validation, and correlation.

India too has recognized the importance of digitizing - and thus improving the reliability of - its healthcare. The Ministry of Health and Family Welfare (MoHFW) first came out with standards for EHRs for India in September 2013. It also proposed to set up the National eHealth Authority (NeHA) in 2015, aiming to promote the setting up of state health records repositories and health information exchanges to facilitate interoperability. NeHA also looked to formulate and manage all health informatics standards for India. The MoHFW also put forward the Digital Health Information in Healthcare Security (DISHA) Act in 2018. DISHA has been drafted to "provide for the establishment of National and State eHealth Authorities and Health Information Exchanges; to standardize and regulate the processes related to collection, storing, transmission and use of digital health data; and to ensure reliability, data privacy, confidentiality and security of digital health data and such other matters related and incidental thereto" [46]. The same year, NITI Aayog proposed to create digital health records for all citizens by 2022 with the National Health Stack (NHS) [23]. In September 2021, the National Digital Health Mission (NDHM) - which will provide

every citizen with a health ID - was announced. Stored on the NHS, this ID will contain the individual's complete medical history and will be accessible by the entire healthcare industry. The NHS is basically an ecosystem of cloud-based services and has in the past couple years empanelled a number of private entities, including AI companies. A 2020 Centre for Sustainable Development paper on EHRs in India goes into further detail about the topic [39].

6.2.1 Challenges to adopting AI for mental health

A challenge to IoT-enabled AI applications (sensors, monitoring devices) is the interoperability of systems [8]. Since the IoT industry currently lacks technical standards, the variation in hardware and software leads to an inconsistent technology ecosystem. The systems may not, again, be interoperable with government infrastructure and this diversity could potentially cause problems with system maintenance and scalability.

A second problem is data privacy and security. AI applications, like traditional ones, are vulnerable to cybersecurity threats. Data consent is also an issue. Users may not be aware of (1) what data is collected, (2) how it is stored and processed and by whom, and (3) who is benefitting from this data. This is visible with the National Health Stack too. Researchers remain wary of the possible security concerns a data breach might bring up: as of October 2021, there were no standards set on data anonymisation or further use of this data by private entities [44]. Theft of medical identity is also a growing concern.

Clarifying data ownership also comes into question. The fact that AI applications are typically "black box" learning systems means that there may be a lot of ambiguity about their societal outcomes, and users' misgivings in their use is justified.

A third, vital challenge is ethics. On top of the existing inherent bias in machine algorithms, it's difficult to decide the limits of AI function. For instance, in a critical area such as mental health, can we code - and trust - chatbots to be reliable mandatory reporters? Can AI be

held accountable for its decisions? Further, because studies involve human subjects, data collection must also be meet ethical requirements.

A fourth challenge is environmental sustainability. Collecting, storing, and analysing the massive amounts of data required by AI applications will lead to significant consumption of energy and power,

And finally, there's the existing digital divide and the fact that struggling for "equitable access" will only worsen the conditions of a few. This was visible with the Aadhar process, and given the NHS's current infrastructure, it's a valid fear.

Further work on applying AI ethically has been detailed by NITI Aayog's 2021 Approach Document on "Responsible AI" [43]. A research agenda on AI for smart government [40] stresses on ensuring four principles to address challenges: transparency, accountability, fairness, and ethics.

To maximize the benefits of AI (for application in a general field), West and Allen [14] recommend the following steps: encouraging greater data access for researchers without compromising users' personal privacy; investing more government funding in unclassified AI research; promoting new models of digital education and AI workforce development so employees have the skills needed in the 21st-century economy; creating a federal AI advisory committee to make policy recommendations; engaging with state and local officials so they enact effective policies; regulating broad AI principles rather than specific algorithms; taking bias complaints seriously so AI does not replicate historic injustice, unfairness, or discrimination in data or algorithms; maintaining mechanisms for human oversight and control; penalizing malicious AI behaviour, and promoting cybersecurity.

References

[1] L. A. and C. A. Psychometric validation of geriatric depression scale - short form among bengali-speaking elderly from a rural area of west bengal: Application of item response theory.

[2] A. Agrawal. The economics of artificial intelligence. *McKinsey Quarterly*, 2018.

[3] M. AI. Self-supervised learning: The dark matter of intelligence. 2021.

[4] V. O. H. P. O. M. E. F. M. Alban Maxhuni, Angélica Muñoz-Meléndez. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*, 31:50–66, 2016.

[5] E. Anthes. Mental health: There's an app for that. *Nature*, 532:20–23, 2016.

[6] T. G. Arsenault-Lapierre G., Kim C. Psychiatric diagnoses in 3275 suicides: A meta-analysis. *BMC Psychiatry*, 4:37, 2004.

[7] N. P. e. a. Ashish Vaswani, Noam Shazeer. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[8] S. M. Atreyi Kankanhalli, Yannis Charalabidis. Iot and ai for smart government: A research agenda. *Government Information Quarterly*, 35:304–309, 2019.

[9] T. G. Ayaan Haque, Viraaj Reddi. Deep learning for suicide and depression identification with unsupervised label correction. 2021.

[10] Balani and . H.-W. e. a. . De Choudhury, 2015; Berry et al.

[11] F. A. Bertolote J.M. Suicide and psychiatric diagnosis: A worldwide perspective. *World Psychiatry*, 1:181–185, 2002.

[12] D. F.-S.-G. B. C. K. D. C. J. C. E. B. G. A. C. Cheryl M. Corcoran, Facundo Carrillo. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17:67–75, 2018.

[13] I. S.-L. D. B. I. M. D. Collaborators. The burden of mental disorders across the states of india: the global burden of disease study 1990–2017. *The Lancet Psychiatry*, 7:148–161, 2020.

[14] S. Cristina. The transformer model. *Machine Learning Mastery*, 2022.

[15] V. R. V. . V. D. Damioli, G. The impact of artificial intelligence on labor productivity. *Eurasian Bus Rev*, 11:1–25, 2021.

[16] S. T. e. a. De Man J., Absetz P. Are the phq-9 and gad-7 suitable for use in india? a psychometric analysis. *Frontiers in Psychology*, 2021.

[17] S. E. e. a. Desai N. D., Shah S. N. Validation of gujarati version of 15-item geriatric depression scale in elderly medical outpatients of general hospital in gujarat. *International Journal of Medical Science and Public Health*, 3:1453–1458, 2014.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] J. P. e. a. Edgcomb J. B., Thiruvalluru R. Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalizaton of women with mental illness. *Medical Care*, 59:S58–S64, 2021.

[20] M. R. M. e. a. Eichstaedt J. C., Smith R. J. Facebook language predicts depression in medical records. *PNAS*, 115:11203–11208, 2018.

[21] M. E.-R. et al. Cairodep: Detecting depression in arabic posts using bert transformers. *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 207–212, 2021.

[22] D. R. V. S. e. a. Fernandes, A.C. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci Rep*, 8: 7426, 2018.

[23] N. I. for Transforming India. National health stack. 2018.

[24] T. S. T. K. M. e. a. Gaiha, S.M. Stigma associated with mental health problems among young people in india: a systematic review of magnitude, manifestations and recommendations. *BMC Psychiatry*, 20:538, 2020.

[25] R. P. e. a. Ganguly S., Samanta M. Patient health questionnaire-9 as an effective tool for screening of depression among indian adolescents. *Journal of Adolescent Health*, 52: 546–551, 2013.

[26] C. P. S. Garg K., Kumar C. N. Number of psychiatrists in india: Baby steps forward, but a long way to go.

[27] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15:290–298, 2018.

[28] Huggingface. https://huggingface.co/bert-base-uncased. .

[29] Huggingface. https://huggingface.co/docs/transformers/main$_c lasses/trainer$..

[30] A. Hutchinson. Reddit now has as many users as twitter, and far higher engagement rates. *SocialMediaToday*, 2018.

[31] D. J. Internet households in india 2010-2025. *Statista*, 2021.

[32] H. J. A brief introduction to artificial intelligence.

[33] Q. V. L. e. a. Kevin Clark, Minh-Thang Luong. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv*, 2020.

[34] B. L. Suicide risk and mental disorders. *Int J Environ Res Public Health*, 2018.

[35] D. R. V. L. and C. A. Using matched samples to estimate the effects of exercise on mental health via twitter.

[36] D. Lee and S. N. Yoon. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *Int J Environ Res Public Health*, 18:271, 2021.

[37] Y. Lin. 10 reddit statistics every marketer should know in 2023 [infographic]. *Oberlo*, 2022.

[38] Y. Liu, M. Ott, and e. a. Goyal. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

[39] W. M. Ict india working paper 25: Electronic health records in india. 2018.

[40] W. D. M. and A. J. R. How artificial intelligence is transforming the world.

[41] B. M. e. a. Mehra A., Agarwal A. Evaluation of psychometric properties of hindi versions of the geriatric depression scale and patient health questionnaire in older adults. *Indian Journal of Psychological Medicine*, 43, 2021.

[42] G. Q. e. a. Metzger M., Tvardik N. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: A french pilot study. *Int J Methods Psychiatry Res.*, 26:1522, 2017.

[43] M. MK. Responsible ai aiforall. approach document for india part 1 - principles for responsible ai.

[44] M. MK. The risks of storing health records of 1.3 billion indians on the national health stack. *The News Minute*, 2021.

[45] A. K. . K. R. Nayar. Covid 19 and its mental health consequences. *Journal of Mental Health*, 30:1–2, 2021.

[46] M. of Health and I. Family Welfare. Digital information security in healthcare, act. 2018.

[47] W. H. Organization. Guidelines on translation and adaptation of instruments.

[48] W. H. Organization. Mental health atlas 2017. 2017.

[49] W. H. Organization. Mental health: strengthening our response. *WHO Newsroom*, 2022.

[50] S. T. V. Rajgopal J. and M. M. Psychometric properties of the geriatric depression scale (kannada version): A community-based study. *Journal of Geriatric Mental Health*, 6, 2019.

[51] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

[52] C. B. e. a. Ricard B. J., Marsch L. A. Exploring the utility of community-generated social media content for detecting depression: An analytical study on instagram. *Journal of Medical Internet Research*, 20, 2018.

[53] W. R. J. G. S. K. J. Robins E., Murphy G.E. Some clinical considerations in the prevention of suicide based on a study of 134 successful suicides. *Am. J. Public Health Nations Health*, 49:888–899, 1959.

[54] L. M. e. a. Robinson J., Witt K. Development of a self-harm monitoring system for victoria. *Int. J. Environ. Res. Public Health*, 20, 2020.

[55] K. S. Mobile internet users in india 2010-2040. *Statista*, 2021.

[56] S. S. Smartphone users in india 2010-2040. *Statista*, 2021.

[57] R. G. e. a. Sarkar S., Kattimani S. Validation of the tamil version of the short form geriatric depression scale-15. *Journal of Neurosciences in Rural Practice*, 6:442–1446, 2015.

[58] C. V. Thomas A. M. and A. A. Translation, validation and cross-cultural adaptation of the geriatric depression scale (gds-30) for utilization amongst speakers of malayalam; the regional language of the south indian state of kerala. *Journal of Family Medicine and Primary Care*, 10:1863–1867, 2021.

[59] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019.

[60] Q. un Nisa and R. Muhammad. Towards transfer learning using bert for early detection of self-harm of social media users. *Conference and Labs of the Evaluation Forum*, 2936.

[61] S. I. L. e. a. Zhengping Jiang. Detection of mental health conditions from reddit via deep contextualized representations. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, page 147, 2020.

[62] S. G. e. a. Zhenzhong Lan, Mingda Chen. Albert: A lite bert for self-supervised learning of language representations. *arXiv*, 2019.

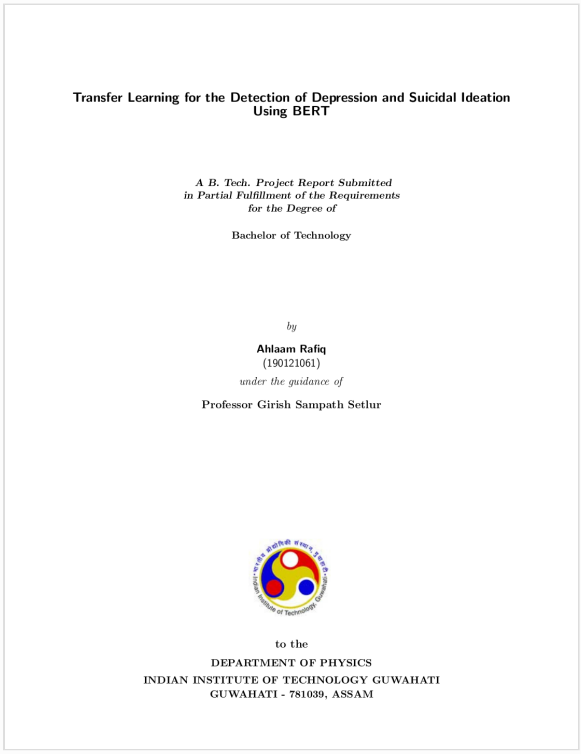
Part II: Turnitin report

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Ahlaam RAFIQ
Assignment title:	BTP 2022-ph421
Submission title:	BTP 4
File name:	thesis_style_4.pdf
File size:	625.27K
Page count:	47
Word count:	10,451
Character count:	57,325
Submission date:	07-Nov-2022 10:31PM (UTC+0530)
Submission ID:	1947255872



Transfer Learning for the Detection of Depression and Suicidal Ideation Using BERT

A B. Tech. Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Ahlaam Rafiq
(190121061)

under the guidance of

Professor Girish Sampath Setlur

to the
DEPARTMENT OF PHYSICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

BTP 4

by Ahlaam RAFIQ

Submission date: 07-Nov-2022 10:31PM (UTC+0530)
Submission ID: 1947255872
File name: thesis_style_4.pdf (625.27K)
Word count: 10451
Character count: 57325

Transfer Learning for the Detection of Depression and Suicidal Ideation Using BERT

A B. Tech. Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Ahlaam Rafiq
(190121061)

under the guidance of

Professor Girish Sampath Setlur

to the

DEPARTMENT OF PHYSICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Transfer Learning for the Detection of Depression and Suicidal Ideation Using BERT**" is a bonafide work of **Ahlaam Rafiq** (**Roll No. 190121061**), carried out in the Department of Physics, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Professor Girish Sampath Setlur**

Professor,

Department of Physics,

Indian Institute of Technology Guwahati,

Nov, 2022

Guwahati.

Assam.

Acknowledgements

I am grateful to Professor Girish S. Setlur for providing me with the opportunity to work in an area I am passionate about, and for his constant guidance throughout the course of the project. I am also grateful to his research team for providing me with very helpful feedback. I would also like to thank the Department of Physics, IIT Guwahati, for providing me with all the necessary support and resources.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Depression is a common mood disorder affecting millions of us everyday - a number that has only shot up during the pandemic. Not all depressed persons are suicidal; however, research indicates that most (possibly at least 90% of) people who have died by suicide have suffered from mental disorders.[53, 6, 11], Further, the risk of suicide is elevated for individuals with several mental disorders, such as depression, schizophrenia, bipolar disorder, and alcoholism.[34] It is, therefore, imperative to be able to identify both depression and suicidal ideation in posts by both individuals with and without mental disorders; in order to ensure at-risk individuals receive timely care. Manual inspection of posts - which can be overwhelmingly large in number, especially when made on social media forums - can be both painstaking and fatally slow in emergency situations.

This project researches the performance of the state-of-the-art natural language processing model, BERT, in classifying "depressive" ideation from "suicidal" ideation in Reddit posts. We find that our basic BERT base model (without any optimizations) outperforms the classical Naïve Bayes and SVM classifiers across all metrics, with a best F1-score of 0.93 (macro-averaged over both classes). Using a relatively small primary dataset (under 2,000

sequences combined), we also find that BERT performs reasonably well in a low resource setting. This study aims to ultimately optimize our models to improve performance scores, and transfer this learning to be able to detect depression and suicidal ideation in posts made by users with various mental illnesses.

1.1 An introduction to mental health and its importance

Defining mental health is not an easy task. An introduction can perhaps be borrowed from the World Health Organization, which defines mental health to be "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community." [49]

The past few years have seen a worldwide recognition of the importance of mental health - be it for overall physical health, community well-being, or economic stability - resulting in a much-needed (albeit relatively small) shift of resources to focus on mental healthcare. There is still, however, a long way to go. Mental health may not be quantifiable, but its numbers are.

1.1.1 Case study: Mental health figures, India

India's increased budget allocation to mental health (2.18%)[**?**] this year serves as a reminder of its poor track record. A few stark figures to consider:

- In 2019, one in every seven Indians was living with a mental health disorder[13]
- In 2017, the contribution of mental health disorders to the national disease burden nearly doubled from 1990.
- The WHO estimates that India will lose over one trillion USD due to the same between 2012-2030,

- The suicide mortality rate (per 100,000 population) in India is higher (16.3) than the global average (10.5) or even the average for Southeast Asia (13.4). [48]
- India is severely understaffed when it comes to mental health workers, with numbers of psychiatrists, nurses, psychologists, social works, therapists, etc. falling far below global and regional averages.[48]

,

It is evident that India's mental health infrastructure is poor, and generally lags behind the rest of South East Asia and the world. A 2019 study [26] estimates that India was then short of 27,000 doctors - requiring 2,700 new psychiatrists (keeping population growth and attrition rates at 0%) annually to fill the gap in the next 10 years - and notes that there are only 700 psychiatrists trained every year in post graduate seats. This gap is also not uniform across the country, and while there has been a an increase in psychiatric facilities in a few states, some states have seen a stagnation or even decline.

The ongoing COVID-19 pandemic has also worsened the current mental health crisis: the Indian Psychiatry Society reports a 20% increase in mental illness cases since the pandemic [45]. In addition, mental health disorders and seeking treatment for the same continue to be shrouded in stigma. A systematic review on the stigma associated with mental health problems among young people in India [24] also finds that one third displays poor knowledge of mental health problems and negative attitudes towards people with mental health problems and one in five has actual/intended stigmatizing behavior.

There is still, thus, a lot of ground to cover. And while there is no tailor-made solution, there have been attempts to streamline the process by involving technology, Numbers provide support: while Indians might lack mental health facilities, their access to technology and the internet is impressive. To give a rough idea about access, an estimated 844.84 million Indians were smartphone users and 107.81 million households had internet access at home in 2021 [56, 31]. As of 2020, India was the world's second-largest internet population at

over 749 million users in 2020, of which 744 million users accessed the internet via their mobile phones [55]. Technology, thus, seems to be primed for utilization to meet the mental health needs of people. A 2015 World Health Organization (WHO) survey of 15,000 mobile health apps revealed that 29% of them emphasized on mental health diagnosis, treatment, or support. [5]

But for a field as intricate as mental health, we need to refine the search for greater effectiveness. This is where artificial intelligence (AI) and its applications can play a major role.

1.2 Artificial intelligence in mental health

AI systems aim to think humanly and act humanly with the ultimate goal of obtaining rational outcomes. [32] It's a field that has seen massive growth in research and development in the past few decades. While not a perfect science by any means, its data-driven and knowledge-based methods - making an extensive use of various kinds of knowledge specific to the domain - have been responsible for significant progress in multiple fields, including expert systems, natural language processing, speech recognition, computer vision, and robotics.[15]

Machine learning is a branch of AI, focusing on using data and algorithms to imitate the way humans learn. Our project focuses on natural language processing (NLP), a field that - standing at a junction of linguistics, computer science, and AI - studies the interactions between computers and human language. NLP remains one of the biggest, and most common, practical applications of machine learning. From an economic point of view, AI either decreases the costs of prediction or improves the quality of predictions available at the same cost. [2]

Its application to mental health is multifold, with research supporting its use as an independent tool or clinical aid in therapy, training, screening, self-management, counseling, and

diagnosing. AI applications in healthcare have previously been found to improve life quality for citizens and efficiency of governance [36], and there's a lot of potential for research in applying it to mental health for reducing the burden on healthcare providers and ensuring equitable access.

1.2.1 Data types and sources

Machine learning is built on data collection and analysis. Choosing, and sourcing, data specific to the mental health domain in question thus opens up challenges in itself.

The "data" in themselves are pretty varied - machine learning enjoys the flexibility of working of data of different modalities. AI-based technologies in psychiatry rely on the identification of specific patterns within highly heterogenous multimodal sets, including: various psychometric scales or mood rating scales, brain imaging data, genomics, blood biomarkers, data based on novel monitoring systems (eg. smartphones), data scraped from social media platforms, speech and language data, facial data, dynamics of the oculometric system, attention assessment based on eye-gaze data, and various features based on the analysis of the peripheral physiological signals (eg. respiratory sinus arrhythmia, startle reactivity).

In this project, I focus on textual data scraped from Reddit and apply natural language processing methods to detect mental health problems (depression and suicidal ideation) in a variety of tasks. We choose text from a non-clinical source (here, online forums) due to its ready availability and extraction. Non-clinical textual data (social media, online forums, instant messaging, etc.) have been, in particular, a hotbed of research for their potential in mapping mental health.

1.3 Machine learning for text classification

Text classification is a classical NLP (natural language processing, a field of machine learning) problem that aims to assign labels to text objects. Background and related work has been

explored in detail in the next chapter.

This project is further strongly centred around transfer learning, which uses the knowledge learned while solving one task and applies to solve another related task.

- Transfer learning is critical to the implementation of BERT (Biderectional Encodeer Representations from Transformers) - a state-of-the-art language model for NLP used throughout the project, and
- We aim to transfer the classification learned in RQ1 and RQ3 to solve RQ4 (see section 1.4).

1.4 Research goals

Briefly, this project aims to solve the following broad research questions:

- RQ1: Use an optimized BERT-based model to classify text in a low-resource setting with depressive sentiment from text with suicidal ideation (two labels: "depressed" or "suicidal").
- RQ2: Compare the performance of BERT-based models with baseline classical machine learning methods, including Naive Bayes and SVM classifiers.
- RQ3: Extend the classification task to include a third, control "neither suicidal nor depressed" label, again using an optimized BERT-based model.
- RQ4: Use transfer learning to classify mental health disorder-specific textual data into the aforementioned three labels.

Chapter 2

Background and related work

This section contains a brief overview of the background of the project and related work. We cover (a) machine learning for text classification, (b), the application of text classification for detection of mental health disorders, (c) transfer learning and its success in low-resource settings, and (d) an additional literature review of the different data sources used for the application of machine learning for mental health solutions.

2.1 Machine learning for text classification

Text classification is a popular machine learning (or more specifically, natural language processing) task that entails categorizing text into organized groups that are defined using labels. "Text" can include sentences, paragraphs, documents, etc. Typical text classification problems include sentiment analysis, natural language inference, question answering, etc.

While text classification can be performed through manual annotation, the increasing scale of data calls for automatic labeling. This can be performed by either rule-based methods (which classify text based on a set of predefined, domain-specific rules) or by machine-learning based methods (which are data- and observation-driven, and try to capture the

inherent relationship between text and label).

Machine learning methods for text classification can largely be divided into two types:

1. Classical models, such as Naïve Bayes, random forests, support vector machines (SVMs) and gradient boosting trees. These models first extract features from the text - common features include bag of words (BoW) and term frequency-inverse document frequency (tfidf) - and then feed these features into a classifier. Classical models, although widely used, have the same limitation as rule based methods: substantial domain knowledge is required to choose appropriate features, thus making it hard to apply learned models for cross-domain tasks. We currently use classical Naïve Bayes and SVM classifiers as our baseline for our classification task.

2. Neural network approaches, which replace extracted features with an embedding model that uses machine learning to map text into a continuous, low-dimensional feature vector. There are several deep learning models with varying architectures built for text classification, including RNN-based models, CNN-based models, graph neural networks, and Transformers. Transformers form the basis for Devlin et al's BERT (Bidirectional Encoder Representations from Transformers) [18], which as of 2021 is the state-of-the-art embedding model and has outperformed classical methods in multiple NLP problems. Based on BERT's excellent performance on other text classification tasks, we hypothesize that BERT can successfully separate suicidal ideation from non-suicidal depression and detect the two in test data, and do this more accurately than classical methods. BERT has been explained in greater depth in the following chapter.

2.2 Text classification for mental health

We reviewed 60 papers on the applications of machine learning in mental healthcare. We found a variety of data types and sources used across the research, discussed in brief in section 2.4. Here we focus on textual data.

Sources of textual content may include transcriptions of clinical interviews or sessions and non-clinical text (such as social media, online forums, instant messaging). Language and voice can be incredibly useful as clinical text data. For example, natural language processing analysis is used [12] to discriminate speech in psychosis from normal speech. This involves preprocessing of transcripts, latent semantic analysis, part-of-speech tagging analysis, and then ML classification and validation. Similarly, speech features extracted from phone conversations can be used for the classification of bipolar disorder episodes. [4] Other examples of clinical sources are clinical assessments using questionnaires, self-assessments, and clinical records. Electronic Medical Records or EMRs (containing medical history from individual clinical practices) and Electronic Health Records (EHRs) (containing comprehensive long-term history collecting multiple EMRs) feature in several studies and are of special importance - In France, for example, emergency department EMRs have been used for automated surveillance of suicide attempts [42]; a UCLA study uses EHRs to differentiate risk of suicide attempt and self-harm after general, recurrent medical hospitalization of women with mental illness [19]; and the UK Clinical Record Interactive Search system has provided de-identified information sourced from EHRs for identifying suicidal ideation and attempts [22].

Non-clinical data finds a major source in social media, which has particularly been a hotbed of research for its potential in mapping mental health. A very brief search shows us that Facebook posts can predict depression in medical records [20], analyzing tweets can help estimate the effects of exercise on mental health [35], and a user's Instagram profile can detect major depressive disorder [52]. Most studies use purely textual data, others may also include features regarding the user's profile (such as accounts followed, number of posts, etc.) and their activity. For a low-resource study like this one, sourcing data from social media is attractive because it is potentially free, relatively less time-consuming, and can be abundantly sourced anywhere. Limitations include the introduction of bias due to participants' voluntary responses, the lack of clinical validation, and the impossibility of offering

personalized treatment.

Weighing the benefits and limitations, however, we find social media - specifically, Reddit forum posts - to be the most appropriate sources for our use case. Reddit is a popular online network or forum of communities - i.e., "subreddits", that are built around a single focal subject - where registered users can post multimodal content and interact with other users and their posts. As of 2022, it has 330 million monthly active users. [30] Several studies we reviewed have used subreddit data for text classification in the mental health domain. Reddit offers its users anonymity and security, hence allowing users with various mental illnesses to discuss their experiences without the fear of being stigmatized. [10] Unlike other popular, anonymous social media forums (such as Twitter), it does not cap post length, thus allowing us to build a dataset with a substantial amount of text.

Jiang et al. used a large-scale dataset of Reddit posts containing data from users with eight disorders, and a control user group [61]. They built strong classifiers using deep, contextualized word representations and concluded that these vastly outperformed previously applied statistical models with simple linguistic features. One such model uses BERT encoding and an attention-based classifier. The authors compared F1- and accuracy scores with a non-contextual baseline text analysis program classifier that maps words to psychologically motivated labels. Results across all labels/disorders vary between 0.67-0.8 for the non-contextual classifier and 0.8-0.9 for BERT, which is a significant improvement. BERT was also used by Nisa et al. for the early detection of self-harm by Reddit users [60], with improved recall over baseline logistic regression. El-Ramly et al. implemented CairoDep [21], built using BERT Transformers, to detect depression in Arabic posts. CairoDep achieves high accuracy, precision, recall, and F1-scores (above 96%) as compared to lexicon-based or shallow machine-learning models (80% or below).

A preceding work that is highly relevant to our study is Haque et al.'s *Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction* [9], which uses

(among others) BERT-based models to classify depression from suicidal ideation, as we do. Their model is further optimized using label correction and a control dataset (optimization remains a future area for us to explore).

Our review indicates that learning context greatly improves a classifier's performance on a mental health dataset. Further, BERT is among the state-of-the-art models available for the task. Lastly, Reddit is a competent source of data for our research goals.

Chapter 3

An overview of BERT

3.1 Self-supervised learning: a brief introduction

We begin by introducing self-supervised learning (SSL). SSL acts as an intermediate between supervised and unsupervised learning. It is like humans in the sense that it depends on previously acquired background knowledge of how the world works to make decisions. As researchers at Meta AI put it: "... self-supervised learning is one of the most promising ways to build such background knowledge and approximate a form of common sense in AI systems".[3]

SSL methods are built around two primary steps in neural network training: first, "pretraining" (neural network training is started with a pre-trained model), followed by task-specific "fine-tuning". Because pre-training needs massive amounts of annotated data (such as learning a language for natural language processing tasks), a pretrained model for the required task may not exist. SSL solves this problem by looking for "labels" that are naturally part of the input data rather than requiring separate external labels. The general technique is to predict an unobserved property of the input from an observed part – for example, masked language modelling in natural language processing (described in detail further) or predicting

hidden future/past frames from the current frame in a video. Without relying on labels, thus, SSL can use a variety of supervisory signals across large datasets. So, essentially, it is still "supervised" in a way – just not by humans.

The need for SSL in NLP arises from the fact that it is very difficult for an AI system to grasp the intricacies of text and language from available supervised training data. Consider a binary classification problem. To achieve high accuracy, we would like to train our model on a large train dataset – and this would need to be labelled by hand. Because of human capacity, there is a limit to how large this dataset could be.

However large this figure may appear, it is too small for a model to effectively learn the subtleties of language. This is where our model can benefit from self-supervised, transfer learning. What happens in SSL-based NLP models is this: the models are pre-trained on massive text corpuses with millions of datapoints (for example, BERT has been pre-trained on the English Wikipedia corpus with 2,500 million words and the BooksCorpus with 800 million words) – and these corpuses are unlabelled data! Given some specific tasks to look for, the model learns the underlying structure of text on its own. Pre-trained models are available off-the-shelf for most tasks. Researchers can take these models and "fine-tune" them on smaller, task-specific labelled datasets. This is supervised.

Today, SSL-based models are the state-of-the-art when it comes to NLP problems. One such incredibly popular model is the Bidirectional Encoder Representations from Transformers model, or BERT, explained in detail below.

3.2 BERT: Bidirectional Encoder Representations from Transformers

Language model pre-training was found to be effective for improving NLP tasks, both sentence- and token- level, and two strategies existed to apply pre-trained language repre-

sentations to downstream tasks [18] (i.e., supervised learning tasks that utilize a pre-trained model):

- *Feature-based*, with task-specific architecture and pre-trained representations as additional features, and
- *Fine-tuning-based*, where all pre-trained parameters and minimal task-specific parameters were fine-tuned.

Prior to BERT, NLP models were largely unidirectional (that is, tokens could only attend to previous tokens in the self-attention layers of the Transformer), and thus suboptimal. While other models (for example, bidirectional-CNNs) were introduced to take into account bidirectional context, BERT – introduced in 2018 by researchers at Google – was the first self-supervised deeply bidirectional system for pre-training NLP [18].

BERT pre-trains deep bidirectional representations from unlabelled text by jointly conditioning on both left, right context in all layers. It can be fine-tuned with just one extra output layer and can be used to model for tasks without major task-specific architecture modifications. Applications of BERT are far-ranging: it can perform on sentence-level (inference, paraphrasing, etc.) and token-level (named entity recognition, question-answering, etc.) tasks.

There are two stages in the BERT SSL framework:

1. **Pre-training**: The model is trained on unlabelled data over different pre-training tasks. This is expensive, but is only a one-time procedure.
2. **Fine-tuning**: The model is initialized with pre-trained parameters. All parameters are fine-tuned using labelled data from downstream tasks – and each downstream task has separate fine-tuned models despite being initialized with the same pre-trained parameters. This is the inexpensive stage.

BERT's mantra, therefore, is building a "unified architecture across different tasks".

3.2.1 BERT and transfer learning

There is an additional benefit to the general pre-training/ specific fine-tuning process. Due to a large number of parameters, training BERT from scratch leads to overfitting. Fine-tuning, on the other hand, takes the pre-trained model as a starting point and further trains on a relatively small dataset.

BERT is pre-trained using 2 unsupervised tasks [18]:

1. *Masked language modelling (MLM)*

   In the initial BERT implementation, the training generator chooses 15% token positions at random. The procedure is as follows: if the i^{th} token is chosen,

   - replace the i^{th} token with [MASK] 80% of the time,
   - replace the i^{th} token with a random token 10% of the time, and
   - leave unchanged 10% of the time.

   Then, a final hidden vector is chosen to predict the original token with cross-entropy loss. This involves predicting only the masked words, not reconstructing the entire input. Further, the underlying Transformer keeps a distributional contextual representation of every input token. It does not know which words it will be asked to predict or which have been replaced randomly. The more recent BERT implementation uses whole word masking instead of randomly selecting WordPiece tokens to mask, as the previous implementation did. Whole word masking works as follows: first, it masks all tokens corresponding to a word at once. The overall masking rate remains the same. Next, each masked WordPiece token is predicted independently.

2. *Next sentence prediction (NSP)*

   Pretraining for binarized NSP is useful to understand sentence relationships. When choosing sentences (i.e., two spans of text from the corpus) "A" and "B", for each

pre-training example:

- 50% of the time, B is the actual next sentence (label *isNext*)
- 50% of the time, B is a random sentence (label *isNotNext*)

The sample is such that the combined length less that or equal to 512 tokens.

Finally, the net training loss is the sum of the mean MLM likelihood and the mean NSP likelihood.

When it comes to fine-tuning, there are three ways to fine-tune [18]:

- Train the entire architecture on the dataset and feed the output to the softmax layer. Error is back-propagated through entire architecture, and pre-trained weights are updated based on the new dataset.
- Train some layers while freezing others – that is, train partially, retraining only higher layers.
- Freeze the entire architecture. That is, freeze all the layers, attach a few neural network layers and train this new model.

The underlying procedure for fine-tuning, however, remains the same. First, for each task, task-specific inputs and outputs are plugged into BERT. Next, all parameters are fine-tuned end-to-end. At the input, the sentences A and B from pre-training are equivalent to: sentence pairs (paraphrasing), question passage pairs (question answering), hypothesis premise (entailment), and text-label pair (text classification/ sequence tagging). At the output, token representations are fed into the output layer for token level tasks. A special representation token is fed into the output layer for classification.

For further optimization, the optimal hyperparameter values (hyperparameters include learning rate, number of training epochs, batch size, etc.) can be tuned and are task-specific.

Ablation studies on BERT reveal that [18]: Removing NSP as a pre-training task hurts

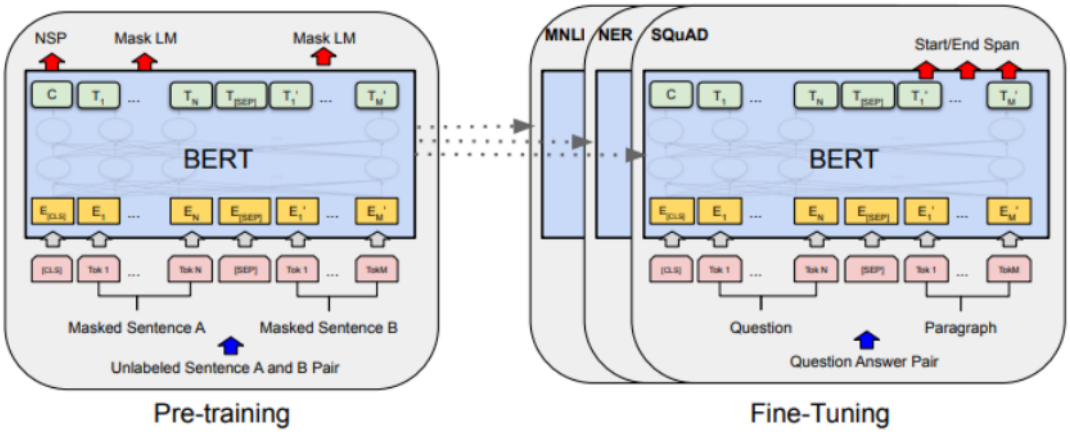


Fig. 3.1 Overall pre-training and fine-tuning procedures for BERT [18]. The same designs are utilised for pre-training and fine-tuning, with the exception of output layers. Models are initialised for various down-stream activities using the same pre-trained model parameters. All parameters are adjusted during fine-tuning. Every input example now has the special symbol [CLS] before it, and [SEP] is a special token for separation.

performance significantly on multiple benchmarks. BERT needs a large amount of pre-training to achieve high fine-tuning accuracy. Fine-tuning is robust to different masking strategies. The purpose of masking is to reduce mismatch between pre-training and fine-tuning. A larger model size leads to strict improvement in accuracy.

The BERT hypothesis [18], in conclusion, is as follows: *when the model is fine-tuned directly on downstream tasks, it uses only a very small number of randomly initialized additional parameters. Thus, task-specific models can benefit from larger, more expressive pre-trained representations even when the downstream task data is very small.*

There are multiple off-the-shelf BERT models available for use. We use the BERT-base (uncased) variant, which consists of 110M parameters and has been trained on the English Wikipedia and BookCorpus - that is, a total of 3.3 billion words.

Many improvements have been made to BERT over the years. Variants include:

- Small BERTS, which have the same general architecture as BERT but fewer or smaller

Transformer blocks,

- RoBERTa [38], a more robust implementation introduced as an "improved recipe" for training BERT models in response to BERT being found to be significantly undertrained,
- ALBERT [62], "A Lite BERT", that reduces model size without affecting computation time by sharing parameters between layers, and
- Electra [33], which has a setup resembling a generative adversarial network (GAN). Electra and BERT share the same architecture, but Electra is pre-trained as a discriminator.

To simplify our initial research, we continue to use the original BERT models over their optimized variants.

Chapter 4

Research experiments and results

Our initial research goal, completed in this study, is to train a BERT-based classifier to separate text showing suicidal ideation from those showing depression.

4.1 Dataset and data preprocessing

We choose Reddit as our data source for the following reasons:

1. Reddit is anonymous. Research indicates that this gives posters a sense of security and anonymity to discuss their experiences and struggles without the fear of being stigmatized or discriminated against.

2. Unlike Twitter and other platforms, Reddit has no limit on the size of a text post, thus providing a larger amount of *contextual* data available to our model.

3. As of 2021, Reddit has a 2.8 million subreddits [37] (smaller communities where people with shared interests interact). We found over a hundred subreddits devoted to specific mental health disorders or mental health in general. For example, in "r/depression" users talk about their struggles with depression; "r/BipolarReddit" is specific to bipolar disorder, "r/anxietyhelp" is for users with anxiety, "r/add" is for users with ADD, etc.

We find this especially useful for our future research work, where mental illness-specific data will be required.

4. Data posted publicly on Reddit is freely accessible and can be downloaded via the Reddit API.

4.1.1 Subreddit overview

For the current study, we downloaded data from the following subreddits:

1. *r/depression*: This community offers a peer-support space for users who struggle with, or know someone who struggles with, a depressive disorder. At the time of writing this paper, it has over 914,000 members.

2. *r/SuicideWatch*: This community offers a peer-support space for at-risk users struggling with suicidal thoughts. At the time of writing this paper, it has over 388,000 members

Using the Reddit Pushshift API, we analysed the top keywords of each subreddit. For r/depression, top keywords include: *self-harming*, *self-hatred*, *dejected*, *quietness*, *numbs*, *friendless*, and *later*. For r/SuicideWatch, top keywords include: *fantasise*, *offing*, *mourned*, *dead-end*, *unlivable*, *outlived*, *hurted*, and *clear-cut*.

While not immediately obvious, it is clear that there is *some* distinction between posts by users with depression and suicidal, at-risk users. Training a machine to automatically detect the difference can be of critical help in emergency settings, and can help provide emergency care to patients based on the severity of their textual responses.

4.1.2 Dataset

We scraped Reddit using the author's personal account as user agent. Most recent data for each subreddit was downloaded in 50 batches. Due to the scale the study and our goal to confirm the effectiveness of BERT in low-resource settings, we kept our dataset small.

- 1230 posts were scraped from *r/depression*, of which 979 were unique and retained.
- 1251 posts were scraped from r/SuicideWatch, of which 974 were unique and retained.

For our study, we only require the title of the post, the body of the post, and the subreddit name. All other information - such as number of comments, post awards, URL, date, post author - is discarded, effectively ensuring our data is de-identified. We concatenate the title of the post to its body and use this as our "text". The subreddit name is used to define the labels - if a post belongs to r/depression, it is automatically labeled as having *depressive* ideation and labeled 0; if a post belongs to r/SuicideWatch, it is automatically labeled as having *suicidal* ideation and labeled 1.

To reduce noise in the datasets, we perform basic data preprocessing by removing non UTF-8 characters and extra white spaces. Further typical NLP preprocessing (such as tokenization or stemming) is not necessary and taken care of by BERT's superior encoding mechanism. However, for comparison with classical machine learning models, we perform additional preprocessing steps: (1) all text is lowered, (2) word tokenization is performed, (3) stop words are removed, (4) word lemmatization is performed. We use the NLTK library for performing tokenization and lemmatization, and the NLTK list of English stopwords as our stopwords.

4.2 Training and evaluating the classifiers

We implement two classical machine learning models (section) and one deep learning model (section). After performing the required data preprocessing steps for each model, we split the data into train, test, and validation sets following a 5-fold cross validation method. The split is in a 70:15:15 ratio. Trained classifiers are tested on the test set and we assess the performance of each using 4 *performance metrics*: accuracy, recall (for both classes), precision (for both classes), and F1 (for both classes).

4.2.1 Classical machine learning models

In this study, we consider only two classical models due to limited resources. For implementation, off-the-shelf classifiers available in the *sklearn* library are used.

1. *Support Vector Machine (SVM)*, a linear model that classifies the dataset as "depressive" (0) or "suicidal" (1) by drawing a line (in higher dimensions, this would be a hyperplane) separating the data into either class. We use a linear kernel. [27]

2. *Naïve Bayes (NB)*, a simple probabilistic classifier that applies Bayes' theorem - i.e., it assumes that each feature contributes independently to class assignment or that that presence/ absence of a particular feature is independent of the presence/ absence of another feature. We use sklearn's NB classifier for multinomial models, that is suitable for classification with discrete features (in our case, word counts). [51]

We use fractional word counts (tf-idf/ "term frequency-inverse document frequency") as discrete features for classification.

4.2.2 BERT-based model

We use the uncased BERT base model, available off-the-shelf as *bert-base-uncased* [28] from Hugging Face, as our primary model. While a cased model is also available (with a higher number of parameters), we chose the uncased model because information about casing is not relevant to our study. It is, further, possible to pretrain the model from scratch to be domain-specific, but we choose not to as: (1) the available BERT models show excellent results across domains, and (2) we wish to reduce the risk of overfitting [59]. Classification is performed using a PyTorch backend with the Hugging Face *Transformers* library and the *AutoModelForSequenceClassification* class. *transformers.AutoModelForSequenceClassification* is a generic model class instantiated from the pretrained BERT model (instantiation also loads pretrained model weights) that carries a sequence classification head.

Our input dataset has the following class distribution: 979 sequences in class 0, and 974

sequences in class 1. Since the dataset is almost evenly balanced, we did not see a need for additional class balancing or evaluating more appropriate metrics.

We follow a 5-fold cross-validation technique to improve our metric accuracy - that is, we train five separate times, with the model seeing a different train/ evaluation/ test dataset each time. For each fold, the dataset is split into train, evaluation, and test datasets in a 70:15:15 ratio (stratified along labels, randomly resampled for each fold).

For training, evaluating, and testing, we use the *Transformers* library's *Trainer* class [29]. We choose the following hyperparameters (default - future work includes performing hyperparameter optimization):

- Evaluation strategy: "epoch". That is, the performance of the model is evaluated at the end of each epoch using the specified *evaluation metric*. Note that the *evaluation metric* is not the *performance metric* - the former is only used to update the model weights at the end of each epoch, the latter is used to check the performance of the final model on the test dataset.
- Learning rate: $2e^{-5}$. This is a common initial learning rate, also recommended by Devlin et al. [cite]
- Training and evaluation batch sizes: 4.
- Number of training epochs: 10
- Evaluation metrics: accuracy, precision (macro-average), recall (macro-average), F1-score (macro-average). We use macro-averages for overall evaluation.

4.2.3 Performance metrics

We evaluate performances across models using the following four metrics (here TP, TN, FP, FN represent numbers of true positives, true negatives, false positives and false negatives respectively):

1. **Precision**, which evaluates the ratio of test posts that have been correctly assigned to their target class. That is, $precision = \frac{TP}{TP+FP}$. Precision can take values between 0 and 1. 0 indicates a poor model and 1 indicates a perfect model. We calculate precision separately for each class and also a macro-average over both classes to represent overall performance.

2. **Recall**, which evaluates the ratio of test posts that actually belong to the predicted class (i.e. target class is the predicted class) to all posts predicted as belonging to that class. That is, $recall = \frac{TP}{TP+FN}$. Recall can take values between 0 and 1. 0 indicates a poor model and 1 indicates a perfect model. We calculate recall separately for each class and also a macro-average over both classes to represent overall performance.

3. **F1-score**, which calculates the harmonic mean of precision and recall, useful for combining precision and recall into a single measure (when used in isolation, precision and recall and not particularly informative). $F_1 = 2 \times \frac{precison \times recall}{precision+recall}$. F1 can take values between 0 and 1. 0 indicates an imperfect model and 1 indicates a perfect model with perfect precision and perfect recall. We calculate F1-scores separately for each class and also a macro-average over both classes to represent overall performance.

4. **Accuracy**, which calculates the ratio between the number of correct predictions to the total number of predictions. That is, $accuracy = \frac{TP+TN}{TP+TN+FP+FS}$. Accuracy can take values between 0 and 1. 0 indicates an imperfect model and 1 indicates a good model. It is a useful metric when both classes are balanced, such as in our case.

4.3 Results

The subsequent figures 4.1-4.4 plot the scores for each performance metric across all models. *Green* bars are used to represent BERT base, and *red* bars are used to represent classical models. In the figures,

- BERT/A represents BERT base evaluated by accuracy,
- BERT/F1 represents BERT base evaluated by F1-score,
- BERT/R represents BERT base evaluated by recall,
- BERT/P represents BERT base evaluated by precision,
- NB represents the Naive Bayes classifier, and
- SVM represents the Support Vector Machine classifier.

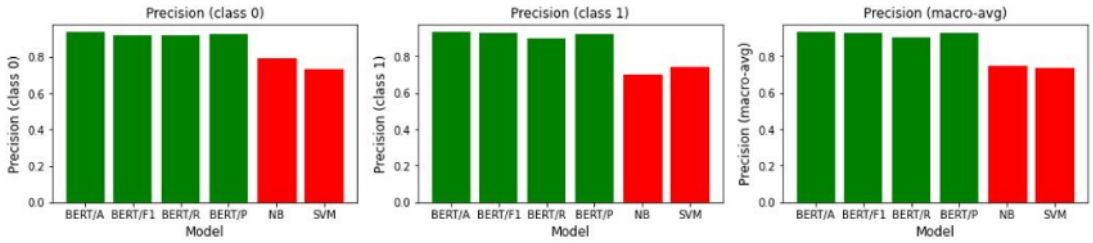


Fig. 4.1 Performance metric scores when model is evaluated on precision

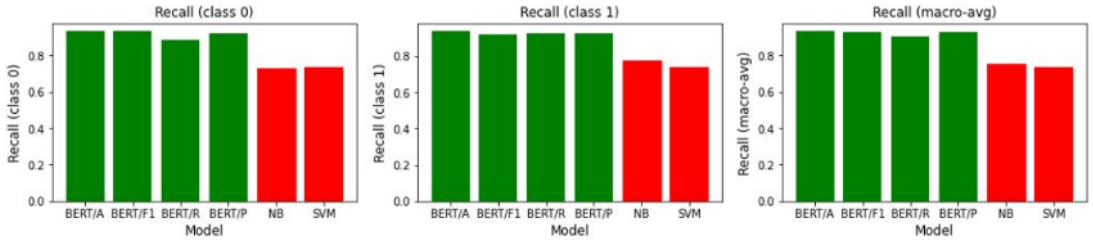


Fig. 4.2 Performance metric scores when model is evaluated on recall

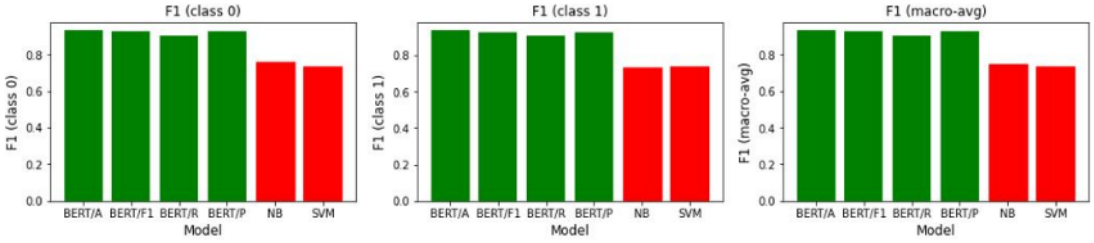


Fig. 4.3 Performance metric scores when model is evaluated on F1-score

The exact scores are given in table 4.1. "MA" represents the macro-average score over both classes.

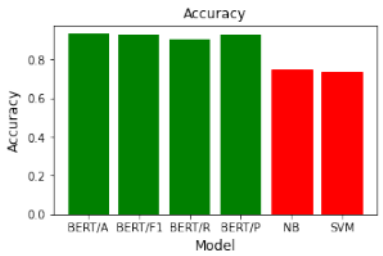


Fig. 4.4 Performance metric scores when model is evaluated on accuracy

4.4 Discussion

On observation, we conclude that BERT is the best classifier for separating posts showing depression from posts showing suicidal ideation. All our four experiments, which involve training the model by evaluating along different metrics, return performance scores for each metric (both macro-averages and scores for each class) above 0.90. The best scores across the board are obtained by evaluating BERT base along accuracy - all metrics (both macro-averages and scores for each class) attain their highest values (all ¿0.93). While classical models perform reasonably well - since no special measures have been taken to optimize them - they perform significantly worse that BERT in terms of all scores.

Model	Evaluation metric	Precision			Recall			F1-score			Accuracy
		0	1	MA	0	1	MA	0	1	MA	
BERT base	Accuracy	0.934	0.932	0.932	0.932	0.934	0.932	0.93	0.934	0.932	0.932
BERT base	F1	0.920	0.930	0.926	0.932	0.918	0.926	0.926	0.922	0.926	0.926
BERT base	Recall	0.916	0.896	0.906	0.888	0.922	0.904	0.900	0.908	0.904	0.904
BERT base	Precision	0.924	0.922	0.926	0.924	0.922	0.926	0.926	0.924	0.926	0.926
Naive Bayes	-	0.794	0.702	0.748	0.728	0.774	0.752	0.758	0.734	0.748	0.748
SVM	-	0.734	0.740	0.738	0.738	0.736	0.738	0.736	0.738	0.738	0.738

Table 4.1 All scores. *0* represents class 0, *depressive*. *1* represents class 1, *suicidal*. *MA* represents the macro-average score over both classes

Chapter 5

Conclusion and Future Work

In this initial stage of the project, we have successfully answered the first two of our research goals. We have shown that, even without optimization and in a low-resource setting, a BERT base model performs exceedingly well in classifying "depressive" posts from "suicidal posts". Further, we have shown that this model outperforms classical Naöve Bayes and SVM classifiers across four metrics: precision, recall, F1-score and accuracy for both "depressive"/0 and "suicidal"/1 classes - and their macro-averages, wherever applicable.

We aim to complete the following experiments or answer the following questions in our future work:

1. Add a control dataset (for instance, data from r/CasualConversation, a subreddit for users to have a friendly, casual conversation on any topic). How do results with a control dataset compare to results now? Further, how does performance compare with Haque et al's results?

2. Increase dataset size. There are two possibilities that can be explored and compared: (a) simply scraping more data, and (b) data augmentation methods, such as random swapping, random deletion, prepending, etc. How does performance differ in a higher-

resource setting?

3. In real life, it is likely that input data will be highly skewed. How capable is a model trained and evaluated on balanced datasets of performing well on an imbalanced test dataset? Does performance improve by introducing a corresponding skewness in the test and evaluation datasets? Class balancing methods such are random resampling may also be considered.

4. Perform hyperparameter optimization to improve overall scores.

5. Our final and primary research goal is to apply this learning to other mental-illness-specific data. For instance, our model should be able to identify "depressive" and "suicidal" ideation in posts by individuals with eating disorders. We believe this can find incredible use in emergency settings.

It's important to highlight, once more, how mental health is absolutely integral to society. It has taken a global crisis to come to terms with the fact that ignoring mental health, and surrounding it with stigma, is no longer an option. It doesn't just concern those with mental health disorders – it concerns all of us.

So far, overlooking mental health has had a huge toll on nearly every aspect of human life and wellbeing, from affecting families to severely costing the national economy. For society as a whole, mental, physical, and social health remain closely interwoven and equally vital. It's imperative, therefore, that mental health be accorded the same importance as, for example, physical health might. The accessibility and adaptability of AI – for all its challenges – does seem to be a promising solution.

Chapter 6

Appendices

6.1 Appendix 1: A brief note on Transformers

Transformers [7] are a network architecture based solely on attention mechanisms, hence doing away with recurrence and convolutions entirely. Because they are more parallelizable, require significantly less time to train, and draw global dependencies between input and output, Transformers are finding increasing use across machine learning problems.

Like LSTMs, Transformers are an architecture to transform one sequence to another, using two components: an encoder and a decoder. Unlike previous Seq2Seq models, however, it does not use any RNNs (GRU, LSTM, etc.). Due to this, there is no need to specify relative positions: positional encodings of different words are added to the embedded representation (this is an n-dimensional vector) of each word.

Model architecture comprises of an encoder and a decoder, both of which are stacks of modules mainly of multi-headed attention and feed-forward layers. The encoder takes the input sequence and maps it to a higher dimensional space. This n-dimensional vector is then fed into the decoder, which outputs a sequence.

While attention mechanisms were introduced prior to Transformers, multi-headed attention

(or self-attention) expands the model's ability to focus on different positions in the text. The attention mechanism looks at the input sequence and decides at each step which other parts of the sequence are important. For each input read in by the encoder, it takes into account several other inputs simultaneously and decides the importance of each by assigning weights.

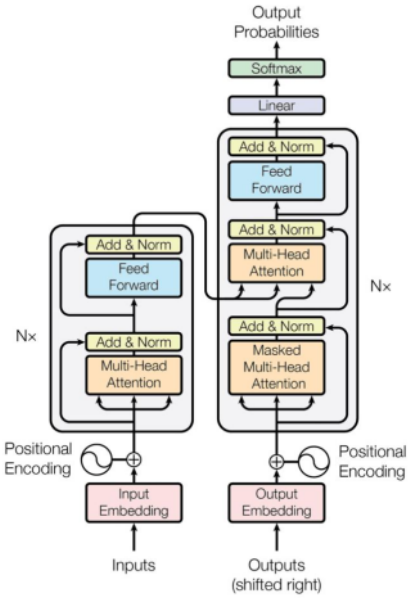


Fig. 6.1 Transformer model architecture

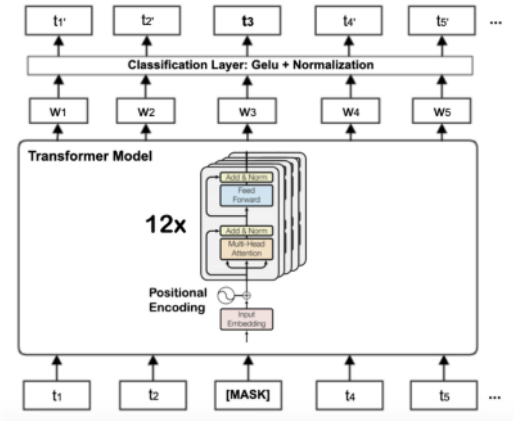


Fig. 6.2 BERT base model architecture, with 12 Transformer layers

Advantages of Transformers over LSTMs include [14]:

- The absence of locality bias due to multi-headed attention.
- A single multiplication per layer is performed - this is more efficient, as effective batch size is now the number of words, not the number of sequences.

The model used in our project, BERT-base, consists of 12 Transformer layers. Its architecture is show in the figure below.

6.2 Appendix 2: A brief note on other mental health data for AI research

The following table attemps to show India's stark mental health figures, as released by the WHO's 2017 Mental Health Atlas [48] (released every three years; India did not participate in the more recent 2020 Atlas). For comparison, we have included figures from the South East Asia region (SEAR), the European Region (EUR), Germany (a world leader in terms of mental health care), and the global average (GLB). Fields marked with "-" are indicative of unavailable data.

Figures	India	GLB*	SEAR*	EUR*	Germany*
Total mental health expenditure per person	4 INR/ ∼0.06 USD	2.5 USD	0.1 USD	21.7 USD	350.58 EUR/ ∼388 USD
Suicide mortality rate**	16.3	10.5	13.4	12.9	13.6
Disability-adjusted life years** (DALYs)	2,433.41	-	-	-	3603.56
Total number of mental health professionals (govt. and non-govt.)	25,312	-	-	-	118,367
Total mental health workers**	1.93	-	-	-	144.87
Number of psychiatrists**	0.29	1.3	0.4	9.9	13.20
Number of child psychiatrists**	0.00	<0.1	-	-	2.76
Total number of child psychiatrists	49	-	-	-	2.76
Number of other specialist doctors**	0.15	-	-	-	3.5
Number of mental health nurses**	0.80	3.5	0.80	23.2	-
Number of psychologists**	0.07	0.9	0.1	4.6	49.55
Number of social workers**	0.06	0.9	0.2	0.8	-
Number of occupational therapists**	0.03	-	-	-	56.43
Number of speech therapists**	0.17	-	-	-	19.41
Number of other paid mental health workers**	0.36	0.5	0.4	11.2	-

** Relative figures, wherever applicable*
** Rate per 100,000 population*

Table 6.1 Figures from the 2017 WHO Mental Health Atlas

Current research and data sources

In our literature review, 35 studies used interviews (and/or interview transcripts) and inventories/questionnaires/assessments (both self-report and clinician-administered) based on psychometric or mood rating scales. These are primarily used for 4 purposes: (1) for screening during participant recruitment, (2) for classification of participants into study subgroups, (3) for initial diagnosis, classification or scoring and follow-ups over the course of the study, and (4) for use as features for the machine learning model.

For implementation in India, the WHO's recommended process of translation and adaptation of instruments [47] may serve as a guide. Previous work on cross-cultural adaptation includes

the translation and validation of several SRIs into multiple Indian languages [58, 41, 50, 57, 17, 1, 25, 16].

EMRs and EHRs, mentioned previously in our discussion of textual sources, are extremely useful in addition to being rich in data for research and development. They ensure anytime/anywhere accessibility of patient records, improve the quality of records and are cost-effective, help track patients' clinical records and improve patient compliance, can be transferred easily within and across healthcare facilities, are easy to update, and facilitate improved healthcare decisions and provide evidence-based care [54]. For research specifically, they serve multiple uses. They can help in the recruitment, identification and screening of the study population and its medical history, and most importantly, they provide a large clinical database with multiple features that can be used for training, validation, and correlation.

India too has recognized the importance of digitizing - and thus improving the reliability of - its healthcare. The Ministry of Health and Family Welfare (MoHFW) first came out with standards for EHRs for India in September 2013. It also proposed to set up the National eHealth Authority (NeHA) in 2015, aiming to promote the setting up of state health records repositories and health information exchanges to facilitate interoperability. NeHA also looked to formulate and manage all health informatics standards for India. The MoHFW also put forward the Digital Health Information in Healthcare Security (DISHA) Act in 2018. DISHA has been drafted to "provide for the establishment of National and State eHealth Authorities and Health Information Exchanges; to standardize and regulate the processes related to collection, storing, transmission and use of digital health data; and to ensure reliability, data privacy, confidentiality and security of digital health data and such other matters related and incidental thereto" [46]. The same year, NITI Aayog proposed to create digital health records for all citizens by 2022 with the National Health Stack (NHS) [23]. In September 2021, the National Digital Health Mission (NDHM) - which will provide

every citizen with a health ID - was announced. Stored on the NHS, this ID will contain the individual's complete medical history and will be accessible by the entire healthcare industry. The NHS is basically an ecosystem of cloud-based services and has in the past couple years empanelled a number of private entities, including AI companies. A 2020 Centre for Sustainable Development paper on EHRs in India goes into further detail about the topic [39].

6.2.1 Challenges to adopting AI for mental health

A challenge to IoT-enabled AI applications (sensors, monitoring devices) is the interoperability of systems [8]. Since the IoT industry currently lacks technical standards, the variation in hardware and software leads to an inconsistent technology ecosystem. The systems may not, again, be interoperable with government infrastructure and this diversity could potentially cause problems with system maintenance and scalability.

A second problem is data privacy and security. AI applications, like traditional ones, are vulnerable to cybersecurity threats. Data consent is also an issue. Users may not be aware of (1) what data is collected, (2) how it is stored and processed and by whom, and (3) who is benefitting from this data. This is visible with the National Health Stack too. Researchers remain wary of the possible security concerns a data breach might bring up: as of October 2021, there were no standards set on data anonymisation or further use of this data by private entities [44]. Theft of medical identity is also a growing concern.

Clarifying data ownership also comes into question. The fact that AI applications are typically "black box" learning systems means that there may be a lot of ambiguity about their societal outcomes, and users' misgivings in their use is justified.

A third, vital challenge is ethics. On top of the existing inherent bias in machine algorithms, it's difficult to decide the limits of AI function. For instance, in a critical area such as mental health, can we code - and trust - chatbots to be reliable mandatory reporters? Can AI be

held accountable for its decisions? Further, because studies involve human subjects, data collection must also be meet ethical requirements.

A fourth challenge is environmental sustainability. Collecting, storing, and analysing the massive amounts of data required by AI applications will lead to significant consumption of energy and power,

And finally, there's the existing digital divide and the fact that struggling for "equitable access" will only worsen the conditions of a few. This was visible with the Aadhar process, and given the NHS's current infrastructure, it's a valid fear.

Further work on applying AI ethically has been detailed by NITI Aayog's 2021 Approach Document on "Responsible AI" [43]. A research agenda on AI for smart government [40] stresses on ensuring four principles to address challenges: transparency, accountability, fairness, and ethics.

To maximize the benefits of AI (for application in a general field), West and Allen [14] recommend the following steps: encouraging greater data access for researchers without compromising users' personal privacy; investing more government funding in unclassified AI research; promoting new models of digital education and AI workforce development so employees have the skills needed in the 21st-century economy; creating a federal AI advisory committee to make policy recommendations; engaging with state and local officials so they enact effective policies; regulating broad AI principles rather than specific algorithms; taking bias complaints seriously so AI does not replicate historic injustice, unfairness, or discrimination in data or algorithms; maintaining mechanisms for human oversight and control; penalizing malicious AI behaviour, and promoting cybersecurity.

References

[1] L. A. and C. A. Psychometric validation of geriatric depression scale - short form among bengali-speaking elderly from a rural area of west bengal: Application of item response theory.

[2] A. Agrawal. The economics of artificial intelligence. *McKinsey Quarterly*, 2018.

[3] M. AI. Self-supervised learning: The dark matter of intelligence. 2021.

[4] V. O. H. P. O. M. E. F. M. Alban Maxhuni, Angélica Muñoz-Meléndez. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*, 31:50–66, 2016.

[5] E. Anthes. Mental health: There's an app for that. *Nature*, 532:20–23, 2016.

[6] T. G. Arsenault-Lapierre G., Kim C. Psychiatric diagnoses in 3275 suicides: A meta-analysis. *BMC Psychiatry*, 4:37, 2004.

[7] N. P. e. a. Ashish Vaswani, Noam Shazeer. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[8] S. M. Atreyi Kankanhalli, Yannis Charalabidis. Iot and ai for smart government: A research agenda. *Government Information Quarterly*, 35:304–309, 2019.

[9] T. G. Ayaan Haque, Viraaj Reddi. Deep learning for suicide and depression identification with unsupervised label correction. 2021.

[10] Balani and . H.-W. e. a. . De Choudhury, 2015; Berry et al.

[11] F. A. Bertolote J.M. Suicide and psychiatric diagnosis: A worldwide perspective. *World Psychiatry*, 1:181–185, 2002.

[12] D. F.-S.-G. B. C. K. D. C. J. C. E. B. G. A. C. Cheryl M. Corcoran, Facundo Carrillo. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17:67–75, 2018.

[13] I. S.-L. D. B. I. M. D. Collaborators. The burden of mental disorders across the states of india: the global burden of disease study 1990–2017. *The Lancet Psychiatry*, 7:148–161, 2020.

[14] S. Cristina. The transformer model. *Machine Learning Mastery*, 2022.

[15] V. R. V. . V. D. Damioli, G. The impact of artificial intelligence on labor productivity. *Eurasian Bus Rev*, 11:1–25, 2021.

[16] S. T. e. a. De Man J., Absetz P. Are the phq-9 and gad-7 suitable for use in india? a psychometric analysis. *Frontiers in Psychology*, 2021.

[17] S. E. e. a. Desai N. D., Shah S. N. Validation of gujarati version of 15-item geriatric depression scale in elderly medical outpatients of general hospital in gujarat. *International Journal of Medical Science and Public Health*, 3:1453–1458, 2014.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] J. P. e. a. Edgcomb J. B., Thiruvalluru R. Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalizaton of women with mental illness. *Medical Care*, 59:S58–S64, 2021.

[20] M. R. M. e. a. Eichstaedt J. C., Smith R. J. Facebook language predicts depression in medical records. *PNAS*, 115:11203–11208, 2018.

[21] M. E.-R. et al. Cairodep: Detecting depression in arabic posts using bert transformers. *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 207–212, 2021.

[22] D. R. V. S. e. a. Fernandes, A.C. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci Rep*, 8: 7426, 2018.

[23] N. I. for Transforming India. National health stack. 2018.

[24] T. S. T. K. M. e. a. Gaiha, S.M. Stigma associated with mental health problems among young people in india: a systematic review of magnitude, manifestations and recommendations. *BMC Psychiatry*, 20:538, 2020.

[25] R. P. e. a. Ganguly S., Samanta M. Patient health questionnaire-9 as an effective tool for screening of depression among indian adolescents. *Journal of Adolescent Health*, 52: 546–551, 2013.

[26] C. P. S. Garg K., Kumar C. N. Number of psychiatrists in india: Baby steps forward, but a long way to go.

[27] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15:290–298, 2018.

[28] Huggingface. https://huggingface.co/bert-base-uncased. .

[29] Huggingface. https://huggingface.co/docs/transformers/main$_c lasses/trainer$..

[30] A. Hutchinson. Reddit now has as many users as twitter, and far higher engagement rates. *SocialMediaToday*, 2018.

[31] D. J. Internet households in india 2010-2025. *Statista*, 2021.

[32] H. J. A brief introduction to artificial intelligence.

[33] Q. V. L. e. a. Kevin Clark, Minh-Thang Luong. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv*, 2020.

[34] B. L. Suicide risk and mental disorders. *Int J Environ Res Public Health*, 2018.

[35] D. R. V. L. and C. A. Using matched samples to estimate the effects of exercise on mental health via twitter.

[36] D. Lee and S. N. Yoon. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *Int J Environ Res Public Health*, 18:271, 2021.

[37] Y. Lin. 10 reddit statistics every marketer should know in 2023 [infographic]. *Oberlo*, 2022.

[38] Y. Liu, M. Ott, and e. a. Goyal. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

[39] W. M. Ict india working paper 25: Electronic health records in india. 2018.

[40] W. D. M. and A. J. R. How artificial intelligence is transforming the world.

[41] B. M. e. a. Mehra A., Agarwal A. Evaluation of psychometric properties of hindi versions of the geriatric depression scale and patient health questionnaire in older adults. *Indian Journal of Psychological Medicine*, 43, 2021.

[42] G. Q. e. a. Metzger M., Tvardik N. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: A french pilot study. *Int J Methods Psychiatry Res.*, 26:1522, 2017.

[43] M. MK. Responsible ai aiforall. approach document for india part 1 - principles for responsible ai.

[44] M. MK. The risks of storing health records of 1.3 billion indians on the national health stack. *The News Minute*, 2021.

[45] A. K. . K. R. Nayar. Covid 19 and its mental health consequences. *Journal of Mental Health*, 30:1–2, 2021.

[46] M. of Health and I. Family Welfare. Digital information security in healthcare, act. 2018.

[47] W. H. Organization. Guidelines on translation and adaptation of instruments.

[48] W. H. Organization. Mental health atlas 2017. 2017.

[49] W. H. Organization. Mental health: strengthening our response. *WHO Newsroom*, 2022.

[50] S. T. V. Rajgopal J. and M. M. Psychometric properties of the geriatric depression scale (kannada version): A community-based study. *Journal of Geriatric Mental Health*, 6, 2019.

[51] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

[52] C. B. e. a. Ricard B. J., Marsch L. A. Exploring the utility of community-generated social media content for detecting depression: An analytical study on instagram. *Journal of Medical Internet Research*, 20, 2018.

[53] W. R. J. G. S. K. J. Robins E., Murphy G.E. Some clinical considerations in the prevention of suicide based on a study of 134 successful suicides. *Am. J. Public Health Nations Health*, 49:888–899, 1959.

[54] L. M. e. a. Robinson J., Witt K. Development of a self-harm monitoring system for victoria. *Int. J. Environ. Res. Public Health*, 20, 2020.

[55] K. S. Mobile internet users in india 2010-2040. *Statista*, 2021.

[56] S. S. Smartphone users in india 2010-2040. *Statista*, 2021.

[57] R. G. e. a. Sarkar S., Kattimani S. Validation of the tamil version of the short form geriatric depression scale-15. *Journal of Neurosciences in Rural Practice*, 6:442–1446, 2015.

[58] C. V. Thomas A. M. and A. A. Translation, validation and cross-cultural adaptation of the geriatric depression scale (gds-30) for utilization amongst speakers of malayalam; the regional language of the south indian state of kerala. *Journal of Family Medicine and Primary Care*, 10:1863–1867, 2021.

[59] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019.

[60] Q. un Nisa and R. Muhammad. Towards transfer learning using bert for early detection of self-harm of social media users. *Conference and Labs of the Evaluation Forum*, 2936.

[61] S. I. L. e. a. Zhengping Jiang. Detection of mental health conditions from reddit via deep contextualized representations. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, page 147, 2020.

[62] S. G. e. a. Zhenzhong Lan, Mingda Chen. Albert: A lite bert for self-supervised learning of language representations. *arXiv*, 2019.

BTP 4

ORIGINALITY REPORT

21%	18%	13%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	deeplearn.org Internet Source	2%
2	www.econstor.eu Internet Source	1%
3	www.abhimanukumar.com Internet Source	1%
4	aclanthology.org Internet Source	1%
5	www.brookings.edu Internet Source	1%
6	www.ncbi.nlm.nih.gov Internet Source	1%
7	medium.com Internet Source	1%
8	doctorpenguin.com Internet Source	1%
9	www.teriin.org Internet Source	1%

10	coek.info Internet Source	1 %
11	drelaineryan.com Internet Source	1 %
12	shreyansh26.github.io Internet Source	<1 %
13	www.mdpi.com Internet Source	<1 %
14	sh-tsang.medium.com Internet Source	<1 %
15	ijip.in Internet Source	<1 %
16	Mohna Chakraborty. "Does reusing pre-trained NLP model propagate bugs?", Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021 Publication	<1 %
17	www.analyticsvidhya.com Internet Source	<1 %
18	www.nhp.gov.in Internet Source	<1 %
19	www.statista.com Internet Source	<1 %

20	www.fast.ai Internet Source	<1%
21	"Artificial Intelligence in Medicine", Springer Science and Business Media LLC, 2021 Publication	<1%
22	Nisha Mani Pandey, Rakesh Kumar Tripathi, Sujita Kumar Kar, K L Vidya, Nitika Singh. "Mental health promotion for elderly populations in World Health Organization South-East Asia Region: Needs and resource gaps", World Journal of Psychiatry, 2022 Publication	<1%
23	orca.cardiff.ac.uk Internet Source	<1%
24	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2020 Publication	<1%
25	Yawen Wang, Lin Shi, Mingyang Li, Qing Wang, Yun Yang. "A Deep Context-wise Method for Coreference Detection in Natural Language Requirements", 2020 IEEE 28th International Requirements Engineering Conference (RE), 2020 Publication	<1%
26	ai.facebook.com Internet Source	<1%

27	curve.carleton.ca Internet Source	<1 %
28	www.coursehero.com Internet Source	<1 %
29	www.igi-global.com Internet Source	<1 %
30	www.kdnuggets.com Internet Source	<1 %
31	Behnaz Balmaki, Masoud A. Rostami, Tara Christensen, Elizabeth A. Leger et al. "Modern approaches for leveraging biodiversity collections to understand change in plant-insect interactions", Frontiers in Ecology and Evolution, 2022 Publication	<1 %
32	www.aclweb.org Internet Source	<1 %
33	www.frontiersin.org Internet Source	<1 %
34	Jiangjiang Zhao, Jie Zhu, Xiaokun Zhang, Xian-Ling Mao, Heyan Huang. "Incorporating Domain Knowledge into Text Classification Diagnosis in Customer Service Dialogue Field", Journal of Physics: Conference Series, 2021 Publication	<1 %

35	Norman, Ian, Ryrie, Iain. "The Art and Science of Mental Health Nursing: Principles and Practice", The Art and Science of Mental Health Nursing: Principles and Practice, 2018 Publication	<1 %
36	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
37	zzd2012victor.medium.com Internet Source	<1 %
38	Rajesh Sagar, Rakhi Dandona, Gopalkrishna Gururaj, R S Dhaliwal et al. "The burden of mental disorders across the states of India: the Global Burden of Disease Study 1990–2017", The Lancet Psychiatry, 2020 Publication	<1 %
39	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2021 Publication	<1 %
40	"Intelligent Systems", Springer Science and Business Media LLC, 2020 Publication	<1 %
41	Lili Bo, Jinting Lu. "Bug Question Answering with Pretrained Encoders", 2021 IEEE International Conference on Software	<1 %

Analysis, Evolution and Reengineering (SANER), 2021
Publication

42	Martin Trapp, Marcin Skowron, Dietmar Schabus. "Retrieving Compositional Documents Using Position-Sensitive Word Mover's Distance", Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '17, 2017 Publication	<1 %
43	ecommons.cornell.edu Internet Source	<1 %
44	opus.uleth.ca Internet Source	<1 %
45	biorobotics.harvard.edu Internet Source	<1 %
46	blog.uny.ac.id Internet Source	<1 %
47	web.archive.org Internet Source	<1 %
48	www.sasop.co.za Internet Source	<1 %
49	Tiago Martinho de Barros, Helio Pedrini, Zanoni Dias. "Leveraging emoji to improve sentiment classification of tweets",	<1 %

	Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021 Publication	
50	Wen-Yuan Zhu, Yu-Wen Wang, Chin-Jie Chen, Wen-Chih Peng, Po-Ruey Lei. "A Bayesian-Based Approach for Activity and Mobility Inference in Location-Based Social Networks", 2016 17th IEEE International Conference on Mobile Data Management (MDM), 2016 Publication	<1 %
51	estudogeral.sib.uc.pt Internet Source	<1 %
52	jogh.org Internet Source	<1 %
53	medinform.jmir.org Internet Source	<1 %
54	www.kantorlaw.net Internet Source	<1 %
55	www.nature.com Internet Source	<1 %
56	"Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery", Springer Science and Business Media LLC, 2022 Publication	<1 %
57	Michael F. Hogan, Julie Goldstein Grumet. "Suicide Prevention: An Emerging Priority For	<1 %

Health Care", Health Affairs, 2016
Publication

58	Vasiliki Voukelatou, Lorenzo Gabrielli, Ioanna Miliou, Stefano Cresci, Rajesh Sharma, Maurizio Tesconi, Luca Pappalardo. "Measuring objective and subjective well-being: dimensions and data sources", International Journal of Data Science and Analytics, 2020 Publication	<1 %
59	digibug.ugr.es Internet Source	<1 %
60	github.com Internet Source	<1 %
61	prisca.unina.it Internet Source	<1 %
62	theses.liacs.nl Internet Source	<1 %
63	"Advanced Computing Technologies and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
64	"Artificial Intelligence in Education", Springer Science and Business Media LLC, 2020 Publication	<1 %

65	"Text, Speech, and Dialogue", Springer Science and Business Media LLC, 2019 Publication	<1 %
66	(7-29-14) http://152.3.140.5/~abhinath/btp.pdf Internet Source	<1 %
67	Hennessy, Eilis, Heary, Caroline, Michail, Maria. "Understanding Youth Mental Health: Perspectives from Theory and Practice", Understanding Youth Mental Health: Perspectives from Theory and Practice, 2022 Publication	<1 %
68	Simon D'Alfonso. "AI in mental health", Current Opinion in Psychology, 2020 Publication	<1 %
69	Sudeep Bhatia, Russell Richie, Wanling Zou. "Distributed semantic representations for modeling human judgment", Current Opinion in Behavioral Sciences, 2019 Publication	<1 %
70	aaltodoc.aalto.fi Internet Source	<1 %
71	artemis.cslab.ece.ntua.gr:8080 Internet Source	<1 %
72	aura.abdn.ac.uk Internet Source	<1 %
73	export.arxiv.org	

	Internet Source	<1%
74	kclpure.kcl.ac.uk Internet Source	<1%
75	mdpi-res.com Internet Source	<1%
76	personales.upv.es Internet Source	<1%
77	www.arxiv-vanity.com Internet Source	<1%
78	"Artificial Neural Networks and Machine Learning – ICANN 2021", Springer Science and Business Media LLC, 2021 Publication	<1%
79	Anshu Malhotra, Rajni Jindal. "Deep learning techniques for suicide and depression detection from online social media: A scoping review", Applied Soft Computing, 2022 Publication	<1%
80	Atreyi Kankanhalli, Yannis Charalabidis, Sehl Mellouli. "IoT and AI for Smart Government: A Research Agenda", Government Information Quarterly, 2019 Publication	<1%
81	Michael Evans. "Recounting the Courts? Applying Automated Content Analysis to	<1%

Enhance Empirical Legal Research : Automated Content Analysis to Enhance Empirical Legal Research", Journal of Empirical Legal Studies, 12/10/2007
Publication

82	Rahul Shidhaye. "Unburden mental health in India: it's time to act now", The Lancet Psychiatry, 2020 Publication	<1 %
83	www.aminer.cn Internet Source	<1 %
84	"Handbook of Health and Well-Being", Springer Science and Business Media LLC, 2022 Publication	<1 %
85	Richard M. Duffy, Brendan D. Kelly. "India's Mental Healthcare Act, 2017", Springer Science and Business Media LLC, 2020 Publication	<1 %
86	Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu et al. "Pre-trained models: Past, present and future", AI Open, 2021 Publication	<1 %
87	privacyinternational.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		

BTP 4

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47