UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

KNOWLEDGE DISTILLATION ON NEURAL NETWORKS FOR ORDINAL SUICIDE IDEATION DETECTION ON SOCIAL MEDIA

Diploma Thesis

Stavros Gazetas

Supervisor: Dimitrios Rafailidis

June 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

KNOWLEDGE DISTILLATION ON NEURAL NETWORKS FOR ORDINAL SUICIDE IDEATION DETECTION ON SOCIAL MEDIA

Diploma Thesis

Stavros Gazetas

Supervisor: Dimitrios Rafailidis

June 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΠΟΣΤΑΞΗ ΓΝΩΣΗΣ ΣΕ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΓΙΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΑΥΤΟΚΤΟΝΙΚΟΥ ΙΔΕΑΣΜΟΥ ΣΤΑ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Διπλωματική Εργασία

Σταύρος Γαζέτας

Επιβλέπων: Δημήτριος Ραφαηλίδης

Ιούνιος 2023

Approved by the Examination Committee:

Supervisor **Dimitrios Rafailidis**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Dimitrios Katsaros**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Michael Vasilakopoulos**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Stavros Gazetas

Diploma Thesis

KNOWLEDGE DISTILLATION ON NEURAL NETWORKS FOR ORDINAL SUICIDE IDEATION DETECTION ON SOCIAL MEDIA

Stavros Gazetas

Abstract

Suicide ideation is a serious and urgent social problem that requires scalable and effective tools to detect it. To provide the appropriate support and intervention, it is crucial to identify those who are having suicidal thoughts as soon as possible. However, the difficulty of efficiently addressing suicidal ideation is made more challenging by the fact that existing models in this field sometimes have significant computational costs, which restricts their accessibility and prevents their widespread deployment. A response-based knowledge distillation (KD) strategy is proposed in an effort to overcome the shortcomings of existing models while also acknowledging the urgency and significance of suicide ideation detection. This method aims to improve the usability and scalability of models for detecting suicidal ideation by lowering computing costs while preserving performance. The knowledge from a high-performing teacher model, SISMO, is condensed into smaller and more computationally effective student models in the suggested response-based KD strategy to overcome these difficulties. SISMO performs better than the state-of-the-art models, especially when it comes to identifying high-risk users, by using an ordinal formulation. The KD process greatly lowers computing costs and the number of parameters needed for detection by transferring the knowledge and performance of the teacher model to the student model, making it possible to use these models on platforms with limited resources. This thesis makes a contribution to the creation of workable and approachable solutions for the detection of suicidal ideation. This development has important ramifications for suicide prevention since it provides prompt identification and support for people experiencing suicidal thoughts.

Keywords:

Suicide ideation detection, Model compression, Knowledge distillation, Machine learning, Deep learning, Teacher-student models, Ordinal formulation, Risk assessment

Διπλωματική Εργασία

ΑΠΟΣΤΑΞΗ ΓΝΩΣΗΣ ΣΕ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΓΙΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΑΥΤΟΚΤΟΝΙΚΟΥ ΙΔΕΑΣΜΟΥ ΣΤΑ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Σταύρος Γαζέτας

Περίληψη

Ο αυτοκτονικός ιδεασμός είναι ένα σοβαρό και επείγον κοινωνικό πρόβλημα που απαιτεί αποτελεσματικά εργαλεία για την ανίχνευσή του. Για την παροχή της κατάλληλης υποστήριξης και παρέμβασης, είναι ζωτικής σημασίας να εντοπίζονται το συντομότερο δυνατό όσοι έχουν αυτοκτονικές σκέψεις. Ωστόσο, η δυσκολία αποτελεσματικής αντιμετώπισης του αυτοκτονικού ιδεασμού καθίσταται πιο δύσκολη από το γεγονός ότι τα υπάρχοντα μοντέλα στον τομέα αυτό έχουν ενίοτε σημαντικό υπολογιστικό κόστος, γεγονός που περιορίζει την προσβασιμότητά τους και εμποδίζει την ευρεία ανάπτυξή τους. Προτείνεται μια στρατηγική απόσταξης γνώσης με βάση την απόκριση σε μια προσπάθεια να ξεπεραστούν οι αδυναμίες των υφιστάμενων μοντέλων, αναγνωρίζοντας παράλληλα τον επείγοντα χαρακτήρα και τη σημασία της ανίχνευσης του ιδεασμού αυτοκτονίας. Η γνώση από ένα μοντέλο δασκάλου με υψηλές επιδόσεις, το SISMO, συμπυκνώνεται σε μικρότερα και πιο αποτελεσματικά από υπολογιστική άποψη μοντέλα μαθητών στην προτεινόμενη στρατηγική με βάση την απόκριση, ώστε να ξεπεραστούν αυτές οι δυσκολίες. Το SISMO αποδίδει καλύτερα από τα μοντέλα τελευταίας τεχνολογίας, ιδίως όσον αφορά τον εντοπισμό χρηστών υψηλού κινδύνου, χρησιμοποιώντας μια διαμορφωμένη διατύπωση. Η διαδικασία μειώνει σημαντικά το υπολογιστικό κόστος και τον αριθμό των παραμέτρων που απαιτούνται για την ανίχνευση μεταφέροντας τη γνώση και την απόδοση του μοντέλου δασκάλου στα μοντέλα μαθητών. Η παρούσα διατριβή συμβάλλει στη δημιουργία εφαρμόσιμων και προσιτών λύσεων για την ανίχνευση του αυτοκτονικού ιδεασμού. Η εξέλιξη αυτή έχει σημαντικές προεκτάσεις για την πρόληψη των αυτοκτονιών, καθώς παρέχει άμεση αναγνώριση και υποστήριξη σε άτομα που βιώνουν αυτοκτονικές σκέψεις.

Λέξεις-κλειδιά:

Ανίχνευση αυτοκτονικού ιδεασμού, Συμπίεση μοντέλου, Απόσταξη γνώσης, Μηχανική μάθηση, Βαθιά μάθηση, Μοντέλα δασκάλου-μαθητή, Εκτίμηση κινδύνου

Table of contents

List of figures

List of tables

Abbreviations

KD	Knowledge Distillation
ML	Machine Learning
DL	Deep Learning
LR	Logistic Regression
SVM	Support Vector Machines
RF	Random Forest
CNN	Convolution Neural Network
LSTM	Long Short Term Memory
BERT	Bidirectional Encoder Representations from Transformers

Chapter 1

Introduction

Suicide is a major issue in the modern world. Many people experience suicidal ideation due to stress and anxiety caused by the fast pace of modern life. The probability that people with suicidal ideation commit suicide is around 30% [1]. This results in more than 700,000 people around the world committing suicide each year [2]. In this era, social media thrives, and people find a way to express their feelings through it. Posts on social media are a very common occurrence, but many of them underlie the hardships that the user is going through. This opens up the path to understanding the user and preventing them from any harmful action. Machine Learning (ML) and specifically Deep Learning (DL) have emerged as methods of interpreting social media posts in an effort to detect ideation and prevent suicide.

There have been many approaches to identifying suicide ideation through social media. O'Dea et al. (2015) [3] proposed using Twitter posts in order to discover patterns that imply suicidal thoughts via classification with Support Vector Machines and Logistic Regression. However, more recent attempts, such as Sawhney et al. (2018) [4] and Aldhyani et al. (2022) [5] have implemented deep learning techniques to tackle this problem (Recurrent Neural Networks, Long Short-Term Memory, Convolutional Neural Networks and Bidirectional Long Short-Term Memory). Deep learning can be a really effective solution to suicide ideation detection. Deep learning architectures perform with very high accuracy without particularly employing complicated techniques. These methods usually outperform classic machine learning approaches, if the text preprocessing is carried out correctly [6].

Neural networks perform at a high rate on text classification, which includes suicide ideation detection. When accuracy needs to be improved, it is very common to increase the number of hidden layers and neurons in the architecture. These changes result in an increase

in the number of the model's parameters (weights and biases associated with each neuron). In deep neural networks, there may be millions of parameters that need to be learned, making the optimization process computationally intensive and challenging. However, these parameters are crucial for the neural network to effectively model complex patterns in the input data and make accurate predictions.

With the prevalence of large neural networks with millions of parameters, model compression is emerging as a promising solution. Knowledge Distillation (KD) is a compression method that shows great potential and effectiveness in tackling this problem [7]. There have been many attempts to utilize KD to significantly reduce the parameters of the model while retaining accuracy. Previous works usually refer to a ”teacher” model, which is a large model with millions of parameters, that is used to train a smaller ”student” model to mimic the teacher's behavior [8, 9, 10]. This way, the knowledge is transferred to the student model, which is able to predict very accurately, despite having significantly fewer parameters.

In the domain of suicide prevention, there have been few attempts to use KD in order to reduce the computational cost. Most models that are used for this purpose are typically very large and complicated. Applying KD to suicide ideation detection models has several potential advantages, such as improved model efficiency, generalization, interpretability, and transferability across different platforms and datasets.

Contribution

This thesis proposes a new distillation framework that is designed to be parameter-efficient. It forges a connection between two separately developing literatures: suicide ideation detection and knowledge distillation. This allows us to use knowledge distillation on the large and complicated models of suicide ideation detection and transfer the knowledge to smaller and more parameter-efficient models. Based on experiments, we can utilize the ground truth labels to diminish the discrepancy in performance between the teacher and student models. In summary, the contributions of this thesis are:

- Proposition of a parameter-efficient knowledge distillation framework on the suicide ideation detection domain.
- Theoretical and experimental evidence shows that the framework can facilitate knowledge transfer by reducing the gap between teacher and student while achieving great accuracy.

Chapter 2

Related Work

2.1 Suicide Ideation Detection Methods

Suicide ideation is a very serious and complex matter with a huge impact on families and communities. Addressing this issue is a priority for many researchers. Detecting suicide ideation can be catalytic for preventing suicide attempts. Various approaches have been made to detect suicide ideation in social media posts. ML and DL techniques have shown very promising results and are commonly used in recent papers.

2.1.1 Machine Learning Methods

Support Vector Machines

SVM is frequently used in the domain of suicide ideation detection [11, 12, 13, 14]. Amini et al. (2016) [13] proposed using SVM in order to assess the high-risk groups for suicide on a dataset derived from a study conducted in Hamadan Province, west of Iran, in 2010 [15]. SVM achieved great accuracy, outperforming other models such as DT, LR, and ANN, especially in the highest risk group.

Logistic Regression

Correspondingly, LR has been employed within the same papers [11, 12, 13, 14]. Specifically, Aladağ et al.(2016) [14] implemented LR in order to detect suicidal ideation in a dataset composed of 508,398 Reddit posts from suicidal subreddits. Along with SVM, LR had the best performance, being the favorite due to its simplicity.

Random Forest

Lastly, another common ML method applied in this domain is RF [11, 14, 16]. RF can perform really well in suicide ideation detection tasks. Usually, RF can match the performance of the most used commonly ML methods in suicide ideation detection, SVM and LR, or even surpass them depending on the task [16].

2.1.2 Deep Learning Methods

Convolution Neural Network

CNN shows great promise in the field of suicide ideation detection [17, 18, 19]. Manas Gaur et al. (2019) [17] proposed a CNN model in an attempt to assess the severity of suicide risk in social media posts, by analyzing a total of 270,000 users and 8 million posts from suicidal subreddits. CNN outperformed the other models (RF, SVM, FFNN) as a result of the use of specific domain knowledge resources and discriminating features.

Long Short Term Memory

LSTM is a really great solution for tackling the problem of suicide ideation detection [18, 19, 20]. It is a widely used model for this matter and can be combined with other models, such as CNN to produce even better results [18]. Lei Cao et al. (2019) proposed SDM, an LSTM model with attention that surpasses other classic ML approaches by performing well on people's implicit and anti-real contrary expressions.

BERT-based

BERT was created to assist computers in comprehending natural language and the context of sentences [21]. This makes it a very useful tool in the domain of suicide ideation detection, and many papers have proposed Bert-based models [22, 23]. Matthew Matero et al. (2019) [22] employed BERT to catch the linguistic features of tweets and fed the results to a RNN-based architecture, which performed better than Bert-based LR.

2.2 Knowledge Distillation Techniques

Neural networks show great promise in the domain of suicide ideation detection, but their large size and complexity make them challenging to deploy. Knowledge distillation is a very common solution to this problem and has been applied several times with different approaches, such as Data-Free KD, Self-Distillation and Channel distillation [24, 25, 26]. While there have been relatively few attempts to use KD for suicide ideation detection, this technique has been a common practice in other very relevant fields like emotion detection, depression detection, and mental health prediction [27, 28, 29].

Recently, Cobeli et al. (2022) [30] introduced the usage of KD in optimism/pessimism detection in tweets. Towards this end, three different strategies were investigated, classic KD, patient KD and multi-task KD [7, 31, 32], by training DL models on other complementary tasks and transferring the knowledge to the student model.

In the field of depression detection, Wu et al. (2023) [33] proposed Mood2Content model, which combines emotional and textual features. Using KD, Mood2Content combines longitudinal contextual information from Twitter posts with the daily emotional status updates of COVID-19 patients to predict their risk of developing depression.

Furthermore, Zeberga et al. (2022) [34] applied KD in the domain of mental health prediction. The knowledge was transferred from a pretrained BERT model to a smaller Distilled_BERT surpassing the other state-of-the-art methods. The application of knowledge distillation (KD) to BERT models has been shown to improve their efficiency and enable them to be deployed on resource-constrained devices [35].

Chapter 3

Methodology

3.1 Proposed Model (SISMO)

The proposed model presented in this thesis is motivated by the work of Ramit Sawhney et al. (2021) [36]. Sawhney introduced SISMO (Suicide Ideation Detection on Social Media using an Ordinal formulation), a hierarchical attention model that accounts for the graded nature of increasing suicide risk levels. In order for this to be achieved, the model utilizes a soft probability distribution approach since each level is of different importance and not all incorrect predictions are flawed to the same degree. SISMO was evaluated on real-world Reddit data annotated by clinical experts. It outperformed all the other methods mentioned in the paper, making it one of the best candidates for tackling the problem of suicide ideation detection on social media.

3.1.1 Defining the Problem

The main objective is to identify the suicide risk level of a particular user through the analysis of their social media posts over time. Based on a modified version of the Columbia-Suicide Severity Rating Scale (C-SSRS) that is suitable for social media, a formal approach is established in order to evaluate the users' suicide risk [37, 38]. According to the scale, there are five different categories a user can belong to. Those categories are Supportive (SU) < Indicator (IN) < Ideation (ID) < Behavior (BR) < Attempt (AT) [36]. Observing the increasing risk between these categories, it can be concluded that their relationship is ordinal. Given this fact, this problem of suicide risk assessment can be addressed as an Ordinal Regression problem, which is essentially a multi-classification task [39]. Each user $u_i \in U = \{u_1, u_2, \ldots, u_n\}$ is

classified based on his posts $P_i = \{p_1^i, p_2^i, \ldots, p_{T_i}^i\}$, which are given in chronological order.

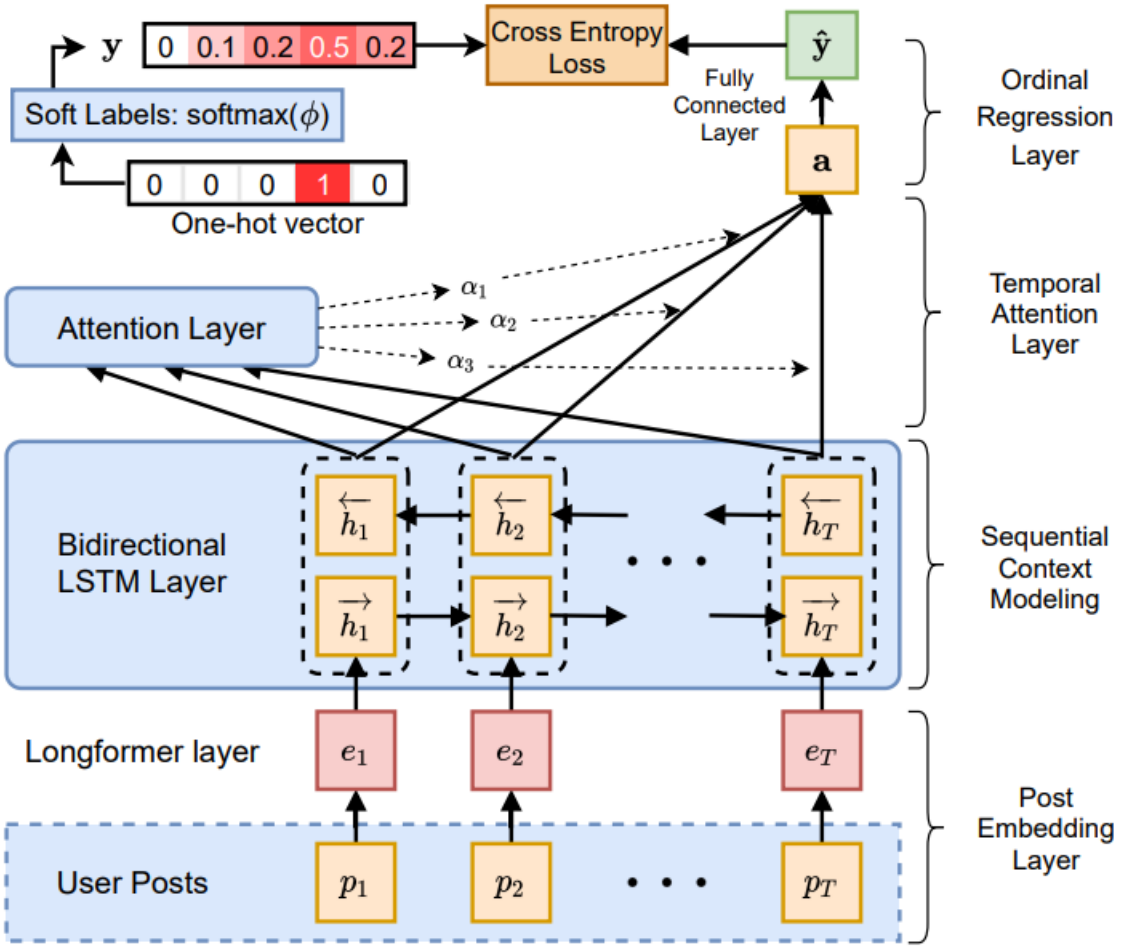


Figure 3.1: SISMO - Model Overview and Components

[36]

3.1.2 SISMO Components

Post Embedding Layer

The first layer of the neural network is the user post embedding layer that incorporates word-level attention [36]. With the intention of gaining insights into the users' mental health, SISMO utilizes Longformer, a pre-trained language model, which generates the post-level embeddings that help discover patterns associated with suicide risk [40]. Each historical post p_t^i is tokenized and then encoded using Longformer, with the encoded representation denoted as $\mathbf{e}_t^i = \text{Longformer}(p_t^i)$.

Sequential Context Modeling

This layer consists of a Bidirectional LSTM, which enables the model to extract time-based features from each user's posts over time. LSTMs are appropriate for modeling a user's social media posts since they are capable of capturing long-term dependencies [41]. The goal is to analyze the posts as a whole and not each one individually, due to the fact that this can provide a more versatile indication of the user's mental health, which can alter over time. Each subsequent post p_t^i is encoded as $\mathbf{e}_t^i$ and transformed into a contextualized representation $\mathbf{h}_t^i$ by considering the context of past and future posts. This is achieved by concatenating the hidden state vectors of the LSTM in both the forward and backward directions [36].

$$\overrightarrow{\mathbf{h}}_t^i = LSTM(\mathbf{e}_t^i, \overrightarrow{\mathbf{h}}_{t-1}^i) \tag{3.1}$$

$$\overleftarrow{\mathbf{h}}_t^i = LSTM(\mathbf{e}_t^i, \overleftarrow{\mathbf{h}}_{t-1}^i) \tag{3.2}$$

$$\mathbf{h}_t^i = [\overrightarrow{\mathbf{h}}_t^i, \overleftarrow{\mathbf{h}}_t^i] \tag{3.3}$$

The BiLSTM network, given that H is the dimension of each $\mathbf{h}_t^i$, converts the encoded posts $[\mathbf{e}_1^i, \ldots, \mathbf{e}_{T_i}^i]$ into contextual representations $[\mathbf{h}_1^i, \ldots, \mathbf{h}_{T_i}^i] \in \mathbb{R}^{H \times T_i}$.

Temporal Attention Layer

As for the attention layer, SISMO utilizes a temporal attention mechanism that is designed to acquire variable weights that are adaptable to each post's contextual representations. The purpose of this mechanism is to emphasize posts containing discernible indicators of suicide risk. These posts are then combined into a larger aggregation.

$$\mathbf{a}_i = \sum_{t=1}^{T_i} a_t^i \mathbf{h}_t^i, a_t^i = \frac{exp(\tilde{a}_t^i)}{\sum_{t=1}^{T_i} exp(\tilde{a}_t^i)} \tag{3.4}$$

$$\tilde{a}_t^i = \mathbf{c}^T tanh(\mathbf{W}_x \mathbf{h}_t^i + \mathbf{b}_x) \tag{3.5}$$

where $\mathbf{W}_x \in \mathbb{R}^{T_i \times H}$, $\mathbf{b}_x \in \mathbb{R}^{T_i}$ and $\mathbf{c} \in \mathbb{R}^{T_i}$ are the network parameters. The attention based representation of a user's posts is denoted with $\mathbf{a}_i$.

Ordinal Regression Layer

The last layer of the model is a fully connected layer with input $\mathbf{a}_i$. The output of this layer is denoted as $\hat{\mathbf{y}}_i = \mathbf{W}_y\mathbf{a}_i + \mathbf{b}_y$, with $\mathbf{W}_y \in \mathbb{R}^{\lambda \times H}$ and $\mathbf{b}_y$ being model parameters. The symbol λ represents the number of the risk levels. The classification confidence vector is produced by utilizing the sequential order of the risk categories.

SISMO is trained by minimizing an ordinal regression loss and using soft labeling in order to represent the ground truth [42]. The soft labels are computed as probability distributions of ground truth labels, denoted as $\mathbf{y} = [y_0, y_1, \ldots, y_4]$, for each true risk level $r_t \in \boldsymbol{\mathcal{Y}} = \{SU = 0, IN = 1, ID = 2, BR = 3, AT = 4\} = \{r_i\}_{i=0}^{4}$. For each risk level r_i the probability is:

$$y_i = \frac{e^{-\phi(r_t, r_i)}}{\sum_{k=1}^{\lambda} e^{-\phi(r_t, r_k)}} \forall r_i \in \boldsymbol{\mathcal{Y}} \tag{3.6}$$

$$\phi(r_t, r_i) = a|r_t - r_i| \forall r_i \in \boldsymbol{\mathcal{Y}} \tag{3.7}$$

The cost function ϕ measures the extent to which the actual risk level r_t differs from the predicted risk level $r_i \in \boldsymbol{\mathcal{Y}}$ and penalizes it accordingly. The parameter a determines the degree of punishment for predicting a risk level that is not accurate.

Once the probability distribution $\mathbf{y}$ for the ground truth label has been computed, the cross-entropy loss is then computed using the classification score $\hat{\mathbf{y}}$.

$$\mathcal{L} = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{\lambda} \mathbf{y}_{ij} log \hat{\mathbf{y}}_{ij} \tag{3.8}$$

3.2 Knowledge Distillation Strategy

The objective of knowledge distillation is to create a smaller student model that can effectively gain the knowledge and insights gathered by a larger teacher model. In this case, the goal is to use SISMO, which is a very complex model with a high number of parameters, and distill the knowledge into a student model with significantly reduced parameters.

In order to distill the knowledge we follow offline distillation [43]. The teacher model (SISMO) is pre-trained, as instructed by the original paper [36]. Then, a variation of SISMO with reduced parameters is created and used as a student model. The student model is trained by adopting a distillation loss function, with the purpose of transferring the knowledge that was acquired during the training of the teacher model.

3.2.1 Response-Based Knowledge Distillation

The knowledge is transferred with a response-based distillation strategy that utilizes the following distillation loss function L^D [44, 45].

$$minL^D = \gamma L^S + (1-\gamma)L^T \tag{3.9}$$

where L^S is the cross-entropy loss of the student model and L^T is the the prediction error of the teacher model in Eq. 3.8 on the training data. The hyper-parameter $\gamma \in [0, 1]$ regulates the relative importance given to the two losses. When γ is set to a high value, the error is biased towards minimizing the errors of the student model, which ignores the knowledge imparted by the teacher model [44]. The distillation loss function L^D allows the student model to emulate the behavior of the teacher, while significantly reducing the model parameters.

3.2.2 Progressive Reduction of Student Model Parameters

To evaluate the effectiveness of knowledge distillation in the proposed SISMO model, three variations of the same model with gradually reduced parameters are created. These models are referred to as SISMO*, SISMO**, and SISMO***, respectively. The initial model has the most extensive set of parameters, followed by the second model with a reduced number of parameters, and finally the third model with the most limited set of parameters. Creating these variations, aims to investigate the impact of reducing model complexity on the knowledge transfer process. Each model is evaluated using the same set of metrics and their performances are compared to that of the original, undistilled model. This analysis provides insights into the effectiveness of knowledge distillation in reducing the model size while maintaining or improving model accuracy.

Chapter 4

Experiments

4.1 Dataset

The data derives from a Reddit annotated dataset of 500 users. The original dataset included 270,000 users with 8 million posts from 15 mental health related subreddits and was processed by Manas Gaur et al. (2019) [38]. The final version of the dataset was developed with the help of four practising psychiatrists. The first step towards that was negation detection, which was used in order to eradicate the non-suicidal users [46]. They proceeded with a user and content overlap analysis, that aimed to remove the redundant users that were very similar to others, as their post content was overlapping. The final users were randomly selected from the remaining users after the above analysis, and each user was annotated in one of the five increasing suicide risk levels under the C-SSRS guidelines. The annotated data was composed of 22% supportive users, 20% users with suicidal indication, 34% users with suicidal ideation, 15% users with suicidal behaviors, and 9% users that attempted suicide.

4.2 Evaluation Protocol

To evaluate the performance of the proposed KD strategy, several evaluation metrics are employed. Specifically, Graded Recall (GR), Graded Precision (GP), and F1-Score are used [47]. GR and GP are modifications of Recall and Precision, which are commonly used in classification tasks. These metrics are calculated through the alterations that are suggested by [38]. In order to integrate the order between the risk levels,the formulations of False Positive (FP) and False Negative (FN) are adjusted as follows:

$$FP = \frac{\sum_{i=1}^{N_T} I(r_i^p > r_i^a)}{N_T}, FN = \frac{\sum_{i=1}^{N_T} I(r_i^a > r_i^p)}{N_T} \tag{4.1}$$

where N_T is the size of the test data, r^a is the actual risk level, and r^p is the predicted risk level. Therefore, FP is defined as the ratio of the number of times r^p is greater than r^a over N_T and FN is defined as the ratio of the number of times r^p is less than r^a over N_T.

Graded Recall

This metric measures how many of the actual positive observations are correctly identified by the algorithm. GR can be computed by dividing the number of true positives (TP) by the sum of true positives and false negatives [47].

$$GR = \frac{TP}{TP + FN} \tag{4.2}$$

Graded Precision

This metric refers to the number of observations that the algorithm has predicted as positive and, among them, how many are actually positive. GP can be computed by dividing the number of true positives by the sum of true positives and false positives [47].

$$GP = \frac{TP}{TP + FP} \tag{4.3}$$

F1-Score

This metric is used to provide an overall measure of the model's performance by combining the GP and GR scores. F1-Score is the harmonic mean of GP and GR [47].

$$F1\text{-}Score = 2 \times \frac{GP \times GR}{GP + GR} \tag{4.4}$$

4.3 Models

4.3.1 Examined models

For the purpose of assessing the performance and effectiveness of the proposed approach, several models are selected. In addition to comparing the distilled models with the original, a number of state-of-the-art methods, that have gained prominence in the field are introduced. These models include the following:

- **SVM+RBF [13]**: Using TF-IDF (Term Frequency Inverse Document Vectorization),a language vector is generated for each user's posts [48]. This vector representation is then fed into a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. The RBF kernel is parameterized with $\sigma = 0.24$, and the SVM is trained with a cost parameter $c = 5$. This configuration will allow the textual data to be efficiently processed and classified by the model.

- **Random Forest [49]**: The language vectors, produced through TF-IDF vectorization, were utilized as input for Random Forest. In this approach, Random Forest is applied with 500 trees, where each tree node consists of 9 predictor variables.

- **BERT+BiLSTM [50]**: The combination of BERT and BiLSTM is employed to enhance the representation of the posts. The embeddings that are produced by BERT are fed to a BiLSTM layer in order to capture the sequential dependencies of the posts. Furthermore, an attention layer is applied to highlight important features in the contextual representations.

- **BiLSTM+Attn [51]**: The BiLSTM+Attn model incorporates Longformer embeddings in its architecture. Longformer embeddings are used for the collection of longline dependencies and contextual information in posts. The model uses a BiLSTM layer to process the embeddings and extract sequential patterns from the data. In addition, an attention mechanism is used for highlighting important characteristics and improving the representation of the posts.

4.3.2 Experimental Models and Settings

The experimental models encompass the teacher (SISMO) and student (SISMO*, SISMO**, and SISMO***) models that are used in the KD procedure. Each student model is composed of distinct configurations based on the original teacher model.

- **SISMO**: The teacher model is fine-tuned to achieve optimal performance. The selected hyperparameters are: hidden layer size $H = 512$, number of BiLSTM layers $n = 2$, learning rate $lr = 0.01$, dropout $\delta = 0.3$ and control parameter $\alpha = 1.8$.
- **SISMO***: The first student model adapts the same hyperparameters except for the number of BiLSTM layers and the hidden layer size, having $n = 1$ and $H = 128$. This achieves a highly reduced set of parameters in comparison with the teacher model.
- **SISMO****: The second student model aims to further reduce the number of parameters by decreasing the hidden layer size H to 64.
- **SISMO*****: The third student model, similarly, has an even smaller number of hidden layer size $H = 32$.

The base version of Longformer is fine-tuned using Hugging-face's Transformers library. All the above models are implemented with PyTorch and optimized using mini-batch AdamW with a batch size of 8. The models are trained for 50 epochs, including early stopping with patience of 5 epochs, as instructed by [36]. The median testing performance is reported over 15 runs, meaning that 15 teacher (pre-trained) models are created in order to train the same number of models for each student. Each run is conducted with random initialization.

4.4 Performance Evaluation

In table 4.1 the performance of all the models is reported. SISMO significantly outperforms all the examined models. This superior performance is primarily the reason SISMO is selected as the teacher model. Although SISMO exhibits superior performance, it becomes apparent that the model incurs a high computational cost, as evidenced by the number of parameters, which is 12.606 million. The proposed KD approach aims to reduce these parameters while maintaining performance.

Model	Parameters (M)	Compression Rate	Evaluation Metrics		
			Graded Precision	Graded Recall	FScore
SVM+RBF [13]	4.836	—	0.565	0.566	0.555
Random Forest [49]	0.179	—	0.545	**0.666**	0.601
BERT+BiLSTM [50]	2.135	—	0.619	0.525	0.581
BiLSTM+Attn [51]	6.305	—	**0.722**	0.538	0.591
SISMO [36]	12.606	—	0.690	0.587	*0.639*
SISMO*	0.986	93.2%	0.681	*0.590*	*0.639*
SISMO**	0.444	96.5%	*0.705*	0.587	**0.648**
SISMO***	0.209	98.4%	0.681	0.569	0.630

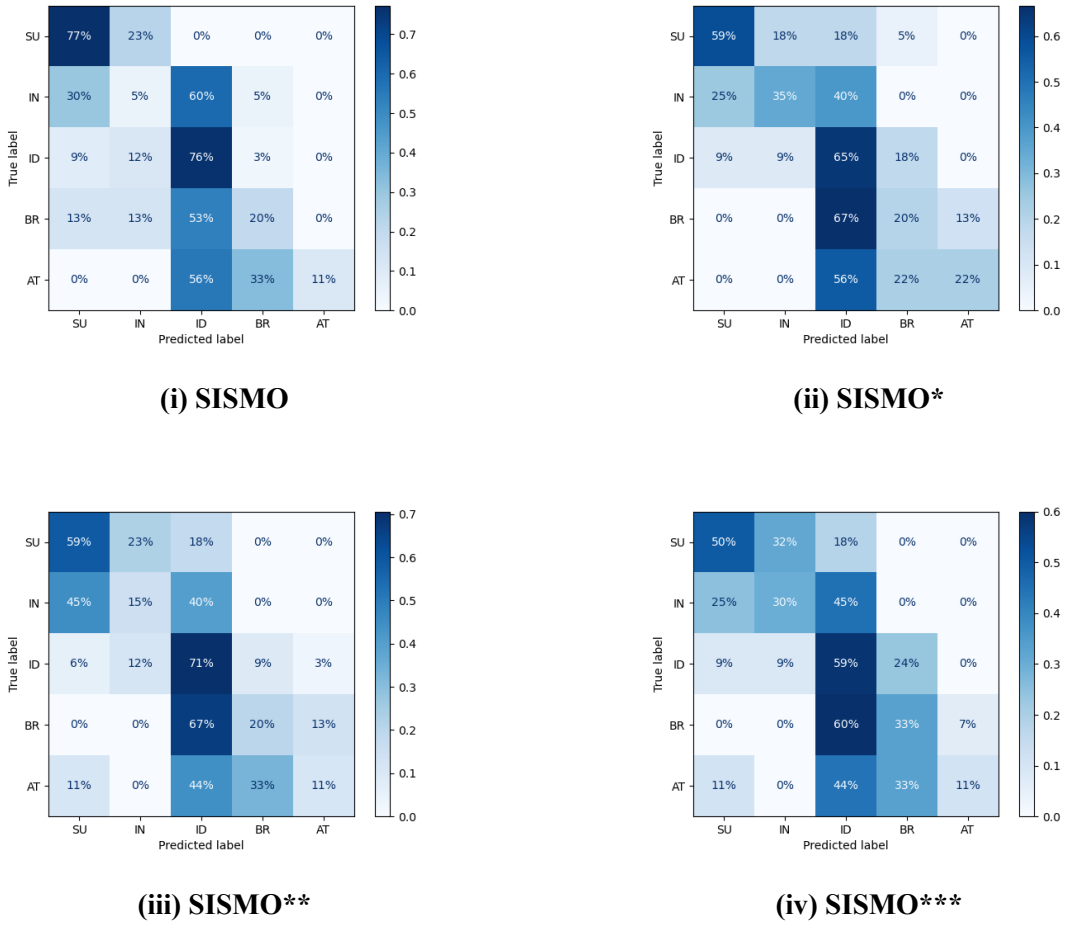
Table 4.1: Model Performance

The median of the results over 15 runs is reported. The student models' median performance is reported for the best γ for each student. Bold denotes the best performance. Italics denote second best. The number of parameters is denoted in millions.

The student models, namely SISMO*, SISMO**, and SISMO***, demonstrate the effectiveness of knowledge distillation in achieving model compression. Through KD, these student models are able to capture and transfer the knowledge from the teacher model while also achieving a reduced parameter count. The performance of the teacher model is maintained by the students, which indicates that the knowledge and decision-making capabilities of the teacher model have been precisely distilled. The student models have a significantly lower number of parameters compared to the teacher. This fact adds up to the effectiveness of the proposed approach, as the student models can match the teacher with limited computational resources.

It is observed that SISMO performs better than the other state-of-the-art models at higher risk levels [36]. This enhancement can be explained by SISMO's ordinal formulation, which punishes predictions based on how far off the anticipated risk is from the actual risk as measured by the C-SSRS scale. The other models, in contrast, take into account every misclassification equally without taking the size of the divergence into account. In Fig. 4.1i it is shown that the misclassified users that belong to AT are classified either in ID or BR, which indicates

that SISMO prioritizes the high risk users.



(i) SISMO **(ii) SISMO***

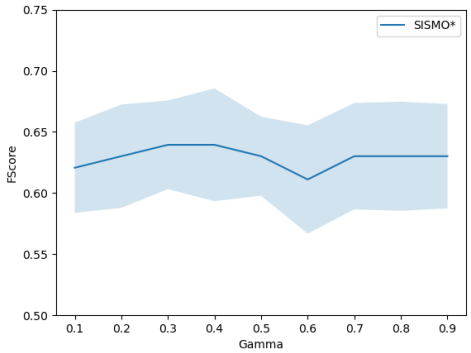
(iii) SISMO** **(iv) SISMO*****

Figure 4.1: Confusion Matrices

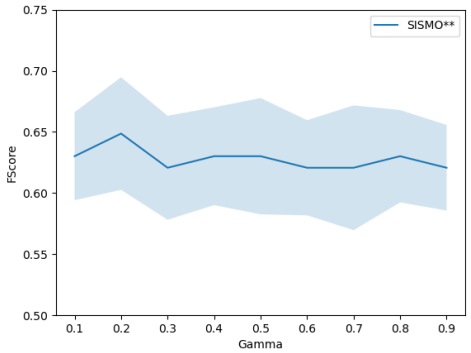
Normalized Confusion matrix for the median of the test results. Instances of the actual risk level are represented by rows, whereas instances of the predicted risk level are represented by columns. The percentage of accurate predictions is represented by the values of the diagonal elements.

Following the behavior of the teacher, the student models act in a comparable manner. The first student, SISMO*, surpasses the teacher in the AT class prediction, which means that the application of KD assists the student in the prediction of high-risk users. It can also be observed that reducing the parameters of the student model results in less accurate predictions at higher risk levels. While SISMO** and SISMO*** match the teacher's performance, SISMO* utilizes KD better than the other student models.

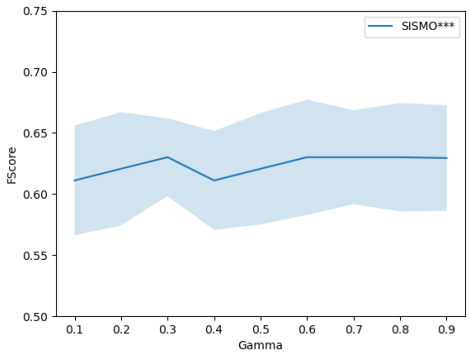
4.5 Parameter Tuning



(i) SISMO*



(ii) SISMO**



(iii) SISMO***

Figure 4.2: Impact of parameter γ on the prediction accuracy.

The impact of the hyperparameter γ (Eq. 3.9) on the performance of the student models is presented in Fig. 4.2. The value of γ ranges from 0.1 to 0.9, with a 0.1 step. This aids in the comprehension of how gamma affects the distillation process and how the teacher influences the students. The median FScore of 15 runs is reported for each γ. For SISMO* the best γ value is found to be 0.3 and 0.4 (the model performs equally well for both γ values), for SISMO** it is 0.2 and for SISMO*** it is 0.3. It is observed that for the value of 0.1 the distillation process focuses on the teacher, meaning that the student's training is limited to the teacher's knowledge. The best performances are reported for γ values in the range of 0.2 to 0.4. As for the higher γ values, the performance of the distilled models is reduced because the student model receives limited guidance from the teacher model during training.

Chapter 5

Conclusion

In this work, a response-based knowledge distillation (KD) approach is proposed for model compression in the domain of suicide ideation detection. The overriding aim is to decrease the computational cost while maintaining the teacher's performance. Several experiments are conducted using various models, including state-of-the-art methods, and their performances are compared with our proposed teacher model (SISMO). The results demonstrate that SISMO, with regard to evaluation metrics such as Graded Precision, Graded Recall and F1 Score, is superior to other models. In particular, the ordinal formulation ,which focuses on high risk users, makes it suitable for this specific task.

However, SISMO is expensive in terms of computational cost and has a large number of parameters. In order to address this, KD is applied to transfer the knowledge from the teacher model into smaller student models (SISMO*, SISMO**, and SISMO***). The knowledge is able to be distilled in an efficient manner, while the number of parameters is greatly reduced. The performance of the teacher model is maintained by the students, indicating the success of the knowledge distillation process. This research indicates that knowledge distillation is an efficient approach for model compression in suicide ideation detection. By compressing the teacher model into smaller student models, a significant reduction in computational cost is achieved while preserving performance. This has practical implications, as the student models can be deployed with limited computational resources, making them more accessible and scalable.

Future Direction

Knowledge Distillation (KD) is a promising strategy with enormous potential for the future of suicide ideation detection. The use of KD can be improved even more as this field of study and development advances by utilizing advanced techniques like continual learning, multimodal fusion, and context-aware modeling. The resulting advances will make it possible to develop more precise and reliable models that can identify subtle patterns and environmental cues linked to suicidal ideation. Additionally, combining KD with other cutting-edge technologies like affective computing and natural language processing can aid in early intervention and provide a deeper understanding of people's mental states.

The future of KD in detecting suicidal ideation lies in its ongoing improvement and integration with advanced methods, which will eventually result in more efficient and widely available options for suicidal prevention and support. This method helps mental health professionals better understand the elements influencing suicidal thoughts while also lowering computational needs and improving model interpretability. KD has the potential to increase scalability, democratize access to reliable detection technologies, and offer prompt assistance to people who are contemplating suicide.

Bibliography

[1] Schreiber Jennifer and Culpepper Larry. Suicidal ideation and behavior in adults, 2022.

[2] World Health Organization. Suicide prevention. https://www.who.int/health-topics/suicide, 2023.

[3] Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188, 2015.

[4] Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[5] Theyazn H. H. Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad A. T. Ahmed. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International Journal of Environmental Research and Public Health*, 19(19):12635, Oct 2022.

[6] Rezaul Haque, Naimul Islam, Maidul Islam, and Md Manjurul Ahsan. A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. *Technologies*, 10(3):57, Apr 2022.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[8] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 7037–7049, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Wenxian Shi, Yuxuan Song, Hao Zhou, Bohan Li, and Lei Li. Learning from deep model via exploring local targets, 2021.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[11] Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha D. Badal, K. P. Subbalakshmi, R. Chandramouli, and Ellen Lee. Learning models for suicide prediction from social media posts, 2021.

[12] Cheng Q, Li TM, Kwok CL, Zhu T, and Yip PS. Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study. *J Med Internet Res*, July 2017.

[13] Amini P, Ahmadinia H, Poorolajal J, and Moqaddasi Amiri M. Evaluating the high risk groups for suicide: A comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iran J Public Health*, 45:1179–1187, Sep 2016.

[14] Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O, and Bingol HO. Detecting suicidal ideation on forums: Proof-of-concept study. *J Med Internet Res*, Jun 2016.

[15] Behzad Amiri, Abolghasem Pourreza, Abbas Rahimi Foroushani, Seyyed Mohammad Hosseini, and Jalal Poorolajal. Suicide and associated risk factors in hamadan province, west of iran, in 2008 and 2009. *Journal of Research in Health Sciences*, 12(2), 2012.

[16] Shaoxiong Ji, Celina Ping Yu, Sai fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

[17] Knowledge-aware assessment of severity of suicide risk for early intervention, 2019.

[18] Priyamvada B., Singhal S., and Nayyar A. Stacked cnn - lstm approach for prediction of suicidal ideation on social media. *Multimedia Tools and Applications*, 2023.

[19] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online, November 2020. Association for Computational Linguistics.

[20] Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China, November 2019. Association for Computational Linguistics.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[22] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[23] Ulya Bayram and Lamia Benhiba. Determining a person's suicide risk by voting on the short-term history of tweets for the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 81–86, Online, June 2021. Association for Computational Linguistics.

[24] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks, 2017.

[25] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022.

[26] Zaida Zhou, Chaoran Zhuge, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation, 2020.

[27] Mahshid Hosseini and Cornelia Caragea. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[28] Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multitask emotion recognition model with knowledge distillation and task discriminator, 2022.

[29] Jorge Pool-Cen, Hugo Carlos-Martínez, Gandhi Hernández-Chan, and Oscar Sánchez-Siordia. Detection of depression-related tweets in mexico using crosslingual schemes and knowledge distillation. *Healthcare*, 11(7), 2023.

[30] Ștefan Cobeli, Ioan-Bogdan Iordache, Shweta Yadav, Cornelia Caragea, Liviu P. Dinu, and Dragoș Iliescu. Detecting optimism in tweets using knowledge distillation and linguistic analysis of optimism. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2032–2041, Marseille, France, June 2022. European Language Resources Association.

[31] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019.

[32] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. Bam! born-again multi-task networks for natural language understanding, 2019.

[33] Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. Exploring social media for early detection of depression in COVID-19 patients. In *Proceedings of the ACM Web Conference 2023*. ACM, apr 2023.

[34] Kamil Zeberga, Muhammad Attique, Babar Shah, Farman Ali, Yalew Zelalem Jembre, and Tae-Sun Chung. A novel text mining approach for mental health prediction using bi-lstm and bert model. *Computational Intelligence and Neuroscience*, 2022.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[36] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. Towards ordinal suicide ideation detection on social media. In *Proceedings of 14th ACM International Conference On Web Search And Data Mining*, WSDM '21, New York, NY, USA, March 2021. Association for Computing Machinery.

[37] Kelly Posner, Gregory Brown, Barbara Stanley, David Brent, Kseniya Yershova, Maria Oquendo, Glenn Currier, Glenn Melvin, Laurence Greenhill, Sa Shen, and John Mann. The columbia-suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, 168(12):1266 – 1277, 2011.

[38] Manas Gaur, Ugur Kursuncu, Amit Sheth, Amanuel Alambo, Krishnaprasad Thirunarayan, Randon S. Welton, Joy Prakash Sain, Ramakanth Kavuluru, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, pages 514–525, May 2019.

[39] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In Luc De Raedt and Peter Flach, editors, *Machine Learning: ECML 2001*, pages 145–156, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[40] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[42] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019.

[43] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, mar 2021.

[44] Stefanos Antaris, Dimitrios Rafailidis, and Sarunas Girdzijauskas. Knowledge distillation on neural networks for evolving graphs. *Social Network Analysis and Mining*, 11(1), oct 2021.

[45] Tao Feng and Mang Wang. Response-based distillation for incremental object detection, 2021.

[46] George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[47] David M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, 2020.

[48] Shahzad Qaiser and Ramsha Ali. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29, jul 2018.

[49] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[50] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *Lecture Notes in Computer Science*, pages 436–447. Springer International Publishing, 2021.

[51] Puneet Mathur, Ramit Sawhney, and Rajiv Ratn Shah. Suicide risk assessment via temporal psycholinguistic modeling (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13873–13874, Apr. 2020.