Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media

Loukas Ilias[ID], Spiros Mouzakitis[ID], and Dimitris Askounis[ID]

Abstract— In today's fast-paced world, the rates of stress and depression present a surge. People use social media for expressing their thoughts and feelings through posts. Therefore, social media provide assistance for the early detection of mental health conditions. Existing methods mainly introduce feature extraction approaches and train shallow machine learning (ML) classifiers. For addressing the need of creating a large feature set and obtaining better performance, other research studies use deep neural networks or language models based on transformers. Despite the fact that transformer-based models achieve noticeable improvements, they cannot often capture rich factual knowledge. Although there have been proposed a number of studies aiming to enhance the pretrained transformer-based models with extra information or additional modalities, no prior work has exploited these modifications for detecting stress and depression through social media. In addition, although the reliability of a machine learning (ML) model's confidence in its predictions is critical for high-risk applications, there is no prior work taken into consideration the model calibration. To resolve the above issues, we present the first study in the task of depression and stress detection in social media, which injects extra-linguistic information in transformer-based models, namely, bidirectional encoder representations from transformers (BERT) and Mental-BERT. Specifically, the proposed approach employs a multimodal adaptation gate for creating the combined embeddings, which are given as input to a BERT (or MentalBERT) model. For taking into account the model calibration, we apply label smoothing. We test our proposed approaches in three publicly available datasets and demonstrate that the integration of linguistic features into transformer-based models presents a surge in performance. Also, the usage of label smoothing contributes to both the improvement of the model's performance and the calibration of the model. We finally perform a linguistic analysis of the posts and show differences in language between stressful and nonstressful texts, as well as depressive and nondepressive posts.

Index Terms— Calibration, depression, emotion, mental health, stress, transformers.

I. INTRODUCTION

ACCORDING to the World Health Organization (WHO) [1], stress can be defined as any type of change that causes physical, emotional, or psychological strain. Stress comes in a number of categories [2], namely, physical, psychological, psychosocial, and psychospiritual stress. Exces-sive stress can lead to anxiety disorders or even depression. Depression entails a great number of symptoms, including loss of interest, anger, pessimism, changes in weight, feelings of worthlessness, thoughts of suicide, and many more. According to the WHO [3], around 280 million people in the world have depression. China, India, the United States, Russia, Indonesia, and Nigeria are some of the countries presenting the highest rates of depression [4]. People with stress and depression use social media platforms, including Twitter and Reddit, and share their thoughts, emotions, feelings, and so on through posts or comments with other users. Therefore, social media constitute a valuable form of information, where linguistic patterns of depressive/stressful posts can be investigated.

Existing research initiatives exploit social media data for identifying depressive and stressful posts. The majority of these research works [5], [6] employ feature extraction approaches and train shallow ML algorithms. Employing feature extraction approaches constitutes a tedious procedure and demands domain expertise, since the authors may not find the optimal feature set for the specific problem. At the same time, the train of shallow ML algorithms does not yield optimal performance and does not generalize well to new data. For addressing these limitations, other approaches [7] use deep neural networks, including convolutional neural networks (CNNs), BiLSTMs, and so on, or transformer-based networks. In addition, there are research studies employing ensemble strategies [8]. However, these approaches increase substantially the training time, since multiple models must be trained separately. In addition, recently there have been studies [9], [10] showing that transformer-based models struggle or fail to capture rich knowledge. For this reason, there have been proposed methods for enhancing these models with external information or additional modalities [11], [12], [13], [14]. However, existing research initiatives in the tasks of stress and depression detection through social media have not exploited any of these approaches yet. In addition, the reliability of an ML model's confidence in its predictions, denoted as calibration [15], [16], is critical for high-risk appli-cations, such as deciding whether to trust a medical diagnosis prediction [17], [18], [19]. Although methods regarding the confidence of models' predictions have been introduced in many studies, including suicide risk assessment [20], sleep stage classification [21], and so on, no prior work for stress and depression detection has taken into account the level of confidence of models' predictions, creating in this way overconfident models.

To tackle the aforementioned limitations, we propose a method, which injects extra-linguistic information into transformer-based models, namely, bidirectional encoder representations from transformers (BERT) and MentalBERT. First, we extract various linguistic features, including NRC Sentiment Lexicon, features derived by latent Dirichlet allocation (LDA) topics, Top2vec, and linguistic inquiry and word count (LIWC) features. Regarding the LDA topic-based features, this is the first study in terms of the tasks of stress and depression detection via social media texts utilizing the Global Outlier Standard Score (GOSS) [22], which captures the text's interest on a specific topic in comparison with other texts. After passing each text through a transformer-based model, we project the linguistic information to the same dimensionality as the outputs of the transformer models. Next, we concatenate the representations obtained by BERT (or MentalBERT) and linguistic information and apply a multimodal adaptation gate [23], where an attention-gating mechanism is used for controlling the importance of each representation. Similar to [24], we modify multimodal BERT (M-BERT) [23] by replacing the multimodal information with linguistic information. Finally, a shifting component is exploited for calculating the new combined embeddings. The new combined embeddings are passed through a BERT (or MentalBERT) model, where the classification [CLS] token is fed to dense layers for getting the final prediction. In addition, for preventing models from becoming too overconfident, we use label smoothing. According to Müller et al. [25], label smoothing has been used successfully to improve the accuracy of deep-learning models across a range of tasks, while at the same time, it implicitly calibrates learned models so that the confidences of their predictions are more aligned with the accuracies of their predictions. We use metrics for assessing both the performance and the calibration of our model. We also demonstrate the efficiency of label smoothing in both calibrating and enhancing the performance of our model. We test our proposed approaches on three publicly available datasets, which differentiate: 1) stressful from nonstressful texts; 2) depressive from nondepressive posts; and 3) posts indicating the severity of depression, namely, minimal, mild, moderate, and severe. We demonstrate the robustness of our model and advantages over state-of-the-art approaches. Finally, we conduct an extensive linguistic analysis and show common linguistic patterns between stress and depression.

Our main contributions can be summarized as follows.

1) We introduce a method, which injects linguistic features into transformer-based neural models.
2) We perform model calibration by using label smoothing. We evaluate the calibration of our approaches by using two metrics. To the best of our knowledge, this is the first study exploiting label smoothing and utilizing calibration metrics.
3) We contribute to the existing literature by performing a detailed linguistic analysis, which reveals significant differences in language between stressful/depressive and nonstressful/nondepressive posts.

II. RELATED WORK

A. Stress Detection

Existing research initiatives build on feature extraction and train of shallow ML algorithms. Muñoz and Iglesias [26] introduced three approaches for detecting psychological stress in a social media text. In terms of the first approach, the authors extracted affective, social, syntactic, and topic-related features. Support vector machines (SVMs), logistic regression (LR), and stochastic gradient descent (SGD) classifiers were trained. Regarding the second approach, the authors employed word embeddings, namely, word2vec, GloVe, and fastText, and trained the abovementioned ML classifiers. With regard to the third approach, the authors introduced an early fusion approach, where they concatenated the features and the word embeddings and trained the aforementioned classifiers. Guntuku et al. [6] extracted LIWC, topic, TensiStrength, and engagements, i.e., time of posts, number of posts, number of posts between 12 A.M.–6 A.M. features and trained linear regression with regularization methods, including ridge, elastic-net, least absolute shrinkage and selection operator (LASSO), and L2 penalized SVMs. They found that elastic-net showed marginally superior performance over the others. The authors proposed also domain adaptation methods, including EasyAdapt and transfer component analysis.

Deep learning approaches and transformer-based models are being used for detecting stress. For instance, Turcan and McKeown [27] introduced a dataset consisting of stressful and nonstressful text and applied ML techniques for differentiating stressful from nonstressful texts. Specifically, the authors extracted lexical, syntactic, social media features, word2vec, and BERT embeddings and trained various ML classifiers, including SVMs, LR, Naive Bayes, perceptron, and decision trees (DTs). Also, deep neural networks were trained, including a two-layer bidirectional gated recurrent neural network (GRNN), CNN, and BERT. Yang et al. [28] introduced a deep neural network to detect stress and depression in a social media text. Specifically, the authors exploited MentalRoBERTa for obtaining token-level embeddings and commonsense transformers (COMET) for extracting mental state knowledge. Next, the authors exploited a GRU layer along with a scaled dot-product attention module. Finally, contrastive learning in a supervised manner was employed by the authors for making sentences with the same label cohesive and different labels mutually exclusive. Also, methods for increasing the training set have been applied. Specifically, Winata et al. [29] proposed (bi)LSTMs coupled with an attention mechanism to classify psychological stress from self-conducted interview transcriptions. For expanding the size of the corpus, the authors applied distant supervision, where they automatically labeled tweets based on their hashtag content.

Hybrid models have also been proposed. For instance, Lin et al. [30] experimented with capturing the users' social interactions in social media. Specifically, the authors proposed a deep neural network consisting of CNNs with cross autoencoders and a partially labeled factor graph. The authors exploited both tweet-level and user-level attributes.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ILIAS et al.: CALIBRATION OF TRANSFORMER-BASED MODELS FOR IDENTIFYING STRESS AND DEPRESSION 3

In terms of the tweet-level attributes, they exploited linguistic, i.e., positive and negative emotion words, punctuation marks, emoticons, and so on, visual, i.e., saturation, brightness, five-color theme, and so on, and social features, i.e., number of comments, retweets, and likes. Regarding user-level attributes, they extracted features pertinent to their posting behavior and social interaction.

Emotion-enhanced approaches in conjunction with explainability techniques have been introduced. Turcan et al. [31] proposed three emotion-enhanced models that incorporate emotional information in various ways to enhance the task of binary stress prediction. In terms of the first approach, the authors exploited two single-task models sharing the same BERT representation layers. Specifically, the authors trained the stress task with the Dreaddit data and the emotion task with the GoEmotions or Vent data. Regarding the second model, the authors proposed a multitask learning framework, where the two tasks were performed at the same time. With regards to the third model, the authors fine-tune first a BERT model using the emotion task and then apply this fine-tuned BERT model to stress detection. Finally, the authors introduced a new framework for model interpretation using local interpretable model-agnostic explanations (LIME).

B. Depression Detection

Some studies have focused on the extraction of features and then the train of shallow ML classifiers. For instance, Tadesse et al. [5] extracted n-grams via the tf–idf approach, LIWC features, and LDA topics. Then, they trained LR, SVM, random forest (RF), AdaBoost, and multilayer perceptron (MLP). Results showed that the bigram features trained on an SVM classifier achieved 80.00% accuracy, while the best accuracy accounting for 91.00% was achieved by exploiting the MLP classifier with all the features, i.e., LIWC, LDA, and bigrams. Liu and Shi [32] extracted a set of textual features, namely, part-of-speech, emotional words, personal pronouns, polarity, and so on, and a set of features indicating the posting behavior of the user, i.e., posting habits and time. Next, feature selection techniques were applied, including recursive elimination, mutual information, and extreme random tree. Finally, Naive Bayes, the k-nearest neighbor, regularized LR, and SVM were used as base learners, and a simple LR algorithm was used as a combination strategy to build a stacking model. Nguyen et al. [33] extracted a set of features, including LDA topics, LIWC features, and affective features by using the affective norms for english words (ANEW) lexicon, and mood labels. The authors trained a LASSO regression classifier for detecting depressive posts and analyzing the importance of each feature. The authors applied also statistical tests and found significant differences between depressive and nondepressive posts. Tsugawa et al. [34] extracted features and trained an SVM classifier to detect depression in Twitter. Specifically, the authors extracted the frequency of words used in tweets, the ratio of tweet topics found by LDA, the ratio of positive and negative words, and many more. Pirina and Çöltekin [35] collected several corpora and trained an SVM classifier using character and word n-grams. Doc2vec and tf–idf features were extracted and given as input to AdaBoost, LR, RF, and SVM for identifying the severity of depression.

Recently, deep learning approaches have been introduced, since they obtain better performance than the traditional ML algorithms and do not often require the tedious procedure of feature extraction. For example, Wani et al. [36] represented words as word2vec and tf–idf approach and trained a deep neural network consisting of CNNs and LSTMs. Kim et al. [37] collected a dataset consisting of posts written by people, who suffer from mental disorders, including depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism. This study developed six binary classification models for detecting mental disorders, i.e., depression versus nondepression, and so on. Specifically, the authors utilized the tf–idf approach and trained an XGBoost classifier. Next, the authors used the word2vec and trained a CNN model. Naseem et al. [38] reformulated depression identification as an ordinal classification problem, where they used four depression severity levels. The authors introduced a deep neural network consisting of a text graph convolutional network, bidirectional long short-term memory (BiLSTM), and attention layer. A similar approach was proposed by Ghosh and Anwar [39], where the authors extracted features and trained LSTMs for estimating the depression intensity levels. A hybrid deep neural network consisting of CNN and BiLSTM was introduced by Kour and Gupta [40]. Zogan et al. [41] introduced the first dataset including posts from users with and without depression during COVID-19 and presented a new hierarchical CNN. An emotion-based attention network model was proposed by Ren et al. [42], where the authors extracted the positive and negative words and passed them through two separate BiLSTM layers followed by attention layers.

Ensemble strategies have also been explored in the literature. This means that multiple models are trained separately and the final decision is taken usually by a majority voting approach. For instance, an ensemble strategy was introduced by Ansari et al. [8]. First, the authors exploited some sentiment lexicons, including AFINN, NRC, SenticNet, and multiperspective question answering (MPQA), extracted features, and applied principal component analysis for reducing the dimensionality of the feature set. An LR classifier was trained using the respective feature set. Next, the authors trained an LSTM neural network coupled with an attention mechanism. Finally, the authors combined the predictions of these two approaches via an ensemble method. Also, an ensemble approach was proposed by Trotzek et al. [43]. First, the authors trained an LR classifier using input user-level linguistic metadata. Specifically, the authors extracted LIWC features, the length of the text, four readability scores, and so on. Next, the authors trained a CNN model. Finally, the authors combined the outputs of these approaches via a late fusion strategy, i.e., by averaging the predictions of the classifiers. Figuerêdo et al. [7] designed a CNN along with early and late fusion strategies. Specifically, the authors exploited fastText and GloVe embeddings. In the early fusion approach, multiple-word embeddings were concatenated and passed to the CNN model. In the late fusion strategy, a majority-vote approach was performed based on the predictions of multiple CNN models. The CNN model comprised a simple convolution

layer, max-pooling, fully connected layers, and concatenated rectified linear units as the activation function.

Explainable approaches have also been introduced. Souza et al. [44] introduced a stacking ensemble neural network, which addresses a multilabel classification task. Specifically, the proposed architecture consists of two levels. In the first level, binary base classifiers were trained with two distinct roles, i.e., expert and differentiating. The expert base classifiers were used for differentiating between users belonging to the control group and those diagnosed with anxiety, depression, or comorbidity. The differentiating base models aimed at distinguishing between two target conditions, e.g., anxiety versus depression. In the second level, a meta-classifier uses the base models' outputs to learn a mapping function that manages the multilabel problem of assigning control or diagnosed labels. The authors used LSTMs and CNNs. Finally, this study explored Shapley additive explanations (SHAP) metrics for identifying the influential classification features. Zogan et al. [45] proposed also an explainable approach, where textual, behavioral, temporal, and semantic aspect features from social media were exploited. A hierarchical attention network was used in terms of explainable purposes. A hierarchical attention network was also used by Uban et al. [46], where the authors extracted a feature set consisting of content, style, LIWC, and emotions/sentiment features. An interpretable approach was proposed by Song et al. [47], where the authors introduced the feature attention network. The feature attention network consists of four feature networks, each of which analyzes posts based on an established theory related to depression and a postlevel attention on top of the networks. However, this method did not attain satisfactory results.

Recently, transformer-based models have been applied to the task of depression detection in social media. Specifically, Boinepelli et al. [48] introduced a method for finding the subset of posts that would be a good representation of all the posts made by the user. First, they employed BERT and computed the embeddings for all posts made by the user. Next, they used a clustering and ranking algorithm. After finding the representative posts per user, the authors added domain-specific elements by exploiting RoBERTa. Finally, the authors experimented with two ways of diagnosing depression, i.e., by either employing a majority-vote approach or training a hierarchical attention network. Anantharaman et al. [49] fine-tuned a BERT model for classifying the signs of depression into three labels, namely, "not depressed," "moderately depressed," and "severely depressed." Similarly, Nilsson and Kovács [50] exploited a BERT model and used abstractive summarization techniques for data augmentation. Zogan et al. [51] presented an abstractive–extractive automatic text summarization model based on BERT, k-means clustering, and bidirectional auto-regressive transformers (BART). Then, they proposed a deep learning framework, which combines user behavior and user post history or user activity.

Multimodal approaches combining both text and images have also been proposed. For instance, a multimodal approach was introduced by Ghosh et al. [52] for detecting depression in Twitter. Specifically, the authors utilized the user's description and profile image. The authors used the IBM Watson NaturalLanguageUnderstanding tool and extracted sentiment and emotion information for all user descriptions along with the possible categories (at most three) that the description may belong to. Next, the authors designed a neural network consisting of BiGRU, attention layers, convolution layers, and dense layers. The authors used GloVe embeddings. The proposed architecture can predict whether the user suffers from depression or not as well as predict the sadness, joy, fear, disgust, and anger score. Li et al. [53] exploited text, pictures, and auxiliary information (post time, dictionary, and social information) and used attention mechanisms within and between the modalities at the same time. The authors exploited TextCNN, ResNet-18, and fully connected layers for extracting representation vectors of text, images, and auxiliary information, respectively. A multimodal approach was proposed by Cheng and Chen [54], where the authors exploited texts, images, posting time, and the time interval between the posts on Instagram. Shen et al. [55] collected multimodal datasets and extracted six depression-oriented feature groups, namely, social network, user profile, visual, emotional, topic-level, and domain-specific features. Gui et al. [56] combined texts and images and proposed a new cooperative multiagent reinforcement learning method.

Multitask approaches have been introduced. A multitask approach was introduced by Zhou et al. [57]. Specifically, the authors proposed a hierarchical attention network consisting of BiGRU layers and integrated LDA topics. The main task was the identification of depression, i.e., binary classification task, while the auxiliary task was the prediction of the domain category of the post, i.e., multiclass classification task. Both multitask and multimodal approaches were introduced by Wang et al. [58]. The authors extracted a total of ten features from the text, social behavior, and pictures. XLNet and BiGRU coupled with attention layers, and dense layers were used.

C. Related Work Review Findings

Existing research initiatives rely on the feature extraction process and the train of shallow machine classifiers targeting diagnosing mental disorders in social media. This fact demands domain expertise and does not generalize well to new data. Other existing approaches train CNNs, BiL-STMs, or employ hybrid models and ensemble strategies. Recently, transformer-based models have been used also. Only a few works have experimented with injecting linguistic, including emotion, features into deep neural networks. These approaches employ multitask learning models, fine-tuning, or multimodal approaches. All these approaches employing transformer-based models usually fine-tune these models. None of these approaches have used modifications of BERT aiming to enhance its performance by injecting into it external knowledge. Also, no prior work has taken into account model calibration creating in this way overconfident models.

Therefore, our work differs from the existing research initiatives, since we: 1) present a new method, which injects linguistic features into transformer-based models; 2) apply label smoothing for calibrating our model and evaluate both the performance and calibration of our proposed models; and 3) present a methodology for gaining linguistic insight into

the tasks of depression and stress investigating their common characteristics.

III. METHODOLOGY

A. Architecture

In this section, we describe our proposed approach for detecting stressful and depressive posts on social media. Our proposed method is based on the work introduced by Rahman et al. [23] and Jin and Aletras [24]. Instead of cross-modal interactions, we inject extra-linguistic information as alternative views of the data into pretrained language models. Our proposed architecture is illustrated in Fig. 1.

Specifically, we use the following feature vectors.

1) *NRC:* The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [59]. Each text is represented as a 10-d vector, where each element is the proportion of tokens belonging to each category.

2) *LIWC:* LIWC is a dictionary-based approach to count words in linguistic, psychological, and topical categories [60]. We use LIWC 2022 [61] to represent each text as a 117-d vector.

3) *LDA Topics:* Before training the LDA model, we remove stop words and punctuation. We exploit LDA (with 25 topics) and extract 25 topic probabilities per text [62]. These probabilities describe the topics of interest in each text. Inspired by Liu et al. [22], we use the following feature vector.

   a) *Global Outlier Standard Score:* For evaluating the ith text's interest on a certain topic k, compared to the rest of the texts, we use the GOSS feature

$$\mu(x_k) = \frac{\sum_{i=1}^{n} x_{ik}}{n} \qquad (1)$$

$$\text{GOSS}(x_{ik}) = \frac{x_{ik} - \mu(x_k)}{\sqrt{\sum_i (x_{ik} - \mu(x_k))^2}}. \qquad (2)$$

   Therefore, each text is represented as a 25-d vector.

4) *Top2Vec:* Top2Vec [63] is an algorithm for topic modeling, which automatically detects topics present in the text and generates jointly embedded topic, document, and word vectors. After training Top2Vec by exploiting the universal sentence encoder, each text is represented as a 512-d vector.

We experiment with the following pretrained models: BERT [64] and MentalBERT [65].

First, we pass each text through the aforementioned transformer-based models. Let $C \in \mathcal{R}^{N \times d}$ be the output of the transformer-based models, where N denotes the sequence length, while d denotes the dimensionality of the models. We have omitted the dimension corresponding to the batch size for the sake of simplicity.

Then, we project the feature vectors to dimensionality equal to 128. We repeat the feature vector N times, so as to ensure that the feature vector and the output of the transformer-based models can be concatenated. Given the word representation

$e^{(i)}$, we concatenate $e^{(i)}$ with feature vectors, i.e., $h_v^{(i)}$

$$w_v^{(i)} = \sigma\left(W_{\text{hv}}[e^{(i)}; h_v^{(i)}] + b_v\right) \qquad (3)$$

where σ denotes the sigmoid activation function, W_{hv} is a weight matrix, and $w_v^{(i)}$ corresponds to the gate. b_v is the scalar bias.

Next, we calculate a shift vector $h_m^{(i)}$ by multiplying the embeddings with the gate

$$h_m^{(i)} = w_v^{(i)} \cdot \left(W_v h_v^{(i)}\right) + b_m^{(i)} \qquad (4)$$

where W_v is a weight matrix, and $b_m^{(i)}$ is the bias vector.

Next, we apply the multimodal shifting component aiming to dynamically shift the word representations by integrating the shift vector $h_m^{(i)}$ into the original word embedding

$$e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)} \qquad (5)$$

$$\alpha = \min\left(\frac{\|e^{(i)}\|_2}{\|h_m^{(i)}\|_2} \beta, 1\right) \qquad (6)$$

where β is a hyperparameter. Then, we apply a layer normalization [66] and dropout layer [67] to $e_m^{(i)}$. Next, the combined embeddings are fed to a BERT/MentalBERT model.

We get the classification [CLS] token of this model and pass it through a dense layer consisting of 128 units with a ReLU activation function. Finally, we use a dense layer consisting of either two units (binary classification task) or four units (multiclass classification task).

We denote our proposed models as M-BERT and multi-modal MentalBERT (M-MentalBERT) followed by the linguistic features which are integrated into them. For example, the injection of LIWC features into a BERT model is denoted as M-BERT (LIWC).

B. Model Calibration

To prevent the model from becoming too overconfident, we use label smoothing [25], [68]. Specifically, label smoothing calibrates learned models so that the confidences of their predictions are more aligned with the accuracies of their predictions.

For a network trained with hard targets, the cross-entropy loss is minimized between the true targets y_k and the network's outputs p_k, as in $H(y, p) = \sum_{k=1}^{K} -y_k \log(p_k)$, where y_k is "1" for the correct class and "0" for the other. For a network trained with label smoothing, we minimize instead the cross-entropy between the modified targets $y_k^{\text{LS}_u}$ and the network's outputs p_k

$$y_k^{\text{LS}_u} = y_k \cdot (1 - \alpha) + \frac{\alpha}{K} \qquad (7)$$

$$H(y, p) = \sum_{k=1}^{K} -y_k^{\text{LS}_u} \cdot \log(p_k) \qquad (8)$$

where α is the smoothing parameter, and K is the number of classes.

IV. EXPERIMENTS

A. Datasets

1) Dreaddit: This dataset includes stressful and nonstressful texts posted by users of Reddit [27]. Specifically, these posts

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6   IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS
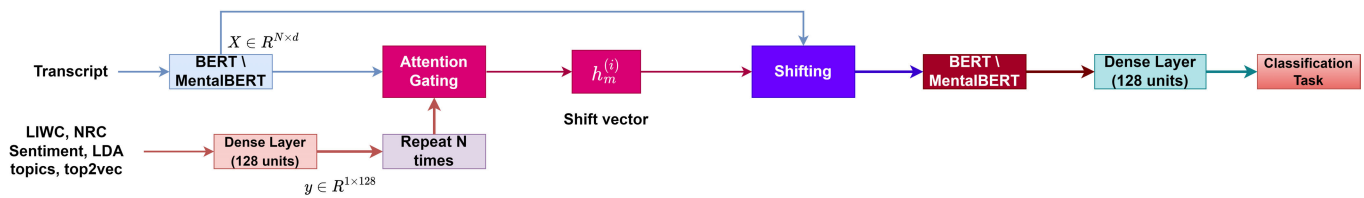

Fig. 1.   Our proposed architecture.

belong to five domains, namely, abuse, anxiety, financial, social, and posttraumatic stress disorder (PTSD). A subset of data has been annotated using Amazon Mechanical Turk. This dataset includes also lexical, syntactic, and social media features per post. The dataset has been divided by the authors into a train and a test set. The train set comprises 1488 stressful texts and 1350 nonstressful ones, while the test set includes 369 stressful texts and 346 nonstressful ones.

2) Depression_Mixed: This dataset [35] consists of 1482 nondepressive posts and 1340 depressive posts. These posts have been written by users on Reddit and English depression forums.

3) Depression_Severity: This dataset includes posts in Reddit [38] and assigns each post to a severity level, i.e., minimal (2587 posts), mild (290 posts), moderate (394 posts), and severe form of depression (282 posts).

B. Experimental Setup

We use the Adam optimizer with a learning rate of 0.001. We apply *StepLR* with a step size of 5 and a gamma of 0.1. We use a batch size of 8. With regards to *Depression_Mixed* dataset, we split the dataset into a train and a test set (80%–20%) similar to Ansari et al. [8]. In terms of *Dreaddit* dataset, we use the test set provided by Turcan and McKeown [27]. Regarding *Depression_Severity* dataset, we use fivefold stratified cross-validation, since the study [38] has also exploited cross-validation. All train sets are divided into a train and a validation set. Regarding *Depression_Severity* dataset, we apply *EarlyStopping* with patience of seven epochs based on the validation loss. In terms of *Depression_Mixed* and *Dreaddit* dataset, we train our introduced model for a maximum of 30 epochs, choose the epoch with the smallest validation loss, and test the model on the test set. We set β of (6) equal to 0.0001.[1] We choose α of (7) equal to 0.001. We use the Python library, namely, transformers [69], for BERT and MentalBERT. Specifically, we use the BERT base uncased version and the MentalBERT base uncased version. We use PyTorch [70] for performing our experiments. All experiments are trained on a single Tesla P100-PCIE-16GB GPU.

C. Evaluation Metrics

1) Performance: In terms of the binary classification tasks, i.e., 0 for nonstressful and 1 for stressful texts or 0 for nondepressive and 1 for depressive texts, we use precision, recall, F1-score, and accuracy to evaluate the performance

of our proposed approach. We use these metrics similar to Wani et al. [36].

Regarding the multiclass classification task reported on the Depression_Severity dataset, we use weighted precision, weighted recall, and weighted F1-score. We use these metrics similar to Mishra et al. [71].

2) Calibration: We evaluate the calibration of our model using the metrics proposed by the relevant literature [72], [73], [74]. Specifically, we use the metrics mentioned as follows.

1) *Expected Calibration Error (ECE):* The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence). First, we divide the predictions into M equally spaced bins (size $1/M$)

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{9}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{10}$$

where y_i and $\hat{y}_i$ are the true and predicted labels for the sample i, and $\hat{p}_i$ is the confidence (predicted probability value) for sample i

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{11}$$

where N is the total number of data points, and B_m is the group of samples whose predicted probability values fall into the interval $I_m = [((m-1)/M), (m/M)]$. Perfectly calibrated models have an ECE of 0.

2) *Adaptive Calibration Error (ACE):* It uses an adaptive scheme that spaces the bin intervals so that each contains an equal number of predictions

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |\text{acc}(r, k) - \text{conf}(r, k)| \tag{12}$$

where $\text{acc}(r, k)$ and $\text{conf}(r, k)$ are the accuracy and confidence of adaptive calibration range r for class label k, respectively, and N is the total number of data points. Calibration ranges r are defined by the $[N/R]$th index of the sorted and thresholded predictions.

D. Baselines

We use the following baselines as comparisons with our proposed approaches.

1) *BERT, MentalBERT:* We fine-tune these pretrained language models in order to explore whether our method of injecting linguistic information into pretrained models

---

[1]We experimented with values of β, including 0.01 and 0.001, but setting β equal to 0.0001 yielded the best results.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ILIAS et al.: CALIBRATION OF TRANSFORMER-BASED MODELS FOR IDENTIFYING STRESS AND DEPRESSION 7

TABLE I

PERFORMANCE COMPARISON AMONG PROPOSED MODELS AND BASELINES USING THE DEPRESSION_MIXED AND DREADDIT DATASETS

| Model | Depression_Mixed | | | | | | | Dreaddit | | | | | |
	Prec.	Rec.	F1-score	Acc.	ECE	ACE		Prec.	Rec.	F1-score	Acc.	ECE	ACE
Baselines													
BERT	91.40	91.40	91.40	-	-	-		75.18	86.99	80.65	-	-	-
MentalBERT	89.27	93.14	91.17	91.15	-	-		-	80.28	80.04	-	-	-
Baselines - Proposed Approaches (without label smoothing)													
M-BERT (NRC)	90.56	91.84	91.20	91.15	0.072	0.081		76.70	85.64	80.92	79.16	0.117	0.125
M-BERT (LIWC)	90.98	92.02	91.49	92.04	0.054	0.055		79.74	84.28	81.95	80.84	0.118	0.110
M-BERT (LDA topics)	88.07	95.80	91.77	92.04	0.071	0.071		76.87	86.68	81.48	79.64	0.103	0.108
M-BERT (top2vec)	90.97	92.99	91.97	92.21	0.057	0.069		78.25	84.82	81.40	80.00	0.075	0.133
M-MentalBERT (NRC)	90.65	92.65	91.64	91.86	0.031	0.054		80.31	82.93	81.60	80.69	0.099	0.101
M-MentalBERT (LIWC)	93.49	87.78	90.55	91.50	0.057	0.056		79.85	85.91	82.77	81.53	0.158	0.153
M-MentalBERT (LDA topics)	87.97	93.09	90.46	90.44	0.089	0.086		78.16	85.60	81.71	80.19	0.109	0.134
M-MentalBERT (top2vec)	91.63	93.77	92.69	93.27	0.058	0.054		79.70	85.09	82.31	81.12	0.123	0.149
Proposed Approaches (with label smoothing)													
M-BERT (NRC)	89.82	94.81	92.25	92.39	0.059	0.065		76.51	89.16	82.35	80.28	0.094	0.105
M-BERT (LIWC)	93.06	91.78	92.41	92.21	0.034	0.044		77.21	89.97	83.10	81.12	0.114	0.125
M-BERT (LDA topics)	90.16	92.71	91.42	92.39	0.063	0.067		80.10	84.24	82.12	81.04	0.063	0.080
M-BERT (top2vec)	90.34	94.93	92.58	92.57	0.049	0.056		78.41	85.64	81.87	80.42	0.066	0.083
M-MentalBERT (NRC)	91.44	92.52	91.98	92.74	0.042	0.057		79.70	86.18	82.81	81.53	0.082	0.101
M-MentalBERT (LIWC)	94.96	89.42	92.11	92.57	0.055	0.057		79.41	87.80	83.40	81.96	0.116	0.112
M-MentalBERT (LDA topics)	94.81	90.78	92.75	92.92	0.047	0.049		79.11	87.50	83.10	81.60	0.096	0.111
M-MentalBERT (top2vec)	96.12	90.18	93.06	93.45	0.033	0.043		78.92	87.26	82.88	81.39	0.121	0.137

TABLE II

PERFORMANCE COMPARISON AMONG PROPOSED MODELS AND BASELINES USING THE DEPRESSION_ SEVERITY DATASET

Model	W. Prec.	W. Rec.	W. F1-score	ECE	ACE
Baselines					
BERT	72.99	71.97	71.00	-	-
MentalBERT	73.35	70.81	71.67	-	-
Baselines - Proposed Approaches (without label smoothing)					
M-BERT (NRC)	74.48	70.08	69.96	0.107	0.076
M-BERT (LIWC)	73.77	71.74	72.13	0.110	0.078
M-BERT (LDA topics)	74.25	71.80	71.28	0.114	0.079
M-BERT (top2vec)	72.93	71.97	71.00	0.086	0.071
M-MentalBERT (NRC)	74.43	72.58	69.96	0.097	0.069
M-MentalBERT (LIWC)	72.39	72.53	71.95	0.112	0.075
M-MentalBERT (LDA topics)	73.83	72.58	72.58	0.118	0.078
M-MentalBERT (top2vec)	74.63	72.39	72.06	0.103	0.075
Proposed Approaches (with label smoothing)					
M-BERT (NRC)	74.04	72.84	72.81	0.102	0.074
M-BERT (LIWC)	73.68	72.16	72.37	0.094	0.069
M-BERT (LDA topics)	73.24	71.46	71.42	0.112	0.078
M-BERT (top2vec)	73.36	72.64	72.30	0.113	0.074
M-MentalBERT (NRC)	73.03	71.23	71.46	0.112	0.079
M-MentalBERT (LIWC)	73.21	73.15	72.43	0.099	0.071
M-MentalBERT (LDA topics)	73.74	73.23	73.16	0.111	0.075
M-MentalBERT (top2vec)	73.68	72.70	72.67	0.094	0.071

leads to performance improvement. In terms of the Depression_Mixed dataset, we report the performance of BERT obtained by Yang et al. [28]. We fine-tune MentalBERT and report its performance on this dataset. With regards to the Dreaddit dataset, we report the performance of BERT and MentalBERT obtained by Turcan and McKeown [27] and Ji et al. [65], respectively. Regarding the Depression_Severity dataset, we fine-tune BERT and MentalBERT and report their performances. We do not report calibration metrics for these models, since our goal, in this case, is to compare only the performances of these models with our proposed approaches.

2) *Proposed Approaches (Without Label Smoothing):* We train the proposed models introduced in Section III without label smoothing. We explore whether label smoothing leads to performance improvement and better calibration of our models.

V. RESULTS

The results of our proposed approach are reported in Tables I and II. Specifically, Table I reports the performances of our proposed approaches on the Depression_Mixed and Dreaddit datasets, while Table II reports the results on the Depression_Severity dataset.

In terms of the Dreaddit dataset, we first compare our proposed approaches without label smoothing with the BERT and MentalBERT models. First, we observe that the integration of linguistic features into the BERT model improves the performance obtained by BERT. Specifically, M-BERT (LIWC) yields the highest F1-score accounting for 81.95% surpassing BERT by 1.30%. At the same time, M-BERT (LIWC) outperforms M-BERT (NRC), M-BERT (LDA topics), and M-BERT (top2vec) in precision, F1-score, and accuracy. Similarly, the injection of LIWC features into the MentalBERT model yields an F1-score of 82.77% outperforming MentalBERT by 2.73% and the other approaches by 0.46%–1.17%. M-MentalBERT (NRC) yields the lowest F1-score accounting for 81.60%. With regards to the proposed approaches with label smoothing, we observe that they outperform the proposed approaches without label smoothing in terms of both performance and calibration metrics. Specifically, we observe that M-BERT (LIWC) with label smoothing attains the highest F1-score and accuracy accounting for 83.10% and 81.12%, respectively, outperforming M-BERT (LIWC) without label smoothing in F1-score by 1.15% and in accuracy by 0.28%. The integration of LIWC features into the MentalBERT model with label smoothing obtains better performance than MentalBERT in F1-score by 3.36% and better performance than M-MentalBERT (LIWC) without label smoothing in F1-score by 0.63%. In addition, we observe that label smoothing leads to improvement in calibration metrics. For instance, M-BERT (NRC) with label smoothing improves both ECE and ACE by 0.023 and 0.020, respectively, in comparison with M-BERT (NRC) without label smoothing.

Regarding the Depression_Mixed dataset, we first compare our proposed approaches without label smoothing with the BERT and MentalBERT models. We observe that the injection of linguistic features, except for NRC features, into the BERT model improves the F1-score. Specifically, we observe that the injection of top2vec features yields the highest F1-score and accuracy accounting for 91.97% and 92.21%, respectively,

surpassing the performance of the BERT model in F1-score by 0.57%. We speculate that the injection of top2vec features obtains better performance than the injection of features derived by LDA topics, i.e., GOSS features since the top2vec algorithm is capable of identifying the number of topics automatically. In terms of MentalBERT, we observe that the injection of top2vec features obtains an F1-score of 92.69% surpassing MentalBERT by 1.52%. We observe that the integration of NRC and top2vec features improve the performance obtained by MentalBERT. Regarding the proposed approaches with label smoothing, we observe that these models attain better performances than the ones obtained by the models without label smoothing. Specifically, we observe that M-BERT (top2vec) with label smoothing surpasses the respective model without label smoothing in F1-score and Accuracy by 0.61% and 0.36% respectively. Similarly, M-MentalBERT (top2vec) with label smoothing obtains the highest F1-score and accuracy accounting for 93.06% and 93.45%, respectively. This model surpasses the respective model without label smoothing in F1-score and accuracy by 0.37% and 0.18%. Except for the improvement of the performance metrics, i.e., precision, recall, F1-score, and accuracy, we observe that the models with label smoothing obtain better results in terms of the calibration metrics, i.e., ECE and ACE, than the ones obtained by the models without label smoothing. For example, we observe that M-BERT (top2vec) with label smoothing improves the ECE and ACE scores obtained by M-BERT (top2vec) without label smoothing by 0.008 and 0.013, respectively. Similarly, M-MentalBERT (LDA topics) with label smoothing improves the ECE and ACE scores obtained by M-MentalBERT (LDA topics) without label smoothing by 0.042 and 0.043, respectively.

With regards to the Depression_Severity dataset, we first compare our proposed approaches without label smoothing with the BERT and MentalBERT models. We observe that the integration of LIWC features and features extracted by LDA topic modeling, i.e., GOSS features, into the BERT model leads to a performance surge in comparison with the BERT model. Specifically, M-BERT (LIWC) outperforms BERT in weighted F1-score by 1.13%. At the same time, the integration of all the features, except NRC, to a MentalBERT model yields a performance improvement compared to the MentalBERT model. Specifically, M-MentalBERT (LDA topics) attains the highest weighted F1-score accounting for 72.58% surpassing MentalBERT by 0.91%. When it comes to proposed models with label smoothing, we observe an improvement in both the performance metrics and calibration ones. More specifically, the integration of NRC features to a BERT model obtains a weighted F1-score of 72.81% outperforming BERT by 1.81%, M-BERT (NRC) without label smoothing by 2.85%, and M-BERT (LIWC) without label smoothing by 0.68%. In addition, M-MentalBERT (LDA topics) with label smoothing obtains the highest F1-score accounting for 73.16% surpassing MentalBERT by 1.49% and M-MentalBERT (LDA topics) without label smoothing by 0.58%. In terms of the calibration metrics, we observe that both ECE and ACE scores are improved when we apply label smoothing. For example, M-BERT (LIWC) with label smoothing obtains an ECE score of 0.094 and an ACE score of 0.069, which are improved by

0.016 and 0.009, respectively, compared with the respective model without label smoothing.

VI. Linguistic Analysis

We finally perform an analysis on the Depression_Mixed and Dreaddit datasets to uncover the peculiarities of stress and depression. Specifically, we seek to find the correlations of LIWC features with stressful/depressive and nonstressful/nondepressive texts. To do this, we adopt the methodology by Ilias and Askounis [75]. First, we normalize LIWC features, so as to ensure that they sum up to 1 across each post. Next, we use the point-biserial correlation between each LIWC category and the label of the post. The output of the point-biserial correlation is a number ranging from -1 to 1. Positive correlations mean that the specific LIWC category is correlated with the stressful/depressive class (label 1), while negative correlations mean that the specific LIWC category is correlated with the nonstressful/nondepressive class (label 0). We consider the absolute values of the correlations. Results are reported in Table III. All the correlations are significant at $p < 0.05$ with Benjamini–Hochberg correction [76] for multiple comparisons.

In terms of the *Depression_Mixed* dataset, we observe that the control group tends to use words with positive tone and emotion, i.e., good, well, happy, hope, and so on. In addition, the healthy control group discusses topics of everyday life, including lifestyle (work, home, and school), culture (car and phone), politics (govern and congress), family, and friends (boyfriend, girlfriend, and dude). Also, these people make plans for the future, thus using words indicating a focus on the future (correlation equal to 0.0666). However, it must be noted that this is a very weak correlation. On the other hand, people with depression focus on the present and do not make plans for the future. They discuss negative topics, including death, illnesses, mental health, and substances. This can be justified by the fact that people with depression often have tendencies to suicide and believe that they cannot achieve anything. In addition, they use swear words, i.e., shit, fuck, and damn, since they think that everything goes wrong in their life. Also, their posts are full of sadness, anxiety, and negative tone.

Regarding the *Dreaddit* dataset, we observe similar patterns. Specifically, nonstressful posts include words with positive tone and emotion. People in nonstressful conditions discuss topics pertinent to lifestyle, money, leisure, technology, food, work, religion, and many more. Also, people use personal pronouns, including third-person singular, second-person, and first-person plural. Also, their posts include words indicating politeness. On the contrary, people in stressful conditions use swear words or words indicating interpersonal conflict, including fight, kill, attacking, and so on. The first person singular constitutes the LIWC category with the highest degree of correlation with the stressful class. This result agrees with previous work [77], where it is mentioned that self-references by individuals present a surge in emotionally vulnerable conditions. Therefore, the use of the first-person singular indicates increased self-focus. In addition, similar to the depressive posts, we observe that stressful posts include words with

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ILIAS et al.: CALIBRATION OF TRANSFORMER-BASED MODELS FOR IDENTIFYING STRESS AND DEPRESSION 9

TABLE III
LIWC FEATURES ASSOCIATED WITH DEPRESSIVE/STRESSFUL AND NONDEPRESSIVE/NONSTRESSFUL POSTS, SORTED BY POINT-BISERIAL CORRELATION. ALL CORRELATIONS ARE SIGNIFICANT AT $p < 0.05$ AFTER BENJAMINI–HOCHBERG CORRECTION

Depression_Mixed				Dreaddit			
Non-Depressive		Depressive		Non-Stressful		Stressful	
LIWC	corr.	LIWC	corr.	LIWC	corr.	LIWC	corr.
Tone	0.3156	health	0.4108	Tone	0.4232	1st person singular	0.3975
Clout	0.3022	mental health	0.3674	clout	0.3501	tone_neg	0.3807
Social referents	0.2914	physical	0.3603	tone_pos	0.2574	emo_neg	0.3434
shehe	0.2634	emo_sad	0.3506	analytic	0.2269	dic	0.3122
we	0.2415	tone_neg	0.3274	social	0.2084	authentic	0.2964
social	0.2401	1st person singular	0.2974	you	0.2082	linguistic	0.2955
male references	0.2199	Authentic	0.2961	Social referents	0.1890	function	0.2928
affiliation	0.1960	cognition	0.2957	emo_pos	0.1630	emo_anx	0.2409
female references	0.1923	emo_neg	0.2843	affiliation	0.1495	emotion	0.2354
conversation	0.1794	cognitive processes	0.2601	article	0.1338	pronoun	0.2272
netspeak	0.1741	feeling	0.2507	we	0.1334	feeling	0.2162
culture	0.1732	focuspresent	0.2139	prosocial	0.1290	personal pronouns	0.2150
allpunc	0.1667	insight	0.2138	number	0.1284	focuspresent	0.2110
family	0.1589	emotion	0.2120	OtherP	0.1231	adverb	0.1925
technology	0.1524	negations	0.2116	Social behavior	0.1229	negations	0.1900
exclam	0.1519	verb	0.2076	drives	0.1130	verb	0.1862
analytic	0.1507	linguistic	0.1893	polite	0.1121	apostro	0.1854
period	0.1458	death	0.1842	comma	0.1079	affect	0.1811
drives	0.1168	function	0.1834	curiosity	0.0990	conjunctions	0.1630
OtherP	0.1156	all-or-none	0.1748	reward	0.0971	auxverb	0.1626
number	0.1075	dic	0.1708	female	0.0909	WC	0.1502
assent	0.1013	affect	0.1609	determiners	0.0896	emo_sad	0.1482
tone_pos	0.0998	adverb	0.1588	visual	0.0867	swear	0.1424
leisure	0.0980	illness	0.1584	exclam	0.0839	cognition	0.1325
Social behavior	0.0937	emo_anx	0.1458	culture	0.0782	emo_anger	0.1260
communication	0.0920	auxverb	0.1442	lifestyle	0.0777	causation	0.1188
lifestyle	0.0917	discrepancy	0.1373	money	0.0750	allure	0.1188
friend	0.0812	apostro	0.1256	attention	0.0709	risk	0.1099
curiosity	0.0792	want	0.1146	allpunc	0.0672	cognitive processes	0.1081
you	0.0774	achieve	0.1136	leisure	0.0635	health	0.1013
determiners	0.0758	pronoun	0.1092	they	0.0590	all-or-none	0.1008
politic	0.0751	lack	0.1054	technology	0.0574	insight	0.0988
relig	0.0673	differ	0.1004	food	0.0561	conflict	0.0984
focusfuture	0.0666	prepositions	0.0951	politic	0.0547	physical	0.0947
visual	0.0660	risk	0.0928	communication	0.0532	illness	0.0935
motion	0.0647	allure	0.0923	quantity	0.0528	words per sentence (WPS)	0.0904
money	0.0626	causation	0.0916	BigWords	0.0474	need	0.0897
ethnicity	0.0591	tentative	0.0880	wellness	0.0471	mental health	0.0893
article	0.0493	time	0.0811	work	0.0465	death	0.0849
emo_pos	0.0480	personal pronouns	0.0801	religion	0.0432	impersonal pronoun	0.0813
home	0.0420	impersonal pronoun	0.0789	shehe	0.0389	fatigue	0.0768
food	0.0401	perception	0.0747	ethnicity	0.0387	memory	0.0746
words per sentence (WPS)	0.0391	swear	0.0697	-	-	time	0.0726
Nonfluencies	0.0385	substances	0.0686	-	-	certitude	0.0557
-	-	memory	0.0673	-	-	preception	0.0397
-	-	BigWords	0.0588	-	-	motion	0.0385
-	-	adj	0.0575	-	-	differ	0.0374
-	-	certitude	0.0547	-	-	-	-
-	-	wellness	0.0457	-	-	-	-
-	-	moral	0.0433	-	-	-	-
-	-	conflict	0.0411	-	-	-	-
-	-	acquire	0.0394	-	-	-	-
-	-	QMark	0.0384	-	-	-	-

negative emotions, anger, sadness, and anxiety. Also, topics discussed by these users are pertinent to illness, health, and death.

VII. DISCUSSION

Our study contributes to the literature by introducing the first approach of integrating extra-linguistic information into pretrained language models based on transformers, namely, BERT and MentalBERT. Specifically, we adapt M-BERT [23] by replacing multimodal information with linguistic information. To be more precise, we extract NRC, LIWC, features derived by LDA topics, and top2vec features. We apply a multimodal adaptation gate and exploit also a shifting component for creating new combined embeddings which are given as input to BERT (and MentalBERT) models. In addition, motivated by the fact that in real-world decision-making systems, classification networks must not only be accurate but also should indicate when they are likely to be incorrect, we apply label smoothing and evaluate our proposed approaches both in terms of classification and calibration.

Therefore, our study is different from the state-of-the-art approaches described in Section II.

1) Prior works having proposed multimodal, multitask, and ensemble strategies in conjunction with transformer-based models, have just fine-tuned these pretrained transformer-based models instead of using some modifications of them. Thus, this study is the first attempt to inject extra knowledge into BERT (and MentalBERT), in order to enhance its performance.
2) All the prior works evaluate only the classification performance of their approaches neglecting the confidence of the prediction. To tackle this, this is the first study in the task of stress and depression detection through social media posts utilizing label smoothing and evaluating both the classification performance and the calibration of the models.
3) Finally, this is the first study utilizing features derived from LDA topics, namely, the GOSS, which captures the text's interest compared to other texts.

From the results of this study, we found the following.

1) *Finding 1:* The integration of linguistic features into transformer-based models yields an increase in classification performance. However, it is worth noting that in some cases this improvement is limited. For instance, the integration of LIWC features into the MentalBERT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10   IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS

model with label smoothing obtains better performance than MentalBERT in F1-score by 3.36% and better performance than M-MentalBERT (LIWC) without label smoothing in F1-score by 0.63%. However, we believe that even a small improvement can make a difference.

2) *Finding 2:* Label smoothing improves both the performance and the calibration of the proposed approaches. The calibration of the proposed approaches is measured via two metrics, namely, ECE and ACE.

3) *Finding 3:* Findings from a linguistic analysis revealed that people in stressful and/or depressive conditions use words belonging to specific LIWC categories more frequently than others.

There are several limitations related to this study.

1) *Hyperparameter Tuning:* Due to limited access to GPU resources, we were not able to perform hyperparameter tuning. On the contrary, we tried some combinations of parameters. We believe that the adoption of the hyperparameter tuning procedure through access to GPU resources would increase further the classification performance.

2) *Explainability:* The present study is not accompanied by explainability techniques, i.e., integrated gradients [78], and so on. Therefore, we aim to apply explainability techniques in the future.

3) Due to limited access to GPU resources and similar to prior work [8], [36], [39], we were not able to perform multiple runs for testing for statistical significance in terms of the *Depression_Mixed* and *Dreaddit* datasets.

VIII. CONCLUSION

In this article, we present a new method for identifying stress and depression in social media text by injecting linguistic information into transformer-based models. Also, it is the first study exploiting label smoothing, in order to ensure that our model is calibrated. We evaluate our proposed methods on three publicly available datasets, which include a depression detection dataset (binary classification), a stress detection dataset, and a depression detection dataset (multiclass classification—severity of depression). Findings suggest that transformer-based networks combined with linguistic information lead to performance improvement in comparison with transformer-based networks. Also, applying label smoothing yields both the performance improvement and better calibration of the proposed models. Specifically, in terms of the Depression_Mixed dataset, we found that the injection of top2vec features into BERT and MentalBERT models along with label smoothing obtained the highest F1-score and accuracy. Regarding the Dreaddit dataset, results showed that the integration of LIWC features into language models based on transformers in conjunction with label smoothing yielded the highest F1-score and accuracy. With regards to the Depression_Severity dataset, findings showed that the injection of NRC features into the BERT model and the integration of features derived by LDA topics, namely, GOSS features, into the MentalBERT model yielded the highest weighted F1-scores. We also conduct a linguistic analysis and show

that stressful and depressive posts present high correlations with common LIWC categories.

In the future, we plan to exploit transfer learning and domain adaptation methods. Also, employing explainable multimodal models is one of our future plans. In addition, we plan to exploit more methods for enhancing transformer-based models with external knowledge. Finally, we aim to contribute further to the uncertainty estimation by exploiting Monte Carlo dropout [79].

REFERENCES

[1] World Health Organization. (2023). *Stress*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/stress

[2] W. J. Friedman. (2023). *Types of Stress and Their Symptoms*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.mentalhelp.net/blogs/types-of-stress-and-their-symptoms/

[3] World Health Organization. (2021). *Depression*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[4] World Health Organization. (2017). *Depression and Other Common Mental Disorders*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.who.int/publications/i/item/depression-global-health-estimates

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[6] S. Chandra Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, and L. H. Ungar, "Understanding and measuring psychological stress using social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, no. 1, pp. 214–225, Jul. 2019.

[7] J. S. L. Figuerêdo, A. L. L. M. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," *Online Social Netw. Media*, vol. 31, Sep. 2022, Art. no. 100225.

[8] L. Ansari, S. Ji, Q. Chen, and E. Cambria, "Ensemble hybrid learning methods for automated depression detection," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 1, pp. 211–219, Feb. 2023.

[9] N. Poerner, U. Waltinger, and H. Schütze, "E-BERT: Efficient-Yet-Effective entity embeddings for BERT," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2020, pp. 803–818.

[10] N. Kassner and H. Schütze, "Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7811–7818.

[11] N. Peinelt, M. Rei, and M. Liakata, "GiBERT: Enhancing BERT with linguistic information using a lightweight gated injection method," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2021, pp. 2322–2336.

[12] R. Wang et al., "K-adapter: Infusing knowledge into pre-trained models with adapters," in *Proc. Findings Assoc. Comput. Linguistics, (ACL-IJCNLP)*, 2021, pp. 1405–1418.

[13] J. Lu, D. Batra, D. Parikh, and S. Lee, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Red Hook, NY, USA: Curran Associates, 2019.

[14] M. E. Peters et al., "Knowledge enhanced contextual word representations," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 43–54.

[15] A. P. Dawid, "The well-calibrated Bayesian," *J. Amer. Stat. Assoc.*, vol. 77, no. 379, pp. 605–610, Sep. 1982.

[16] A. H. Murphy and E. S. Epstein, "Verification of probabilistic predictions: A brief review," *J. Appl. Meteorol.*, vol. 6, no. 5, pp. 748–755, Oct. 1967.

[17] C. S. Crowson, E. J. Atkinson, and T. M. Therneau, "Assessing calibration of prognostic risk scores," *Stat. Methods Med. Res.*, vol. 25, no. 4, pp. 1692–1706, Aug. 2016.

[18] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado, "Calibrating predictive model estimates to support personalized medicine," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 2, pp. 263–274, Mar. 2012.

[19] M. Raghu et al., "Direct uncertainty prediction for medical second opinions," in *Proc. Mach. Learn. Res.*, vol. 97, Long Beach, CA, USA, Jun. 2019, pp. 5281–5290.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ILIAS et al.: CALIBRATION OF TRANSFORMER-BASED MODELS FOR IDENTIFYING STRESS AND DEPRESSION 11

[20] R. Sawhney, A. Neerkaje, and M. Gaur, "A risk-averse mechanism for suicidality assessment on social media," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2022, pp. 628–635.

[21] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[22] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, and J. Lu, "Detecting 'smart' spammers on social network: A topic model approach," in *Proc. NAACL Student Res. Workshop.* San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 45–50.

[23] W. Rahman et al., "Integrating multimodal information in large pre-trained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[24] M. Jin and N. Aletras, "Complaint identification in social media with transformer networks," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1765–1771.

[25] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–13.

[26] S. Muñoz and C. A. Iglesias, "A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations," *Inf. Process. Manage.*, vol. 59, no. 5, Sep. 2022, Art. no. 103011.

[27] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*, 2019, pp. 97–107.

[28] K. Yang, T. Zhang, and S. Ananiadou, "A mental state knowledge—Aware and contrastive network for early stress and depression detection on social media," *Inf. Process. Manage.*, vol. 59, no. 4, 2022, Art. no. 102961.

[29] G. I. Winata, O. P. Kampman, and P. Fung, "Attention-based LSTM for psychological stress detection from spoken language using distant supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6204–6208.

[30] H. Lin et al., "Detecting stress based on social interactions in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1820–1833, Sep. 2017.

[31] E. Turcan, S. Muresan, and K. McKeown, "Emotion-infused models for explainable psychological stress detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 2895–2909.

[32] J. Liu and M. Shi, "A hybrid feature selection and ensemble approach to identify depressed users in online social media," *Frontiers Psychol.*, vol. 12, p. 6571, Jan. 2022.

[33] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 217–226, Jul. 2014.

[34] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from Twitter activity," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 3187–3196.

[35] I. Pirina and Ç. Çöltekin, "Identifying depression on Reddit: The effect of training data," in *Proc. EMNLP Workshop SMM4H, 3rd Social Media Mining Health Appl. Workshop Shared Task*, 2018, pp. 9–12.

[36] M. A. Wani, M. A. ELAffendi, K. A. Shakil, A. S. Imran, and A. A. A. El-Latif, "Depression screening in humans with AI and deep learning techniques," *IEEE Trans. Computat. Social Syst.*, early access, Sep. 28, 2022, doi: 10.1109/TCSS.2022.3200213.

[37] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Sci. Rep.*, vol. 10, no. 1, pp. 1–6, Jul. 2020.

[38] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on Reddit using ordinal classification," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2563–2572.

[39] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 6, pp. 1465–1474, Dec. 2021.

[40] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," *Multimedia Tools Appl.*, vol. 81, no. 17, pp. 23649–23685, Jul. 2022.

[41] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic," *IEEE J. Biomed. Health Informat.*, early access, Feb. 9, 2023, doi: 10.1109/JBHI.2023.3243249.

[42] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun, "Depression detection on Reddit with an emotion-based attention network: Algorithm development and validation," *JMIR Med. Informat.*, vol. 9, no. 7, Jul. 2021, Art. no. e28754.

[43] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 588–601, Mar. 2020.

[44] V. B. de Souza, J. C. Nobre, and K. Becker, "DAC stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3303–3311, Jul. 2022.

[45] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, Jan. 2022.

[46] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Gener. Comput. Syst.*, vol. 124, pp. 480–494, Nov. 2021.

[47] H. Song, J. You, J.-W. Chung, and J. C. Park, "Feature attention network: Interpretable depression detection from social media," in *Proc. 32nd Pacific Asia Conf. Lang., Inf. Comput.* Hong Kong: Association for Computational Linguistics, Dec. 2018, pp. 1–3.

[48] S. Boinepelli, T. Raha, H. Abburi, P. Parikh, N. Chhaya, and V. Varma, "Leveraging mental health forums for user-level depression detection on social media," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 5418–5427.

[49] K. Anantharaman, A. S, R. Sivanaiah, S. Madhavan, and S. M. Rajendram, "SSN_MLRG1@LT-EDI-ACL2022: Multi-class classification using BERT models for detecting depression signs from social media text," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, 2022, pp. 296–300.

[50] F. Nilsson and G. Kovács, "FilipN@LT-EDI-ACL2022-detecting signs of depression from social media: Examining the use of summarization methods as data augmentation for text classification," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, 2022, pp. 283–286.

[51] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: Learning multi-modalities with user post summarization for depression detection on social media," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 133–142.

[52] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "What does your bio say? Inferring Twitter users' depression status from multimodal profile information using deep learning," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 5, pp. 1484–1494, Oct. 2022.

[53] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: A multimodal hierarchical attention model for depression detection in social media," *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 6, Jan. 2023.

[54] J. C. Cheng and A. L. P. Chen, "Multimodal time-aware attention networks for depression detection," *J. Intell. Inf. Syst.*, vol. 59, no. 2, pp. 319–339, Oct. 2022.

[55] G. Shen et al., "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3838–3844.

[56] T. Gui et al., "Cooperative multimodal approach to depression detection in Twitter," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, Jul. 2019, pp. 110–117.

[57] D. Zhou, J. Yuan, and J. Si, "Health issue identification in social media based on multi-task hierarchical neural networks with topic attention," *Artif. Intell. Med.*, vol. 118, Aug. 2021, Art. no. 102119.

[58] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "Online social network individual depression detection using a multitask heterogenous modality fusion approach," *Inf. Sci.*, vol. 609, pp. 727–749, Sep. 2022.

[59] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word—Emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013.

[60] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[61] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, *The Development and Psychometric Properties of LIWC-22*. Austin, TX, USA: Univ. Texas Austin, 2022.

[62] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[63] D. Angelov, "Top2vec: Distributed representations of topics," 2020, *arXiv:2008.09470*. [Online]. Available: https://github.com/ddangelov/Top2Vec

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[65] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 7184–7190.

[66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[67] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[68] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[69] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations* Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[70] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[71] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Author profiling for abuse detection," in *Proc. 27th Int. Conf. Comput. Linguistics.* Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 1088–1098.

[72] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[73] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–4.

[74] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 1321–1330.

[75] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4153–4164, Aug. 2022.

[76] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc., Ser. B, Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

[77] J. W. Pennebaker and T. C. Lay, "Language use and personality during crises: Analyses of mayor Rudolph Giuliani's press conferences," *J. Res. Personality*, vol. 36, no. 3, pp. 271–282, Jun. 2002.

[78] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 3319–3328.

[79] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, NY, USA, vol. 48, Jun. 2016, pp. 1050–1059.

Loukas Ilias received the integrated master's degree from the School of Electrical and Computer Engineering (SECE), National Technical University of Athens (NTUA), Athens, Greece, in June 2020, where he is currently pursuing the Ph.D. degree with the Decision Support Systems (DSS) Laboratory, SECE. He has completed a Research Internship with University College London (UCL), London, U.K.

He is a Researcher with the DSS Laboratory, NTUA, where he is involved in EU-funded research projects. He has published in numerous journals, including IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, *Expert Systems With Applications* (Elsevier), *Applied Soft Computing* (Elsevier), *Computer Speech and Language* (Elsevier), IEEE ACCESS, and *Frontiers in Aging Neuroscience*. His research has also been accepted for presentation at international conferences, including the IEEE-Engineering in Medicine and Biology Society (EMBS) International Conference on Biomedical and Health Informatics (BHI'22) and the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASP) 2023. His research interests include speech processing, natural language processing, social media analysis, and the detection of complex brain disorders.

Spiros Mouzakitis received the Ph.D. degree from the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, in 2009.

He has more than ten years of industry experience in software engineering for banking, e-commerce systems, and energy suppliers. He worked as a Project Manager for 16 years in more than 20 projects (H2020, FP7, FP6, and FP5 research projects) related to big data, open data, market/impact analysis in the context of ICT technologies, web development, and enterprise interoperability. He is a Senior Research Analyst with the Decision Support Systems Laboratory, National Technical University of Athens. He has published in numerous journals and presented his research at international conferences. His current research is focused on decision analysis in the field of decision support systems based on machine learning/deep learning, big and linked data analytics, as well as optimization systems and algorithms.

Dimitris Askounis was the Scientific Director of over 50 European research projects in the above areas (FP7, Horizon2020, and so on). For a number of years, he was an Advisor to the Minister of Justice and the Special Secretary for Digital Convergence for the introduction of information and communication technologies in public administration. Since June 2019, he has been the President of the Information Society SA, Kallithea, Greece. He is currently a Professor at the School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Athens, Greece, and the Deputy Director of the Decision Support Systems Laboratory. He has over 25 years of experience in decision support systems, intelligent information systems and manufacturing, e-business, e-government, open and linked data, big data analytics, Artificial Intelligence (AI) algorithms, and the application of modern Information Technology (IT) techniques in the management of companies and organizations.