Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features

Sourabh Zanwar
RWTH Aachen University
sourabh.zanwar@rwth-aachen.de

Daniel Wiechmann
University of Amsterdam
d.wiechmann@uva.nl

Yu Qiao
RWTH Aachen University
yu.qiao@rwth-aachen.de

Elma Kerz
RWTH Aachen University
elma.kerz@ifaar.rwth-aachen.de

Abstract

In recent years, there has been increased interest in building predictive models that harness natural language processing and machine learning techniques to detect emotions from various text sources, including social media posts, micro-blogs or news articles. Yet, deployment of such models in real-world sentiment and emotion applications faces challenges, in particular poor out-of-domain generalizability. This is likely due to domain-specific differences (e.g., topics, communicative goals, and annotation schemes) that make transfer between different models of emotion recognition difficult. In this work we propose approaches for text-based emotion detection that leverage transformer models (BERT and RoBERTa) in combination with Bidirectional Long Short-Term Memory (BiLSTM) networks trained on a comprehensive set of psycholinguistic features. First, we evaluate the performance of our models within-domain on two benchmark datasets: GoEmotion (Demszky et al., 2020) and ISEAR (Scherer and Wallbott, 1994). Second, we conduct transfer learning experiments on six datasets from the Unified Emotion Dataset (Bostan and Klinger, 2018) to evaluate their out-of-domain robustness. We find that the proposed hybrid models improve the ability to generalize to out-of-distribution data compared to a standard transformer-based approach. Moreover, we observe that these models perform competitively on in-domain data.

1 Introduction

Emotions are a key factor affecting all human behavior, which includes rational tasks such as reasoning, decision making, and social interaction (Parrott, 2001; Loewenstein and Lerner, 2003; Lerner et al., 2015; Bericat, 2016). Although emotions seem to be subjective by nature, they appear in objectively derivable ways in texts. Text-based emotion detection (henceforth TBED) is a branch of sentiment analysis that aims to extract textual features to identify associations with various emotions such as anger, fear, joy, sadness, surprise, etc. TBED is a rapidly developing interdisciplinary field that brings together insights from cognitive psychology, social sciences, computational linguistics, natural language processing (NLP) and machine learning (Canales and Martínez-Barco, 2014; Acheampong et al., 2020a; Alswaidan and Menai, 2020; Deng and Ren, 2021). TBED has a wide range of real-world applications, from healthcare (Cambria et al., 2010a), recommendation systems (Majumder et al., 2019), empathic chatbot development (Casas et al., 2021), offensive language detection (Plaza-del Arco et al., 2021), social data analysis for business intelligence (Cambria et al., 2013; Soussan and Trovati, 2020), and stock market prediction (Xing et al., 2018).

The differentiation of emotions and their classification into specific groups and categories is a subfield of affective research and has yielded several theories and models (Borod et al., 2000; Scherer et al., 2000; Cambria et al., 2012; Sander and Nummenmaa, 2021; Susanto et al., 2020). The grouping of models for the classification of emotions generally differs according to whether emotions are conceived as discrete/categorical or as dimensional. Categorical models of emotions, like Ekman's six basic emotions (anger, disgust, fear, joy, sadness, and surprise) (Ekman, 1992, 1999), assume physiologically distinct basic human emotions. Plutchik's Wheel of Emotion (Plutchik, 1984) is another categorical model that assumes a set of eight discrete emotions expressed in four opposing pairs (joy–sadness, anger–fear, trust–disgust, and anticipation–surprise). Dimensional emotion models, like the Circumplex Model of Russell (1980), groups affective states into a vector space of valence (corresponding to senti-

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 1 - 13
November 7, 2022 ©2022 Association for Computational Linguistics

ment/polarity), arousal (corresponding to a degree of calmness or excitement), and dominance (perceived degree of control over a given situation).

Current approaches to TBED take the advantage of recent advances in NLP and machine learning, with deep learning techniques achieving state-of-the-art performance on benchmark emotion datasets (see Acheampong et al. 2020a for recent reviews). However there still remains the issue of out-of-domain generalizability of the existing emotion detection models. The way emotions are conveyed in texts may differ from domain to domain, reflecting differences in topics, communicative goals, target audience, etc. This makes the deployment of such models in real-world sentiment and emotion applications difficult. The importance of this issue has been increasingly recognized in the TBED literature. For example, Bostan and Klinger (2018) emphasize that "[j]ournalists ideally tend to be objective when writing articles, authors of microblog posts need to focus on brevity", and that "emotion expressions in tales are more subtle and implicit than, for instance, in blogs". To support future transfer learning and domain adaptation work for TBED, the authors constructed a unified, aggregated emotion detection dataset that encompasses different domains and annotation schemes.

In this work, we contribute to the improvement of the generalizability of emotion detection models as follows: We build hybrid models that combine pre-trained transformer language models with Bidirectional Long Short-Term Memory (BiLSTM) networks trained, to our knowledge, on the most comprehensive set of psycholinguistic features. We evaluate the performance of the proposed models in two ways: First, we conduct within-corpus emotion classification experiments (training on one corpus and testing on the same) on two emotion benchmark datasets, GoEmotion (Demszky et al., 2020) and ISEAR (Scherer and Wallbott, 1994), to show that such hybrid models outperform pre-trained transformer models. Second, we conduct transfer learning experiments on six popular emotion classification datasets of the Unified Emotion Dataset (Bostan and Klinger, 2018) to show that our approach improves the generalizability of emotion classification across domains and emotion taxonomies. The remainder of the paper is organized as follows: In Section 2, we briefly review recent related work on TBED. Then, in Section 3, we present popular benchmark datasets for emotion detection. Section 4 details the extraction of psycholinguistic features using automated text analysis based on a sliding window approach. In Section 5, we describe our emotion detection models, and in Section 6, we present our experiments and discuss the results. Finally, we conclude with possible directions for future work in Section 7.

2 Related Work

In this section, we focus on previous TBED research conducted on two popular benchmark datasets (GoEmotions, ISEAR) to compare the performance of our models with state-of-the-art emotion recognition models, as well as previous attempts to improve generalizability using transfer learning techniques.

Current work on TBED typically utilizes a variety of linguistic features, such as word or character n-grams, affect lexicons, and word embeddings in combination with a supervised classification model (for recent overviews see, Sailunaz et al., 2018; Acheampong et al., 2020b; Alswaidan and Menai, 2020). While earlier approaches relied on shallow classifiers, such as a naive Bayes, SVM or MaxEnt classifier, later approaches increasingly relied on deep learning models in combination with different word embedding methods. For example, Polignano et al. (2019) proposed an emotion detection model based on the use of long short-term memory (LSTM) and convolutional neural network (CNN) mediated through the use of a level of attention in combination with different word embeddings (GloVe, Pennington et al. 2014, and Fast-Text, Bojanowski et al. 2017).

In experiments performed on the ISEAR dataset, Dong and Zeng (2022) proposed a text emotion distribution learning model based on a lexicon-enhanced multi-task convolutional neural network (LMT-CNN) to jointly solve the tasks of text emotion distribution prediction and emotion label classification. The LMT-CNN model is an end-to-end multi-module deep neural network that utilizes semantic information and linguistic knowledge to predict emotion distributions and labels. Based on comparative experiments on nine commonly used emotion datasets, Dong and Zeng (2022) showed that the LMT-CNN model can outperform two previously introduced deep-neural-network-based models: TextCNN, a convolutional neural network for text emotion classification (Kim, 2014) and MT-CNN (Zhang et al., 2018), a multi-task convolutional neural network model that simultaneously predicts the distribution of text emotion and the dominant emotion of the text (see Table 1 for numerical details on the performance of these models on the datasets used in the present work). In recent years, TBED research has increasingly relied on transformer-based pre-trained language models (Acheampong et al., 2020a; Demszky et al., 2020; **?**): For example, Acheampong et al. (2020a) perform comparative analyses of BERT (Devlin et al., 2019), RoBERTA (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019) for text-based emotion recognition on the ISEAR dataset. While all models were found to be efficient in detecting emotions from text, RoBERTa achieved the highest performance with a detection accuracy of 74.31%. The currently best-performing model on the ISEAR dataset, reaching a micro-average F1 score of 75.2%, is Park et al. (2021). In this work a RoBERTa-Large model was finetuned to learn conditional VAD distributions – obtained from the NRC-VAD lexicon (Mohammad, 2018) – through supervision of categorical labels. The learned VAD distributions were then used to predict the emotion labels for a given sentence.

For the recently introduced GoEmotions dataset, Demszky et al. (2020) already provided a strong baseline for modeling emotion classification by fine-tuning a BERT-base model. Their model achieved an average F1-score of 64% over an Ekman-style grouping into six coarse categories. **?** conducted comparative experiments with additional transformer-based models – BERT, DistilBERT, RoBERTa, XLNet, and ELECTRA (Clark et al., 2020) – on the GoEmotions dataset. As in the case of ISEAR, the best performance was achieved by RoBERTa, with an F1-score of 49% on the full GoEmotions taxonomy (28 emotion categories).

Previous TBED work has also proposed combinations of different approaches. For example, Seol et al. (2008) proposed a hybrid model that combines emotion keywords in a sentence using an emotional keyword dictionary with a knowledge-based artificial neural network that uses domain knowledge. To our knowledge, however, almost no TBED research has investigated hybrid models that combine transformer-based models with (psycho)linguistic features (see, however, De Bruyne et al. 2021, for an exception in Dutch). This is surprising, as such an approach has been successfully applied in related areas, for example personality prediction (Mehta et al., 2020; Kerz et al., 2022).

The available research aimed at improving the generalizability of transformer-based models using transfer learning techniques has so far focused on demonstrating that training on a large dataset of one domain, say Reddit comments, can contribute to increasing model accuracy for different target domains, such as tweets and personal narratives. Specifically, using three different finetuning setups – (1) finetuning BERT only on the target dataset, (2) first finetuning BERT on GoEmotions, then perform transfer learning by replacing the final dense layer, and (3) freezing all layers besides the last layer and finetuning on the target dataset –, Demszky et al. (2020) showed that the GoEmotions dataset generalizes well to other domains and different emotion taxonomies in nine datasets from the Unified Emotion Dataset (Bostan and Klinger, 2018).

3 Datasets

We conduct experiments on a total of eight datasets. The within-domain experiments are performed on two benchmark corpora: The GoEmotions dataset (Demszky et al., 2020) and the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset (Scherer and Wallbott, 1994). GoEmotions is the largest available manually annotated dataset for emotion prediction. It consists of 58 thousand Reddit comments, labeled by 80 human raters for 27 emotion categories plus a neutral category. While 83% of the items of the dataset have received a single label, GoEmotions is strictly speaking a multilabel dataset, as raters were free to select multiple emotions. The dataset has been manually reviewed to remove profanity and offensive language towards a particular ethnicity, gender, sexual orientation, or disability. The ISEAR dataset is a widely used benchmark dataset consisting of personal reports on emotional events written by 3000 people from different cultural backgrounds. It was constructed by collecting questionnaires answered by people that reported on their own emotional events. It contains a total of 7,665 sentences labeled with one of seven emotions: joy, fear, anger, sadness, shame, guilt and disgust. The transfer-learning experiments are conducted on six benchmark datasets from Unified Emotion Dataset (Bostan and Klinger, 2018) that were chosen based on their diversity in size and domain: (1) The **AffectiveText** dataset (Strapparava and Mihalcea, 2007) consists of 1,250 news headlines. The annotation schema follows Ekman's basic emotions, complemented by valence. It is multi-label annotated via expert annotation and emotion categories are assigned a score from 0 to 100. (2) The **CrowdFlower** dataset consists of 39,740 tweets annotated via crowdsourcing with one label per tweet. The dataset was previously found to be noisy in comparison with other emotion datasets (Bostan and Klinger, 2018). (3) The dataset **Electoral-Tweets** (Mohammad et al., 2015) targets the domain of elections. It consists of over 100,000 responses to two detailed online questionnaires (the questions targeted emotions, purpose, and style in electoral tweets). The tweets are annotated via crowdsourcing. (4) The Stance Sentiment Emotion Corpus **SSEC** (Schuff et al., 2017) is an annotation of 4,868 tweets from the SemEval 2016 Twitter stance and sentiment dataset. It is annotated via expert annotation with multiple emotion labels per tweet following Plutchik's fundamental emotions. (5) The Twitter Emotion Corpus **TEC** (Mohammad, 2012) consists of 21,011 tweets. The annotation schema corresponds to Ekman's model of basic emotions. They collected tweets with hashtags corresponding to the six Ekman emotions: #anger, #disgust, #fear, #happy, #sadness, and #surprise, therefore it is distantly single-label annotated. (6) The Emotion-Stimulus dataset (Ghazi et al., 2015) has 1,549 sentences with their emotion analysed. The set of annotation labels comprises of Ekman's basic emotions with the addition of shame. (7) The ISEAR$_{\text{UED}}$ dataset that is part of the Unified Emotion Dataset has 5,477 sentences with single emotion annotations. This dataset is a filtered version of the original ISEAR dataset described above. Bostan and Klinger (2018) filter and keep the texts with the labels anger, disgust, joy, sadness and fear for the Unified Emotion Dataset.

4 Sentence-level measurement of psycholinguistic features

The datasets were automatically analyzed using an automated text analysis (ATA) system that employs a sliding window technique to compute sentence-level measurements (for recent applications of this tool across various domains, see Qiao et al. (2020) for fake news detection, Kerz et al. (2021) for predicting human affective ratings) and Wiechmann et al. (2022) for predicting eye-moving patterns during reading). We extracted a set of 435 psycholinguistic features that can be binned into four groups: (1) features of morpho-syntactic complexity (N=19), (2) features of lexical richness, diversity and sophistication (N=77), (3) readability features (N=14), and (4) lexicon features designed to detect sentiment, emotion and/or affect (N=325). Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014).

The group of **morpho-syntactic complexity features** includes (i) surface features related to the length of production units, such as the average length of clauses and sentences, (ii) features of the type and frequency of embeddings, such as number of dependent clauses per T-Unit or verb phrases per sentence and (iii) the frequency of particular structure types, such as the number of complex nominals per clause. This group also includes (iv) information-theoretic features of morphological and syntactic complexity based on the Deflate algorithm (Deutsch, 1996). The group of **lexical richness, diversity and sophistication features** includes six different subtypes: (i) lexical density features, such as the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text, (ii) lexical variation, i.e. the range of vocabulary as manifested in language use, captured by text-size corrected type-token ratio, (iii) lexical sophistication, i.e. the proportion of relatively unusual or advanced words in a text, such as the number of words from the New General Service List (Browne et al., 2013), (iv) psycholinguistic norms of words, such as the average age of acquisition of the word (Kuperman et al., 2012) and two recently introduced types of features: (v) word prevalence features that capture the number of people who know the word (Brysbaert et al., 2019; Johns et al., 2020) and (vi) register-based n-gram frequency features that take into account both frequency rank and the number of word n-grams ($n \in [1, 5]$). The latter were derived from the five register subcomponents of the Contemporary Corpus of American English (COCA, 560 million words, Davies, 2008): spoken, magazine, fiction, news and academic language (see Kerz et al., 2020, for details see e.g.). The group of **readability features** combines a word familiarity variable defined by a prespecified vocabulary resource to estimate semantic difficulty along with a syntactic variable, such as average sentence length. Examples of these measures include the Fry index (Fry, 1968) or the

SMOG (McLaughlin, 1969). The group of **lexicon-based sentiment/emotion/affect features** was derived from a total of ten lexicons that have been successfully used in personality detection, emotion recognition and sentiment analysis research: (1) The Affective Norms for English Words (ANEW) (Bradley and Lang, 1999), (2) the ANEW-Emo lexicons (Stevenson et al., 2007), (3) DepecheMood++ (Araque et al., 2019), (4) the Geneva Affect Label Coder (GALC) (Scherer, 2005), (5) General Inquirer (Stone et al., 1966), (6) the LIWC dictionary (Pennebaker et al., 2001), (7) the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), (8) the NRC Valence, Arousal, and Dominance lexicon (Mohammad, 2018), (9) SenticNet (Cambria et al., 2010b), and (10) the Sentiment140 lexicon (Mohammad et al., 2013).

5 Modeling Approach

We construct a total of five models: (1) a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model, (2) a fine-tuned RoBERTA model (Robustly Optimized BERT pre-training Approach), (3) a bidirectional neural network classifiers trained on sentence-level measurements of psycholinguistic features described in Section 3.1, and (4) and (5) two hybrid models integrating BERT and RoBERTa predictions with the psycholinguistic features. We train all models in a multi-label classification setup. For the within-domain evaluation of the models on the GoEmotions dataset, we follow the procedure specified in Demszky et al. (2020): That is, we filtered out emotion labels selected by only a single annotator. The 93% of the original were randomly split into train (80%), dev (10%) and test (10%) sets. These splits are identical to those used by Demszky et al.. In the transfer learning setting geared to show that our modeling approach improves generalization across domains and taxonomies, we perform experiments on each of the six emotion benchmark datasets presented in section 3 using four approaches: with/without finetuning on target dataset and with/without the inclusion of the label 'neutral'. The performance of these models is evaluated using 5 times repeated 5-fold crossvalidation using a 80/20 split to counter variability due to weight initialization. We report performance metrics averaged over all runs. All models are implemented using PyTorch (Paszke et al., 2019). Unless specifically stated otherwise, we use 'BCELoss' as our loss function, 'AdamW' as optimizer, with learning rate 2×10^{-5} and weight decay of 1×10^{-5}

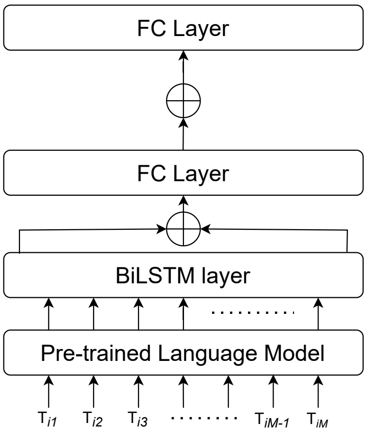

Figure 1: Structure diagram of transformer-based emotion detection models

5.1 Transformer-based models (BERT, RoBERTa)

We used the pretrained 'bert-base-uncased' and 'roberta-base' models from the Huggingface Transformers library (Wolf et al., 2020). The models consist of 12 Transformer layers with hidden size 768 and 12 attention heads. We run experiments with (1) a linear fully-connected layer for classification as well as with (2) an intermediate bidirectional LSTM layer with 256 hidden units (Al-Omari et al., 2020) (BERT-BiLSTM). The following hyperparameters are used for fine-tuning: a fixed learning rate of 2×10^{-5} is applied and $L2$ regularization of 1×10^{-6}. All models were trained for 8 epochs, with batch size of 4 and maximum sequence length of 512 and dropout of 0.2. We report the results from the best performing models, i.e. RoBERTa-BiLSTM and BERT-BiLSTM.

5.2 Bidirectional LSTM trained on psycholinguistic features (PsyLing)

As a model based solely on psycholinguistic features, we constructed a 2-layer bidirectional long short-term model (BiLSTM) with a hidden state dimension of 32, which is depicted in Figure 2. The input to the model is a sequence $CM_1^N = (CM_1, CM_2 \ldots, CM_N)$, where CM_i, the output of the ATA-system, for the ith sentence of a document, is a 435 dimensional vector and N is the sequence length. To predict the labels of a sequence, we concatenate the last hidden states of the last layer in forward ($\overrightarrow{h_n}$) and backward directions ($\overleftarrow{h_n}$). The resulting vector $h_n = [\overrightarrow{h_n}|\overleftarrow{h_n}]$ is

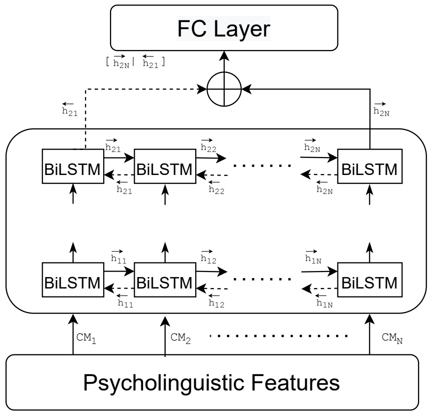

Figure 2: Structure diagram of BiLSTM emotion detection model trained on psycholinguistic features

then transformed through a 2-layer feedforward neural network, whose activation function is Rectifier Linear Unit (ReLU). The output of this is then passed to a Dense Fully Connected Layer with a dropout of 0.2, and finally fed to a final fully connected layer. The output of this is a K dimensional vector, where K is the number of emotion labels.

5.3 Hybrid models (BERT+PsyLing, RoBERTa+PsyLing)

We assemble the hybrid models by (1) obtaining a set of 256 dimensional vector from the PsyLing model and then (2) concatenating these features along with the output from the pre-trained transformer-based model part. To obtain the output of the pre-trained transformer-based model, the given text is fed to a pre-trained language model, its outputs are passed through a 2-layer BiLSTM with hidden size of 512. This is further passed through a fully connected layer to obtain a 256 dimensional vector. This concatenated vector is then fed into a 2-layer feedforward classifier. To obtain the soft labels (probabilities that a text belongs to the corresponding emotion label), sigmoid was applied to each dimension of the output vector.

6 Results

The models were evaluated using accuracy, precision, recall and F1 scores as the performance metrics. The results of the within-domain classification experiments on the GoEmotion and ISEAR datasets are shown in Table 1 (detailed results on all metrics are provided in see Table 4 in the appendix). We focus here on the discussion of F1 scores. For both datasets and for both transformer-based models, we find that the proposed hybrid models outperform the standard transformer-based baseline models: Specifically, in the case of the GoEmotions dataset, the hybrid models (BERT+PsyLing, RoBERTa+PsyLing) exhibit an increase in F1 score of +2% relative to their respective baseline models. In the case of the ISEAR dataset, the RoBERTa+PsyLing model show an increase in F1 score of +2% relative to RoBERTa, while the BERT+PsyLing model show an increase in F1 score of +1% relative to BERT. Our hybrid models show improvements in all emotion categories, except for anger, where they are on par with their respective baseline models. These results indicate that integrating transformer-based models with BiLSTM trained on psycholinguistic features can improve emotion classification within two distinct domains: an online domain (Reddit) as well as the domain of reports of personal events. On the GoEmotion dataset, our best-performing hybrid model, RoBERTa+PsyLing, outperforms the previous SOTA model Roberta-EMD (Park et al., 2021) by +9.9% macro-F1. On the ISEAR dataset, both hybrid models outperform two of the three CNNs presented in Dong and Zeng (2022), TextCNN and MT-CNN, and are competitive with the lexicon-enhanced multi-task CNN (LMT-CNN). In fact, both hybrid models outperform the LMT-CNN on two of the five emotion categories, with an increase on the joy category of +10.31% F1 (LMT-CNN vs. BERT-PsyLing) and an increase on the fear category of +4.05% F1 (LMT-CNN vs. BERT-PsyLing). The results of the comparisons with previous deep-learning TBED models on the two benchmark datasets thus indicate that the proposed approach constitutes a valuable framework for future TBED efforts.

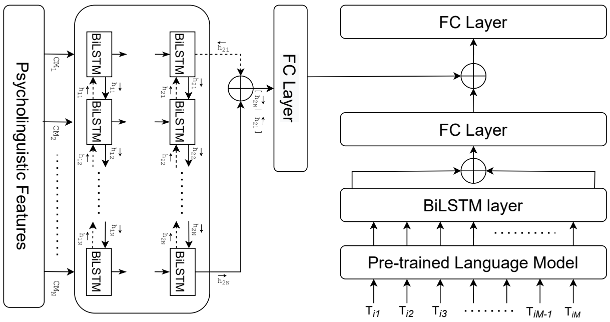

Figure 3: Structure diagram of hybrid emotion detection models

An overview of the results of the out-of-domain experiments is presented in Table 2. Table 3 shows

	GoEmotion Dataset						
Model	Anger	Disgust	Sadness	Surprise	Fear	Joy	Average
RoBERTa-EMD (Park et al., 2021)	–	–	–	–	–	–	61.1
BERT	70	48	64	72	72	90	68
RoBERTa	70	49	63	69	71	90	69
PsyLing	50	24	40	40	34	80	45
BERT+PsyLing (ours)	**71**	49	**65**	72	72	91	70
RoBERTa+PsyLing (ours)	70	**50**	**65**	**74**	**73**	**92**	**71**
	ISEAR Dataset						
TextCNN (Dong and Zeng, 2022)	62.14	65.22	76.39	–	72.09	73.97	69.96
MT-CNN (Dong and Zeng, 2022)	65.68	67.63	77	–	74.25	72.09	71.33
LMT-CNN (Dong and Zeng, 2022)	**66.54**	**70.64**	**80.68**	–	74.95	74.69	73.5
RoBERTa-EMD (Park et al., 2021)	**–**	**–**	**–**	**–**	**–**	**–**	**75.2**
BERT	56	65	71	-	77	84	71
RoBERTa	60	69	71	-	72	84	71
PsyLing	38	36	48	-	48	57	45
BERT+PsyLing (ours)	58	**70**	70	-	78	**85**	72
RoBERTa+PsyLing (ours)	**64**	69	**73**	-	**79**	79	73

Table 1: Results on the two benchmark datasets (GoEmotion (top), ISEAR (bottom)). All scores represent macro-averages of F1 scores(in %).

	Model	TEC	Crowdfl.	$\text{ISEAR}_{\text{UED}}$	elect-tweet	affect-text	SSEC	emo-stimulus
Train GoEmo	BERT	29	**23**	44	26	36	19	53
w/o finetuning	RoBERTa	**31**	**23**	44	**29**	39	21	56
w/o neutral	PsyLing	22	18	25	16	23	11	38
	BERT+PsyLing	**31**	**23**	44	27	36	21	56
	RoBERTa+PsyLing	29	**23**	**47**	27	**40**	**22**	**61**
w/o finetuning	BERT	20	26	35	23	13	16	41
with neutral	RoBERTa	22	27	34	**25**	14	**18**	47
	PsyLing	16	20	17	13	10	08	23
	BERT+PsyLing	21	27	35	24	15	17	45
	RoBERTa+PsyLing	**23**	**28**	**36**	**25**	**16**	17	**49**
with finetuning	BERT	55	31	63	36	54	**32**	92
w/o neutral	RoBERTa	**56**	30	**65**	34	53	**32**	**94**
	PsyLing	34	23	41	32	36	24	46
	BERT+PsyLing	55	32	**65**	39	**57**	**32**	**94**
	RoBERTa+PsyLing	**56**	**31**	**65**	**41**	**57**	**32**	**94**
with finetuning	BERT	46	33	55	33	44	29	96
with neutral	RoBERTa	44	**34**	**56**	30	46	30	95
	PsyLing	24	24	35	28	29	30	53
	BERT+PsyLing	**47**	34	55	34	**48**	31	**97**
	RoBERTa+PsyLing	46	**34**	**56**	**34**	47	**33**	96

Table 2: Results on transfer learning experiments. Values are macro-averaged F1 scores (in %).

Dataset	BERT	RoBERTa	PsyLing	BERT + PsyLing	RoBERTa + PsyLing	Bostan and Klinger, 2018
TEC	63	64	45	**67**	64	48
CrowdFlower	46	**47**	41	**47**	**47**	24
ISEAR$_{\text{UED}}$	76	**78**	49	**78**	**78**	52
elect-tweet	**62**	**62**	58	**62**	**62**	31
affect-text	63	63	48	**67**	**67**	64
SSEC	58	60	45	58	60	**67**
emo-stimulus	94	96	55	**97**	**97**	**97**

Table 3: Comparison of performance with Bostan and Klinger (2018). Values are micro-averaged F1 scores (in %).

comparisons of the results of our best performing model, RoBERTa+PsyLing, in the finetuning setting without the neutral label with the results of maximum entropy classifiers trained on with bag-of-words (BOW) features from Bostan and Klinger (2018). The results in Table 2 reveal that the RoBERTa+PsyLing hybrid model was the best performing model across all four experimental settings. Performance was generally observed to be highest in the finetuning setting without the neutral label. Importantly, the results in Table 2 reveal that the integration of psycholinguistic features matched or improved the performance of the models across all settings, with increases in F1 scores of up to 7% relative to a standard transformer-based approach. The results in Table 3 indicate that our hybrid models pretrained on GoEmotions outperform the results of the baseline models provided by Bostan and Klinger (2018) on five of the seven emotion datasets (TEC, CrowdFLower, ISEAR$_{\text{UED}}$, elect-tweet, and affect text), with increases in performance of up to 31%. The hybrid models tied the near-perfect performance of the baseline model on the emo-stimulus dataset and fell short only on the SSEC dataset. A possible reason for the relatively low performance of our models on the latter may be due to the fact that the SSEC was rated based on Plutchik's fundamental emotions.

7 Conclusion

This paper proposed approaches for text-based emotion detection that leverage transformer models in combination with Bidirectional Long Short-Term Memory networks trained on a comprehensive set of psycholinguistic features. The results of transfer learning experiments performed on six out-of-domain emotion datasets demonstrated that the proposed hybrid models can substantially improve model generalizability to out-of-distribution data compared to a standard transformer-based model. Moreover, we found that these models perform competitively on in-domain data. In future work, we intend to extend this line of work to dimensional emotion models as well as to models that jointly solve the tasks of emotion label classification and text emotion distribution prediction.

Ethical Considerations

The datasets used in this study may contain biases, are not representative of global diversity and may contain potentially problematic content. Potential biases in the data include: Inherent biases in user base biases, the offensive/vulgar word lists used for data filtering, inherent or unconscious bias in assessment of offensive identity labels. All these likely affect labeling, precision, and recall for a trained model.

References

Francisca Adoma Acheampong, Nunoo-Mensah Henry, and Wenyu Chen. 2020a. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020b. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Hani Al-Omari, Malak A. Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*.

Eduardo Bericat. 2016. The sociology of emotions: Four decades of progress. *Current Sociology*, 64(3):491–513.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Joan C Borod et al. 2000. *The neuropsychology of emotion*. Oxford University Press.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (ANEW): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Charles Browne et al. 2013. The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, 37(4):13–16.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2010a. Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. In *Development of multimodal interfaces: active listening and synchrony*, pages 148–156. Springer.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer.

Erik Cambria, Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. 2013. Big social data analysis. *Big data computing*, 13:401–414.

Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. 2010b. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43.

Jacky Casas, Timo Spring, Karl Daher, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 41–47.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Emotional robbert and insensitive bertje: Combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 257–263.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Peter Deutsch. 1996. Rfc1951: Deflate compressed data format specification version 1.3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Yuchang Dong and Xueqiang Zeng. 2022. Lexicon-enhanced multi-task convolutional neural network for emotion distribution learning. *Axioms*, 11(4):181.

Paul Ekman. 1992. Are there basic emotions? *Psychological review*, 99 3:550–3.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Edward Fry. 1968. A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Brendan T Johns, Melody Dye, and Michael N Jones. 2020. Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73(6):841–855.

Elma Kerz, Yu Qiao, and Daniel Wiechmann. 2021. Language that captivates the audience: predicting affective ratings of ted talks in a multi-label classification task. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–24.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 182–194, Dublin, Ireland. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, 44(4):978–990.

Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. Emotion and decision making. *Annual review of psychology*, 66:799–823.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

G. Loewenstein and J.S. Lerner. 2003. *The role of affect in decision making*, pages 619–642. Oxford University Press, Oxford.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

G Harry McLaughlin. 1969. Clearing the smog. *Journal of Reading*.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Saif Mohammad. 2012. # emotional tweets. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380.

W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and brnn. In *Proceedings of the 3rd international workshop on rumours and deception in social media (RDSM)*, pages 14–31.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.

David Sander and Lauri Nummenmaa. 2021. Reward and emotion: an affective neuroscience approach. *Current Opinion in Behavioral Sciences*, 39:161–167.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Klaus R Scherer et al. 2000. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Yong-Soo Seol, Dong-Joo Kim, and Han-Woo Kim. 2008. Emotion recognition from text using knowledge-based ann. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 1569–1572.

Tariq Soussan and Marcello Trovati. 2020. Improved sentiment urgency emotion detection for business intelligence. In *International Conference on Intelligent Networking and Collaborative Systems*, pages 312–318. Springer.

Ryan A Stevenson, Joseph A Mikels, and Thomas W James. 2007. Characterization of the affective norms for english words by discrete emotional categories. *Behavior research methods*, 39(4):1020–1024.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, pages 4595–4601.

A Appendix

Table 4: Detailed Results on the two benchmark datasets (GoEmotion (top), ISEAR (bottom))

		GoEmotion Dataset						
Model	Metric	Anger	Disgust	Sadness	Surprise	Fear	Joy	Average
RoBERTa-EMD (Park et al 2021)	F1	–	–	–	–	–	–	61.1
BERT	Pre	69	38	53	68	68	88	64
	Rec	71	65	80	77	76	91	77
	F1	70	48	64	72	72	90	68
RoBERTa	Pre	70	62	79	78	71	88	75
	Rec	71	41	53	62	70	93	65
	F1	70	49	63	69	71	90	69
PsyLing	Pre	48	28	47	43	42	80	48
	Rec	53	22	34	38	29	80	43
	F1	50	24	40	40	34	80	45
BERT+PsyLing (ours)	Pre	69	65	68	73	81	90	74
	Rec	71	40	63	69	56	90	65
	F1	**71**	49	**65**	72	72	91	70
RoBERTa+PsyLing (ours)	Pre	69	65	68	73	81	90	74
	Rec	71	40	63	69	56	90	65
	F1	70	**50**	**65**	**74**	**73**	**92**	**71**
		ISEAR Dataset						
TextCNN (Dong & Zeng 2022)	Pre	61.36	63.5	76.64	–	70.67	79.3	70.29
	Rec	70.84	64.24	74.21	–	71.66	64.59	69.11
	F1	62.14	65.22	76.39	–	72.09	73.97	69.96
MT-CNN (Dong & Zeng 2022)	Pre	61.31	64.68	80.27	–	72.16	81.13	71.91
	Rec	71.62	64.46	77.37	–	73.66	69.36	71.29
	F1	65.68	67.63	77	–	74.25	72.09	71.33
LMT-CNN (Dong & Zeng 2022)	Pre	62.28	66	82.07	–	72.5	82.15	73
	Rec	72.38	65.1	79.34	–	74.4	71.64	72.57
	F1	**66.54**	**70.64**	**80.68**	–	74.95	74.69	73.5
RoBERTa-EMD (Park et al 2021)	F1	**–**	**–**	**–**	**–**	**–**	**–**	**75.2**
BERT	Pre	51	74	74	-	83	84	73
	Rec	63	60	69	-	74	86	70
	F1	56	65	71	-	77	84	71
RoBERTa	Pre	58	68	77	-	93	86	77
	Rec	61	66	64	-	62	77	66
	F1	60	69	71	-	72	84	71
PsyLing	Pre	26	35	37	-	46	62	41
	Rec	62	34	63	-	48	53	41
	F1	38	36	48	-	48	57	45
BERT+PsyLing (ours)	Pre	55	73	72	-	80	84	73
	Rec	62	68	68	-	77	86	72
	F1	58	**70**	70	-	78	**85**	72
RoBERTa+PsyLing (ours)	Pre	66	72	79	-	80	80	75
	Rec	66	66	68	-	77	77	71
	F1	**64**	69	**73**	-	**79**	79	73