DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media

Mario Ezra Aragón[α β], **A. Pastor López-Monroy**[γ], **Luis C. González**[δ]
David E. Losada[α], **Manuel Montes-y-Gómez**[β]
[α] Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
[β] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
[γ] Centro de Investigación en Matemáticas (CIMAT), Mexico
[δ] Facultad de Ingeniería, UACh, Mexico
{ezra.aragon,david.losada}@usc.es, pastor.lopez@cimat.mx,
lcgonzalez@uach.mx, mmontesg@inaoep.mx

Abstract

Mental disorders affect millions of people worldwide and cause interference with their thinking and behavior. Through the past years, awareness created by health campaigns and other sources motivated the study of these disorders using information extracted from social media platforms. In this work, we aim to contribute to the study of these disorders and to the understanding of how mental problems reflect on social media. To achieve this goal, we propose a double-domain adaptation of a language model. First, we adapted the model to social media language, and then, we adapted it to the mental health domain. In both steps, we incorporated a lexical resource to guide the masking process of the language model and, therefore, to help it in paying more attention to words related to mental disorders. We have evaluated our model in the detection of signs of three major mental disorders: Anorexia, Self-harm, and Depression. Results are encouraging as they show that the proposed adaptation enhances the classification performance and yields competitive results against state-of-the-art methods.

1 Introduction

Mental disorders are among the most common illnesses worldwide. Some estimates[1] indicate that more than 50% of the population will be diagnosed with a mental disorder at some point in their lives. The prevalence of these disorders is highly concerning since they alter the way people think, feel, and take action, resulting in the incapacity of daily life routines. In addition, the recent COVID-19 pandemic triggered a serious global social and economic disruption, which had a direct effect on people's lives, and brought many challenges that can be stressful and overwhelming (Li et al., 2020). This situation was particularly difficult for people with mental health conditions and caused an increase in the prevalence of anxiety and depression (World Health Organization, 2022). There is therefore an increasing need for developing new tools to monitor the presence of mental disorders and to respond to early signs of psychological concerns.

Nowadays, social media content is massive and provides an opportunity to do research on how people undergo difficulties. Many people use online platforms to publicly share their daily routines and important events, while others take advantage of the anonymity of these spaces to explicitly discuss mental health issues and to seek help (Ríssola et al., 2021; Crestani et al., 2022). In this work, we aim to contribute to the detection of signs of mental disorders by automatically analyzing social media posts. This type of analysis is expected to support new technologies able to warn about the onset of mental disorders and provide supporting evidence. As argued by Neuman et al. (2012), these new forms of screening should not be taken as "magic substitutes for the human expert" but, instead, as computational tools that can substantially reduce the workload of public health systems, e.g., by facilitating preventive measures.

Latest developments in Natural Language Processing (NLP) encourage the fine-tuning of pre-trained language models for a wide variety of tasks. This approach often yields good results, but it is problematic for tasks where the language is highly domain-specific (Villa-Cueva et al., 2022), such as in the case of mental disorders. An alternative is to pre-train the model –e.g., BERT (Devlin et al., 2019)– with data from the target domain. However, pre-training is expensive and complex, and collecting a sufficiently large training corpus can

[1] Center for Disease Control and Prevention, https://www.cdc.gov/mentalhealth/learn/index.htm

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15305–15318

be difficult for certain domains, as exposed by the creators of MentalBERT (Ji et al., 2022).

Instead of pre-training the model, we propose to perform a domain adaptation, similar to the one proposed in (Howard and Ruder, 2018) but refined in two stages (Villa-Cueva et al., 2022) that exploit a novel lexicon-driven learning. This process takes an already trained model and continues its training using a (relatively) small corpus focused on the target domain (Gururangan et al., 2020). In particular, we propose DisorBERT, a two-step domain adaptation model of BERT for detecting signs of mental disorders in social media. First, we teach BERT the general structure of the language used in social media texts (e.g., in Reddit posts), then we specialize it in the kind of language used to express information about mental disorders. Furthermore, we exploit lexical knowledge to guide the language model's masking process. Instead of learning the occurrence of general words, this "guided masking" opts to bias the learning process towards words that are important to the target application, which in our case is the detection of Anorexia, Depression, and Self-harm. We can summarize our contributions as follows:

1. We introduce DisorBERT, a simple yet effective double-domain adaptation model for detecting signs of mental disorders in social media.

2. We explore the use of lexical knowledge, extracted from a depression lexicon, to guide and enhance the masking process of the language model.

3. We empirically evaluate the proposed model and provide quantitative and qualitative evidence of its robustness for the detection of signs of Anorexia, Depression, and Self-harm in social media.

2 Related Work

The detection of mental disorders is an interdisciplinary research area that has been fostered thanks to the current availability of a variety of data sources and computational models (Velupillai et al., 2019). In recent years, several works have explored social media platforms to study the manifestation of mental disorders (Skaik and Inkpen, 2020; Ríssola et al., 2021). Social media sources have been exploited to detect features that help to identify signs of need of medical or psychological support (Calvo et al., 2017). For example, expressions of distress or negative feelings, particularly published by young people (Robinson et al., 2016), abound in online media. A variety of methods have been applied to find relevant and discriminative patterns from user-generated text. For example, some studies employed words or word sequences as features (Tsugawa et al., 2015; Schwartz et al., 2014; Ning et al., 2018). Other groups of studies have applied sentiment analysis techniques to model emotional properties of users' posts (Ramírez-Cifuentes and Freire, 2018; Preoţiuc-Pietro et al., 2015), or exploited a set of psychological categories to capture social relationships, thinking styles as well as individual differences (Coppersmith et al., 2015). In Cheng et al. (2017) and O'Dea et al. (2017), the authors explored the association between linguistic inquiry features and suicide risk factors, extracting patterns in different linguistic profiles. Similarly, for the detection of signals related to suicide attempts, Coppersmith et al. (2018) used word embeddings and a bidirectional Long Short-Term Memory (LSTM) to capture contextual information. Furthermore, for detecting suicidal ideation, Ramírez-Cifuentes et al. (2020) explored the incorporation of images into text-based representations. These authors analyzed the combination and relationship of textual and visual information, identifying significant differences in the use of both. More recently, with the increasing popularity of transformers, BERT-based classifiers have been fine-tuned for detecting different mental disorders (Martínez-Castaño et al., 2020; Parapar et al., 2022).

On the other hand, a number of studies have trained language models for specific domains (Zihan et al., 2021). For example, Beltagy et al. (2019) exploited a large-scale annotated dataset with scientific data to adapt BERT to the scientific domain, showing improvements over the BERT base model in multiple classification tasks. Similarly, in Lee et al. (2019), BERT was pre-trained on large-scale biomedical corpora outperforming the original BERT in a variety of biomedical text mining tasks. In recent work, Nguyen et al. (2020) trained a BERT model with approximately 850M tweets achieving outstanding results in several tweet analysis tasks.

Closer to our work, Ji et al. (2022) adapted BERT to the mental health domain by collecting specific

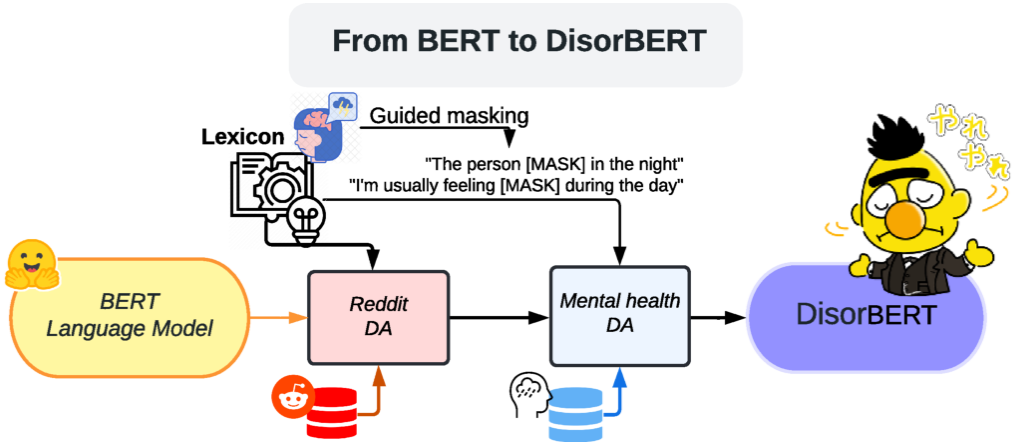

Figure 1: General diagram of the double domain adaptation process. It starts with a base language model, and then it integrates information from Reddit and from mental health information sources. The training of the language model also incorporates some lexicon knowledge to guide the masking process.

data from Reddit. Their model, named MentalBERT, was trained using 13,671,785 sentences for around eight days using four Tesla v100 GPUs. The aforementioned pre-training approach is usually effective, but it is also expensive as collecting a corpus of a suitable size to pre-train BERT can be a big challenge in many domains. To overcome this issue, we propose to perform a double-domain adaptation of BERT, a less-expensive process that takes advantage of the already trained general model. We explain this strategy in detail in the following section.

3 From BERT to DisorBERT: A Mental Disorder Detection Model

This section introduces DisorBERT, a language model specially suited for the detection of signs of mental disorders in Social Media. As mentioned above, the construction of DisorBERT consists of a two-stage domain adaptation of BERT. In short, the idea is to first teach BERT the general structure of the language used in a large social media platform (e.g., in Reddit), and next, to specialize the model in the language of users with mental disorders. This whole process is depicted in Figure 1.

For domain adaptation, we follow the procedure suggested by Howard and Ruder (2018) and Wolf et al. (2022), which consists of continuing BERT's pre-training through the fine-tuning of a Masked Language Model for more epochs. The process first uses the Reddit corpus for fine-tuning, and then, in a second adaptation, uses a collection of documents related to mental health. The idea is to adapt the language model from the general data of Wikipedia and books corpora (BERT's sources for training) to the more specific language of Reddit and mental health. For this process, we trained our model for three epochs, using a batch size of 128, a learning rate of $2e^{-5}$, on a GPU NVIDIA Tesla V100 32GB SXM2. It is particularly relevant to note that, in each of these steps, we employed a depression lexicon to guide the model's masking process and, therefore, to bias the learning towards words that are important to the target application domain.

3.1 Adapting the Model to Reddit

The first step of our domain adaptation involves the adjustment of the model to the language style used in social media sources. To this end, we used the corpus from Kim et al. (2019), which contains pre-processed posts from Reddit for the task of text summarization. There are more than 120k text-summary pairs discussing diverse topics and interests; for our experiments, we used both.

As data pre-processing, we concatenated all the examples and then split the whole corpus into chunks of equal size (128 words). The next step was to mask some of the words in each batch, picking 20% of them. It is worth mentioning that this percentage is within the typical range used for BERT, and it is a common choice in the literature. In Subsection 3.3, we explain in more detail the masking strategy.

3.2 Adapting the Model to Mental Health

Our second step adapts the model to the mental health domain. To accomplish this, we used four datasets extracted from subreddits containing infor-

mation related to mental health and depression[2,3,4]. Overall, we obtained more than 105k posts related to mental disorders. Focusing on these topics, our model can learn to identify how users with mental disorders express themselves as well as the language they use in social networks. Observe that this includes text from a large variety of users (e.g., some users may express negative feelings in their publications but they might not suffer from psychological problems).

In this case, we also concatenated all the examples and split the whole data set into chunks for the fine-tuning of the language model.

3.3 Guided Masking of Language Model

In order to improve the learning process, we incorporate lexical knowledge related to mental disorders. To train masked language models, it is common to employ random masking. This technique consists of selecting a random number of words within a sentence and asking the model to predict the hidden word. With this technique, the main idea is that the model can learn the context in which words occur. For our study, we incorporated knowledge from a lexical resource during the masking process. Instead of randomly masking words, we first checked if the text had words from our lexicon. If so, the lexicon words are masked to begin the training. In the event that the masked words within the original text did not complete the required 20%, we added additional random words to the masking. This new form of masking helps the model pay more attention to words that are related to mental disorders. The hypothesis is that the model built in this way should be able to more easily identify users who show signs of disorders. Once the model has been trained, we can proceed to specialize it in the downstream detection tasks by applying a traditional fine-tuning process.

The reference resource for the proposed "guided masking" is a depression lexicon built by Losada and Gamallo (2020). This is one of the few publicly available lexicons focusing on depression. Its word list resulted from expanding an already existing terminological resource by exploiting distributional strategies and lexical relationships such as synonymy. Here, we augmented this lexicon by adding different verb tenses to all the verbs, for example, the verb "abandon" will lead to words such as "abandons", "abandoned", and "abandoning". This addition helps to cover cases where people describe situations not happening in the present (e.g. when they refer to past events). Observe that we used a single (depression-oriented) lexicon to support the identification of risks for three different tasks, including Anorexia and Self-harm.

[2]https://huggingface.co/datasets/ShreyaR/DepressionDetection

[3]https://huggingface.co/datasets/hugginglearners/reddit-depression-cleaned

[4]https://huggingface.co/datasets/jsfactory/mental_health_reddit_posts

4 Experimental Settings

4.1 Collections

For the evaluation, we used the datasets from the eRisk 2019-2020 evaluation tasks (Losada et al., 2019, 2020). These tasks propose the detection of signs of anorexia, depression, and self-harm. Table 1 shows general information about the collections and how the classes are distributed. For depression, we used the data set from eRisk 2018 (Losada et al., 2018) for training our models.

Data set	**Train**		**Test**	
	P	C	P	C
Anor'19	61	411	73	742
avg # posts	407.8	556.9	241.4	745.1
avg # words	37.3	20.9	37.2	21.7
Dep'20	214	1493	40	49
avg # posts	440.9	660.8	493.0	543.7
avg # words	27.5	22.75	39.2	45.6
SH'20	41	299	104	319
avg # posts	169.0	546.8	112.4	285.6
avg # words	24.8	18.8	21.4	11.9

Table 1: Datasets used for experimentation. P indicates the positive users and C is used for control users.

The eRisk organizers provided datasets containing the post history of several users from Reddit. For each task, we have two categories: *i*) positive users affected by either anorexia, depression, or self-harm, and *ii*) a control group composed of people who do not suffer from these mental disorders. For the anorexia and self-harm tasks, the organizers obtained the positive users searching for people who explicitly mentioned that they were diagnosed by a medical specialist with one of these conditions. Vague expressions like "I think I have anorexia" or "I am anorexic" were not considered as expressions of a diagnostic. On the other hand, the control group contains random users from different subReddits and users who often interact in

the anorexia, depression, or self-harm threads. This adds more realism to the data as the control group includes, for example, expert clinicians who are active in mental health subReddits because they give support and advice to other people. Thus, risk prediction technology cannot be merely based on distinguishing the topic of the conversations.

For the depression task, organizers asked users to fill out the BDI questionnaire (Beck et al., 1961) (thus obtaining the estimated level of severity of their depression). In the eRisk depression task of 2020, participants were given the thread of users' social media postings and they were asked to estimate the severity of the depression (where the real BDI questionnaires acted as the ground truth). For our study, we exclusively focused on a binary task (similar to anorexia and self-harm), i.e., to distinguish between positive and control users. So, we split the users into two categories based on the BDI scores. The positive class contains the users that obtained 21 or more points in the questionnaire (according to the medical literature a score higher than 20 is indicative of the presence of moderate or severe depression). The control group contains the rest of the individuals (BDI scores lower than 21).

4.2 Model Configuration

Pre-processing: We performed a simple preprocessing of the texts by lowercasing all words and removing special characters like URLs, emoticons, and hashtags.

Training and predictions: For each user, we separated the post history into $N = 35$ segments. We selected this value empirically, after testing some sizes of sequences recommended in the literature. For training, we processed each segment of the post history as an individual input or item and trained the model. For the test, each segment receives a label of 1 or 0; then, if the majority of the items are positive, the user is classified as a possible case of risk. The main idea is to consistently detect the presence of major signs of anorexia, depression, or self-harm through all the user posts.

Parameters: We used the models provided by HuggingFace v4.24.0 (Wolf et al., 2022), and Pytorch v1.13.0 (Paszke et al., 2019). In particular, for training the model we used a batch size of 256, Adam optimizer, with a learning rate of $1e^{-5}$, and cross-entropy as a loss function. We trained the models for three epochs using a GPU NVIDIA Tesla V100 32GB SXM2.

4.3 Baseline Approaches

Bag-of-Words: We employed a traditional Bag-of-Words (BoW) approach considering word unigrams and TF-IDF weights. For feature selection, we applied the Chi-Square test and used a Support Vector Machine (SVM) with a linear kernel as a classifier. We also explored alternative BoW classifiers, but an SVM was the best-performing choice in our experiments.

Deep Neural Networks: We used CNN and Bi-LSTM networks. These neural networks used 100 neurons, an Adam optimizer, and Glove (Pennington et al., 2014) embeddings with a dimension of 300. For the CNN, we used 100 random filters of sizes 1 and 2.

BERT: We employed a BERT-based classification model with fine-tuning over each training set.

MentalBERT: This is a pre-trained language model for the mental healthcare domain. It was built from a large collection of sentences extracted from Reddit (Ji et al., 2022). Similar to BERT, we fine-tuned this model over each training set.

For each baseline, we explored different parameters using manual and grid search (depending on the model) and selected the best-performing setting.

In addition to the previous approaches, we also compared our results against those of the participants of the eRisk evaluation shared tasks. For this comparison, we considered the F_1 score, precision, and recall over the positive class, as reported in (Crestani et al., 2022).

5 Evaluation and Discussion

Table 2 shows the results of our approach and all baseline methods. It also includes the results of our approach using only Reddit adaptation, only mental health adaptation, and random masking instead of guided masking.

The first thing to notice is that BERT performed well but MentalBERT and DisorBERT are better choices, highlighting the importance of having domain-oriented models. Going into detail, we can observe that most of our proposed models outperformed the baselines in terms of F_1. Our single-domain adaptations obtained slight improvement in comparison with baselines, while the double-domain adaptation further increased performance, particularly with the incorporation of lexical knowledge. Here it is important to highlight that the lexicon employed is not specific to the language of

Method		Anorexia			Depression			Self-Harm		
	Masking	F1	P	R	F1	P	R	F1	P	R
Baselines										
BoW-SVM	–	0.67	0.85	0.55	0.58	0.56	0.60	0.50	**0.95**	0.34
RNN-GloVe	–	0.65	0.92	0.51	0.58	**0.59**	0.57	0.57	0.62	0.53
CNN-GloVe	–	0.67	0.93	0.52	0.61	0.56	0.68	0.57	0.62	0.53
BERT	Random	0.77	0.70	0.85	0.62	0.55	0.72	0.60	0.44	**0.94**
MentalBERT	Random	0.76	0.66	**0.89**	0.67	0.57	0.80	0.71	0.62	0.84
Our methods: Single and Double Domain Adaptation										
BERT w/Reddit	Random	0.81	0.75	0.88	0.66	0.56	0.80	0.71	0.66	0.76
BERT w/Reddit	Guided	0.82	0.82	0.82	0.68	0.55	0.90	0.72	0.65	0.82
BERT w/Health	Random	0.80	0.77	0.84	0.67	0.53	0.93	0.69	0.60	0.82
BERT w/Health	Guided	0.82	0.81	0.84	0.68	0.57	0.85	**0.74**	0.72	0.76
DisorBERT	Random	0.82	**0.83**	0.81	0.68	0.54	**0.93**	0.72	0.65	0.80
DisorBERT	Guided	**0.83**	0.82	0.85	**0.69**	0.56	0.89	0.72	0.73	0.71

Table 2: F1, precision (P), and recall (R) results over the positive class in three eRisk tasks. For the sake of completeness, we include results corresponding to single domain adaptations, "BERT w/Reddit" indicates the model only adapted to Reddit texts, and "BERT w/Health" is the model only adapted to mental health language.

anorexia or self-harm but, still, it was also beneficial for these two target tasks.

From another perspective, DisorBERT showed a good balance between precision and recall, whereas other variants (e.g., RNN-GloVe) improved precision at the expense of recall. DisorBERT has, therefore, a solid retrieval behavior and it can effectively find multiple traces of psychological risks. This is an important outcome since high recall is essential for clinical screening tools. However, there might be some potential use cases where high recall is not the most preferable choice, e.g., a social network that wants to focus on the riskiest behavior. For these scenarios, it may be necessary to modify our model to prioritize precision. In Figure 2, we plot the precision and recall of DisorBERT and the baselines. Our model tends to locate in the main diagonal region (indicating its good balance), while other methods have high precision or recall but score low in the other dimension.

For a more detailed analysis, we applied McNemar's statistical test (Rotem et al., 2020) to compare the best DisorBERT results with the best baseline results, Table 3 shows this comparison. The symbol '=' means not significantly different ($p > 0.5$), '*' means significantly different ($p < 0.05$), '**' means very significantly different ($p < 0.01$), and '***' (highly significantly different: $p < 0.001$). The results suggest that the proposed approaches differ significantly from the baselines.

Task	MB	BERT	CNN	SVM	BH
Anor	***	***	***	***	*
Dep	*	***	***	***	*
SH	*	***	***	***	**

Table 3: Pairwise significance differences between DisorBERT and the baseline models using McNemar's test comparison. MB = MentalBERT, BH = BERT w/Health.

Comparison against eRisk participants:

Figure 3 presents a boxplot of the F_1 scores of all participants for the anorexia and self-harm eRisk shared tasks[5]. The red circles represent the best DisorBERT model. For both tasks DisorBERT gets to the highest quartile, and, especially, in the anorexia detection task, our result is above the highest-scoring participant. These results indicate that our approach is competitive in comparison with the participants. However, it is important to mention that eRisk participants focused on obtaining early and accurate predictions of the users, while our approach focuses exclusively on determining accurate classifications.

Overall, we can highlight the following conclusions of the experimental results:

- The combined effect of double domain adaptation and guided masking is effective at cap-

[5]We cannot include comparisons on the eRisk 2020 depression detection task because 2020 participants were assessed using other metrics.

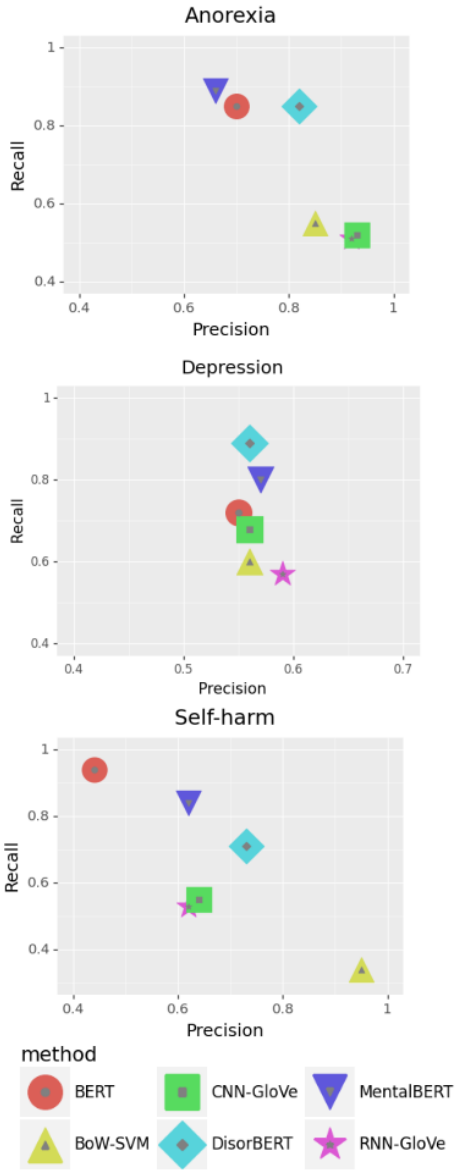

Figure 2: Precision and recall results of DisorBERT and the baselines for the three tasks.

turing signs of mental disorders in social media interactions; DisorBERT performed better than the original BERT model.

- Our approach also obtained better results than those achieved by MentalBERT, a model trained with a larger amount of data and with higher consumption of computational resources. The proposed double-domain adaptation is effective and computationally lightweight.

- The evaluation showed a solid balance between finding users and labeling them correctly, making DisorBERT suitable for clinical detection applications.

Figure 3: Boxplots for the eRisk comparison. The red O represents DisorBERT.

6 Analysis of the models

6.1 BERT vs DisorBERT

BERT is a language model trained from a general corpus, while DisorBERT is a model guided to the mental health domain. Let us illustrate the behavior of the learned model, and the kind of textual segments it tends to pay more attention to. First, we analysed the most likely words the models generate when given a sentence with masked words. As sentences, we used examples from the Beck Depression Inventory (BDI) (Beck et al., 1961). This clinical tool, which consists of 21 items, aims to identify and measure the severity of typical symptoms of depression in adults and adolescents aged 17 and older. For example, it measures mood, pessimism, sense of failure, self-dissatisfaction, and guilt, among others.

This test gives several responses for each item. We selected one of the answers for each one, masked a keyword, and looked at the words predicted by BERT and DisorBERT. In Table 4, we can see some examples of these sentences and the answers returned (ordered by decreasing likelihood). With DisorBERT, the answers tend to have a more negative meaning or psychological orientation compared to BERT. Take, for example, the sentence "I used to be able to [MASK]", where DisorBERT predicts the words "focus", "talk", "breathe", "sleep", and "eat". These words are related to common problems that are associated with mental disorders and cause interference in the thinking and behavior of the affected person. The BERT model is more general whereas DisorBERT learns to focus on issues related to mental disorders.

Let us now look at the models in a different way. For each BDI sentence, we know the target masked word and we can extract the position of this word in the ranked list provided by each model. For example, in the third case of Table 4, DisorBERT made a perfect job because the correct word ("killing") was the top-ranked word, while BERT put the correct word in the second position of the list. Mean reciprocal rank (MRR) is a natural way to quantitatively measure the ability of the models to find the correct word. It is a standard search effectiveness measure that compares systems in terms of their ability to rank the correct answer at top rank positions. To calculate this value, we generated the top 5 words for each sentence and averaged the reciprocal ranks[6] for all the answers (if the correct word is not in the ranked list then the system gets a RR equal to 0 for that sentence). BERT obtained an MRR of 0.2436 and DisorBERT 0.4325. This demonstrates that DisorBERT does a substantially better job at learning the language of the BDI inventory, which is a reference clinical tool to measure the prevalence of depression symptoms. Nevertheless, our model still struggled with several BDI items, showing the difficulty of the task.

Phrase	BERT	DisorBERT
"I used to be able to cry"	read	focus
"I used to be able to "	talk	talk
[MASK].	fly	breathe
	walk	sleep
	breathe	eat
"I hate myself."	asked	hate
"I [MASK] myself."	told	kill
	shook	killed
	brace	blame
	hugged	love
"I have thoughts of	him	killing
killing myself."	killing	hurting
"I have thoughts of	it	dying
[MASK] myself."	death	kill
	her	destroying

Table 4: Comparison on the prediction of masked examples for BERT and DisorBERT (the prediction are ranked based on their likelihood). The examples are taken from Beck Depression Inventory (BDI).

To further analyze these models, in Figure 4 we present all the words predicted by both models,

[6]The reciprocal rank (RR) is the multiplicative inverse of the rank of the correct answer, assigning values of 1 for first place, $1/2$ for the second, $1/3$ for the third, and so on.

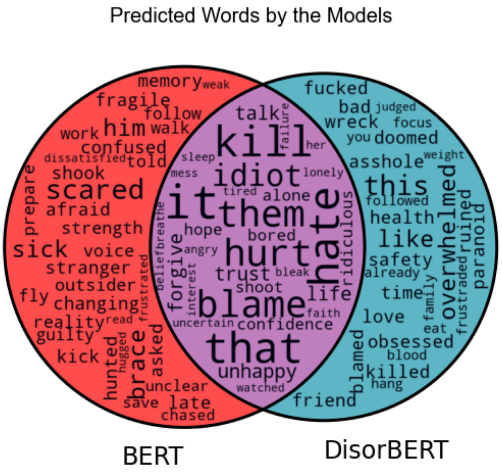

Figure 4: BDI words predicted distribution for both models.

weighted by their frequency. This figure results from applying the two models to the entire set of 21 BDI items. Similar to what happened before, BERT tends to generate more general words, while our model tends to be biased toward words related to mental disorders. Even with words in common such as "kill" or "hate", the frequency of occurrence of those words is higher for DisorBERT (i.e., it predicted several of these words more times than BERT). Finally, to measure how different these two groups of words are, we calculated the cosine similarity between the two sets of words and obtained a value of 0.4767. This suggests that there is some agreement among the proposed completions, but the predictions between the two models largely differ.

6.2 Adding interpretability to the detection of signs of mental disorders

Although incorporating transformers led to enhanced performance in contrast to other techniques, it also complicates the analysis and visualization of the model, which is important to understand its behavior. For models based on transformers, the attention scores in the head modules provide useful information about the words or sequences that are relevant for detection. However, with multiple heads and layers, the analysis becomes more difficult. Pardo-Sixtos et al. (2022) proposed a visualization tool that provides an interactive head view in the form of a graph. This tool starts with the [CLS] token, then, searches for the tokens in the previous layer that are important (highest attention values) for [CLS]. This visualization allows us

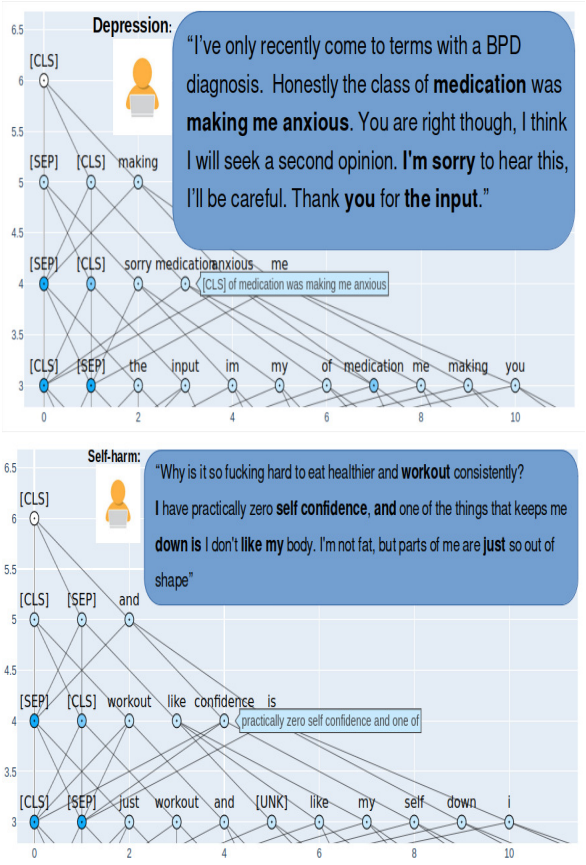

Figure 5: Graph example from a user with depression (upper) and self-harm (lower). The figure shows the most relevant words in the publications, according to the highest attention values.

to analyze the most important sequences of the text by obtaining the most relevant words and sentences in each layer of the transformer module.

For this analysis, we selected a depression user with the highest score in the BDI questionnaire and a user who self-harmed, and computed the attention scores of the user's posts. We used the attention scores in the head module to visualize the parts of the posts that are important for the classification. Figure 5 shows an example of the graphs generated. In this way, we can understand the words and contexts that are relevant to the classification. For example, in the upper graph the most prominent words are related to anxiousness and medication, topics that are highly relevant to depression. In the lower graph (self-harm case), the prominent words are related to low self-confidence. It is interesting to see how the model can focus on mental health issues and pay more attention to related contexts.

7 Conclusion

In this study, we explored a Double Domain Adaptation approach for the tasks of detecting signs of Anorexia, Depression, and Self-harm in social media. The first step of the domain adaptation focused on learning the writing style of social media users. The second step was oriented to learn about mental disorders and how users refer to psychological issues. In both steps, we incorporated lexical knowledge to guide the model toward words that are highly indicative of mental-related topics. Results suggest that combining domain adaptation with lexical knowledge helps in detecting traces of mental disorders. This approach outperformed traditional and state-of-the-art baselines and is competitive with the performance of top early-risk algorithms. Furthermore, the analysis of our method revealed that the context learned by the model is important in getting a better understanding of the concerns expressed by people.

In future work, we want to explore the application of different lexical resources that are even more specialized for the target tasks, as well as the usage of clinical data to train more specialized language models, e.g., MIMIC (Johnson et al., 2016). On the other hand, emojis are often important features for social media analysis, and we want to explore their incorporation into our training process. Also, we are interested in expanding this study to different languages, since most of the work related to mental disorders has focused on English.

Limitations

This study aims to detect signs of Anorexia, Self-harm, and Depression in users of social media environments through a double-domain adaptation of a language model. This study presents some limitations, mainly because these datasets are observational studies and we do not have access to the personal and medical information that is often considered in risk assessment studies. For example, we cannot discard that some users who publicly expressed that they have been diagnosed with anorexia are actually non-anorexia cases. However, the identification of positive users from self-expressions of diagnosis is a common practice in this area (Coppersmith et al., 2014), and the test collections built in this way are regarded as solid experimental benchmarks. There are also some limitations given by the nature of the data, as the users in these datasets might differ from users at

risk who do not have exposure to social media (e.g., elderly people or individuals who do not have an online account or decided to not make their profiles public).

Ethics Statement

When we analyze social media content, we may have concerns regarding individual privacy or certain ethical considerations. These concerns appear due to the usage of information that could be sensitive and personal (e.g., references to emotions and health concerns). It is also important to mention that these datasets could contain biases belonging to the nature of social media data, e.g., a gender, age, or sexual orientation profile that could cause someone to be mislabeled as having a mental disorder. The experiments and usage of this data are for research and analysis only, and the misuse or mishandling of information is prohibited. In any case, the datasets we employed (corpus by Kim (Sect 3.1), Reddit datasets (Sect 3.2), lexicon data set (Sect 3.3), and eRisk collections (Sect 4.1) are publicly available and we strictly followed the terms of use and user agreement of these collections (see e.g. https://tec.citius.usc.es/ir/code/eRisk2019.html). Furthermore, these collections are anonymized and our research does not involve any contact with social media users. Under such conditions, this research does not require review and approval by the Ethics Committee Board.

Acknowledgements

Mario Ezra Aragon and David E. Losada thank the support obtained from (i) project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación Resiliencia, Unión Europea-Next GenerationEU), and (ii) Consellería de Educación, Universidade e Formación Profesional (accreditation 2019–2022 ED431G-2019/04, ED431C 2018/29) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System. Mario Ezra Aragon also thanks to INAOE for the collaboration grant awarded from August to December 2022.

References

Aaron Beck, C.H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *Arch Gen Psychiatry*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Qijin Cheng, Tim Mh Li, Chi-Leung Kwok, Tingshao Zhu, and Paul Sf Yip. 2017. Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study. *J Med Internet Res*, 19(7).

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights*, 10.

Fabio Crestani, David E. Losada, and Javier Parapar. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*. Springer Verlag, Englewood Cliffs, NJ.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Alistair E. Johnson, Tom J. Pollard, Lu Shen, Li Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and G. Mark Roger. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:1–9.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiaoya Li, Mingxin Zhou, Jiawei Wu, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Analyzing COVID-19 on online social media: Trends, sentiments and emotions. *CoRR*, abs/2005.14464.

David Losada and Pablo Gamallo. 2020. Evaluating and improving lexical resources for detecting signs of depression in text. *Lang Resources & Evaluation*, 54:1–24.

David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 343–361, Cham. Springer International Publishing.

David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, Cham. Springer International Publishing.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 272–287, Cham. Springer International Publishing.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers : ilab at clef erisk 2020. *CEUR Workshop Proceedings*, 2696. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020. urn:nbn:de:0074-2696-0.

Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. 2012. Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, 56(1):19–25.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Liu Ning, Zhou Zheng, Xin Kang, and Ren Fuji. 2018. Tua1 at erisk 2018. *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*.

Bridianne O'Dea, Mark E Larsen, Philip J Batterham, Alison L Calear, and Helen Christensen. 2017. A linguistic analysis of suicide-related twitter posts. *Crisis*, 35(5):319–329.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview of erisk 2022: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 233–256, Cham. Springer International Publishing.

L. Fernando Pardo-Sixtos, A. Pastor López-Monroy, Mahsa Shafaei, and Thamar Solorio. 2022. Hierarchical attention and transformers for automatic movie rating. *Expert Systems with Applications*, 209:118164.

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 *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky,

H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Diana Ramírez-Cifuentes and Ana Freire. 2018. Upf's participation at the CLEF erisk 2018: Early risk prediction on the internet. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi González. 2020. Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. *J Med Internet Res*, 22.

Esteban A. Ríssola, David E. Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2).

Jo Robinson, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Rodrigues Maria, Steve Fisher, and Helen Herrman. 2016. Social media and suicide prevention: a systematic review. *Early Interv Psychiatry*, 10(2):103–121.

Dror Rotem, Peled-Cohen Lotem, Shlomov Segev, and Reichart Roi. 2020. *Statistical Significance Testing for Natural Language Processing*. Springer Chamg.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: A review. *ACM Comput. Surv.*, 53(6).

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3187–3196, New York, NY, USA. Association for Computing Machinery.

Sumithra Velupillai, Gergö Hadlaczky, Genevieve M. Gorrell, Nomi Werbeloff, Dong Nguyen, Rashmi Patel, Daniel Leightley, Johnny Downs, Matthew Hotopf, and Rina Dutta. 2019. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Frontiers in Psychiatry*, 10.

Emilio Villa-Cueva, Ivan Gonzalez-Franco, Fernando Sanchez-Vega, and Adrian Pastor Lopez-Monroy. 2022. Nlp-cimat at politices 2022: Politibeto, a domain-adapted transformer for multi-class political author profiling. In *Iberian Languages Evaluation Forum*, volume 69.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2022. Fine-tuning a masked language model.

WHO World Health Organization. 2022. Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.

Liu Zihan, Xu Yan, Yu Tiezheng, Dai Wenliang, Ji Ziwei, Cahyawijaya Samuel, Madotto Andrea, and Fung Pascale. 2021. Crossner: Evaluating cross-domain named entity recognition. *AAAI Conference on Artificial Intelligence*, 35:13452–13460.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We add a limitation section after the conclusions.

☑ A2. Did you discuss any potential risks of your work?
In the Ethics Statement section after the conclusions.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Section 4 "Experimental Settings"

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 - subsection 4.2 "Model Configuration"

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 - subsection 4.2 "Model Configuration"

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5 "Evaluation and Discussion" and Section 6 "Analysis of the Models"

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3 and Section 4 - subsection 4.2 "Model Configuration"

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.