TEAM BIAS BUSTERS@LT-EDI: Detecting Signs of Depression with Generative Pretrained Transformers

Andrew Nedilko
Workhuman
agnedil@gmail.com

Abstract

This paper describes our methodology adopted to participate in the multi-class classification task under the auspices of the Third Workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI) in the Recent Advances in Natural Language Processing (RANLP) 2023 conference. The overall objective was to employ ML algorithms to detect signs of depression in English-language social media content, classifying each post into one of three categories: no depression, moderate depression, and severe depression. To accomplish this, we utilized novel generative pretrained transformers (GPTs), leveraging the full-scale OpenAI API. Our strategy incorporated prompt engineering for zero-shot and few-shot learning scenarios with ChatGPT and fine-tuning a GPT-3 model. The latter approach yielded the best results which allowed us to outperform our benchmark XGBoost classifier based on character-level features on the dev set and score a macro F1 score of 0.419 on the final blind test set.

1 Introduction and Related Works

From a common-sense linguistic perspective, detecting signs of depression in text can be challenging for a number of reasons:

- Variability of language and perception: a) different people may express their feelings differently, b) the same phrase might mean different things in different contexts, c) different people might interpret the same piece of writing in very different ways, d) the way people express emotions and discuss mental health can vary widely across different cultures.
- Privacy: some people may not be eager to openly express their depressive symptoms or feelings, using vague or metaphorical language.
- Absence of non-verbal cues: a lot of non-verbal information is lost in written text, such as tone of voice, facial expression, posture, etc.
- Co-occurrence of depression with other medical conditions which can have its own impact on text.

It should be also noted that text analysis can provide only hints, but should never be used as a definitive diagnostic tool. Only trained mental health professionals can diagnose depression.

Although discovering signs of depression in a written text is challenging because such text is not a direct indicator of someone's mental state, there are certain language patterns which might indicate a higher likelihood of depression.

According to Al-Mosaiwi and Johnstone (2018) and Al-Mosaiwi (2018) the following is typical of texts written by people with depression in the order of increasing importance:

- they use more words for negative emotions - a person dealing with depression often tends to have a more negative tone in their writing;
- depressed individuals often focus heavily on themselves, possibly due to feelings of isolation or self-blame; therefore, they use significantly more first person singular pronouns and significantly fewer second and third person pronouns. For the same reason, ruminations can be also observed in their writing when they repeat the same thought over and over again;
- depressed people use significantly more absolutist words - absolute magnitude or probability (50% greater in anxiety and depression forums and 80% greater in suicidal ideation forums).

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 138–143, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_020

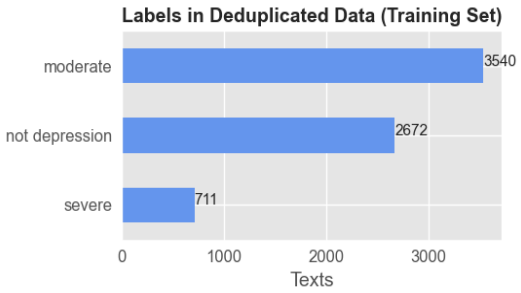

Figure 1: Distribution of Categories - Training Set

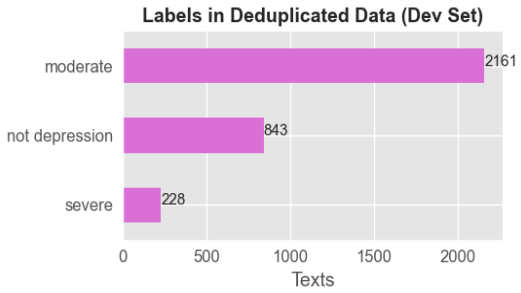

Figure 2: Distribution of Categories - Dev Set

The studies presented in Capecelatro et al. (2013) show that the depressed people prefer words related to sadness, death, avoid positive words. And the longer their depression is (5 years and more), the fewer appetitive or food-related words (eat, chew, drink, hunger) or sexual words (arousal, make out, orgasm) they use. The authors claim that this can be due to long-term changes in their brains.

Similar ideas are repeated in Newell et al. (2018) and Davis (2020). We attempted to quantify these characteristics in the form of counts of phrases that belong to each of these classes and use them as features for machine learning (ML) models. See more details in subsection 3.2 and subsection 4.2.

In addition, Havigerová et al. (2019) states the importance of early detection of the signs of depression. The goal of their research was to study automatic analysis of texts to build predictive models that can identify individuals at risk of a mental disorder. The authors came up with four regression models to predict a higher emotional state of depression using such text features as the ratio of pronouns to nouns, ratio of verbs to nouns (readiness for action), ratio of finite verbs to number of sentences, and ratio of the number of punctuation marks to the number of sentences.

2 Dataset and Task

The dataset consists of social media posts in which people describe their emotions and feelings. The number of examples in each subset is as follows: training set – 7201, development (dev) set – 3245, test set – 499. Given these posts, our task was to classify the signs of depression into three categories: no depression, moderate depression, severe depression. As you can see from Fig. 1 and Fig. 2, the distribution of categories is imbalanced with the majority category being "moderate depression" and the minority category – "severe depression".

Based on the common understanding that the same person cannot be both depressed and not depressed, and that two different people cannot write the same relatively long post in social media, we considered this a multi-class, but not multi-label classification i. e. each text can have only one label.

In line with this, 232 complete duplicates were removed from the training set; complete here means that all the values in these rows in all columns were identical. In addition, there were 158 cases in the training set where the text of the post was the same, but the labels were different. Since the same person cannot be depressed and not depressed at the same time, we decided to remove such cases because the true label was unknown (there were at least two different labels in each case), and we didn't feel to be qualified enough to decide which category each mislabeled text should belong to. There were only complete 23 duplicates in the dev set.

As for the data leakage – there were only three posts that occurred both in the training and dev sets. The test set didn't have any overlap with the training or dev set.

Fig. 3 and Fig. 4 demonstrate that the character length distribution across all datasets reveals a minority of abnormally lengthy texts. The majority, representing the 95th percentile, encompasses texts containing fewer than 2,500 characters. However, a substantial surge in text length is observed beyond this point, extending to and exceeding 20,000 characters. Jumping ahead, we should say that training separate classifiers for different lengths of text didn't improve the aggregate results.

A blind test set without labels was used for testing the model that had the best performance on the dev set. Unlike the training and dev sets, the test set required heavy text cleaning as certain combinations of English characters and even single characters (mostly contractions at the sub-word level)

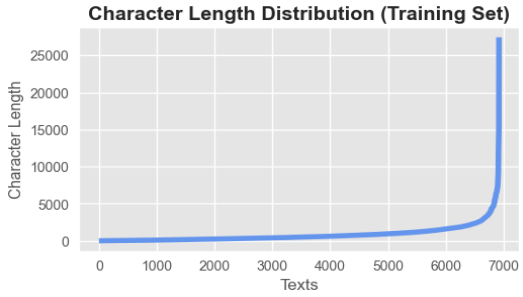

Figure 3: Character Length Distribution - Training Set

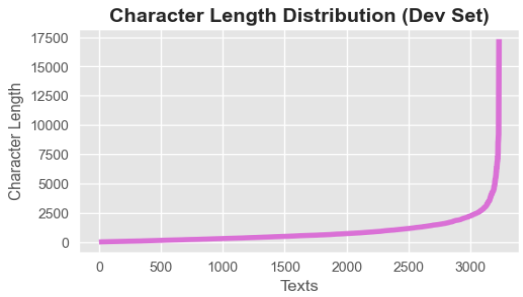

Figure 4: Character Length Distribution - Dev Set

were replaced with Chinese hieroglyphs. Examples: ” ’t ”, ” ’s ”, ” ’m ”, ” ’ve ”, ” ’d ” and many more.

3 System Description

3.1 Baseline Model

The baseline model, an XGBoost classifier employing word n-gram counts as features (utilizing CountVectorizer with an n-gram range of (1,3)), established the initial metrics. The initial macro F1 score was subpar, falling under 0.5. However, after implementing various enhancement techniques, including oversampling, data augmentation, and erroneous label elimination, we managed to elevate the final baseline macro F1 score to 0.54 and the micro F1 score to 0.61, as detailed in Table 1 below. See Fig. 5 for the baseline confusion matrix.

Oversampling was conducted up to the number of data points in the majority class. See subsection 4.2 below for a description of the data augmentation process.

3.2 Numerical Features

Using the sources of information from section 1, for each post we counted the number of occurrences of the following terms and tried to use this information as additional features to improve the classification results:

- words with a strong negative correlation in posts, e.g. self-harm, abominable, hopeless, disgraceful, etc.;
- words related to death;
- words and phrases representing absolutism, e.g. forever, never, no one, always, completely, etc.;
- first person pronouns singular: I, me, myself, mine, etc.;
- other personal pronouns: you, we, they, etc.;
- words related to the apetite and eating: bevarage, buffet, cravings, dining, etc.
- words related to sex;
- special medical terms describing specifically depression and medications for depression: delusional, exausting, mood swings, mental disorder, etc.

3.3 GPT: Iterative Prompt Engineering vs. Fine-Tuning

The effectiveness of transformer models and their ensembles for sequence classification has been validated by Kshirsagar et al. (2022), although with a macro F1 score remaining under 0.55. The recent rise of autoregressive models with the generative pretrained transformer (GPT) architecture and their remarkable ”human-level performance on diverse professional and academic benchmarks” OpenAI (2023) have been widely recognized. Thus, we deemed it pertinent to assess if the latest, novel GPT series models could offer a more efficient solution to the task of depression detection.

For this task, we leveraged a set of OpenAI models due to their comprehensive commercial APIs that offer diverse methods of interaction with pretrained models. Primarily, we employed the ChatGPT API with prompt engineering, generating various prompts to conduct extensive experiments on the development set, with an aim to optimize the macro F1 score. Both zero-shot and few-shot learning methodologies were applied. The training set was used for the sole purpose of concatenating examples for few-shot learning.

Zero-shot prompts asked the model to select the right category from a pre-defined list of categories (no depression, moderate depression, severe depression) for each example from the dev set or test set.

Few-shot prompts followed the same schema, but several labeled examples were appended to them so that the model could learn from such examples and make more accurate classification. The labeled examples were taken randomly from the training set.

Since these APIs didn't outperform our baseline model, we sought to enhance our metrics by fine-tuning a prior GPT-series model. Presently, neither ChatGPT nor GPT-4 offer fine-tuning capabilities. Only the original GPT-3 base models, which lack instruction following training and are smaller than ChatGPT, permit fine-tuning. We chose the largest such model – DaVinci, which led to surpassing our baseline model's score. We fine-tuned the model using the standard OpenAI API, without modifying the predefined hyperparameters. This API allows users to load the training set in a special format, fine-tune the model on this dataset, and then make calls to the fine-tuned model in order to classify new examples from the dev set or test set.

4 Analysis of Results

4.1 ChatGPT

We used zero-shot learning on the basis of the idea that the labels' names are self-descriptive and could be readily understood by a pre-trained model such as ChatGPT. We opted against employing GPT-4 for this experiment due to the lengthy nature of some texts and the multitude of examples in the development set. This decision was cost-driven, as GPT-4 API calls are significantly more expensive than those of ChatGPT.

Among all models, the zero-shot ChatGPT classifier demonstrated the poorest performance. Its highest macro F1 score reached was 0.25, significantly underperforming the baseline classifier (see Table 1). As illustrated by the confusion matrix in Fig. 6 the primary cause of this outcome was the classifier's tendency to excessively classify examples into the "severe depression" category.

To enhance the zero-shot classification results, we next explored few-shot learning. Given that the ChatGPT context window is confined to 4096 tokens, we could only select a finite number of labeled examples from the training set. These examples were randomly sampled for each development set data point to be classified. An alternate strategy could involve selecting the top n most similar training set examples based on a similarity score (e.g., using embeddings), but time constraints prevented us from testing this approach.

The results of the few-shot method were better than zero-shot – the macto F1 score reached 0.39, but you can see from Fig. 7 this method had a tendency to excessively classify examples into the "moderate depression" category.

Also, there are two apparent constraints of the few-shot learning method:

- **Size constraint**: The compact context window size precludes the usage of all examples from the training set in one prompt.
- **Cost constraint**: Being a commercial API, the more examples you utilize for each data point to be classified, the higher the cost.

4.2 Data Augmentation

We attempted to use non-textual features described in subsection 3.2. Due to limited time for this task, our first and quick attempt at using these features alone allowed us to achieve a macro F1 score of 0.43 (micro F1 score = 0.51). Nevertheless, the non-textual features did not provide any benefits when we combined them with the text features.

Two of the three categories in our dataset are underrepresented. To augment the minority classes, we performed data augmentation, adding 2800 new examples to the "severe depression" category and 1311 to the "no depression" category. For this, we deployed GPT-4, providing it with several training set examples from a specific category with similar lengths. The model was then instructed to generate approximately 25 more examples using semantically comparable language and within the same length of text.

We varied the ranges of text length for this exercise, selecting existing examples randomly. This method enabled a slight improvement in training our baseline model, though the uplift was marginal. The semantic similarity of the newly generated examples was validated by making sure their OpenAI embeddings stayed within a certain cosine similarity range when compared with existing examples.

Other types of augmented data that we tried to use as features included a title and a meaningful summary for each text generated by ChatGPT. However, these augmented data did not improve the final results either.

4.3 Improving Labels

After observing consistently low results in several experiments and noting that simple oversampling

Classifier	Macro F1	Micro F1
Baseline on text	0.5030	0.5696
Baseline on numeric feat.	0.4331	0.5127
Text + numeric features	0.4810	0.5628
Baseline on text, cleaned labels	0.5352	0.6094
Zero-shot ChatGPT, cleaned labels	0.2484	0.2560
Few-shot ChatGPT, cleaned labels	0.3885	0.5220
Fine-tuned GPT-3, cleaned labels	0.6018	0.6847

Table 1: Performance of Various Classifiers on Development Set

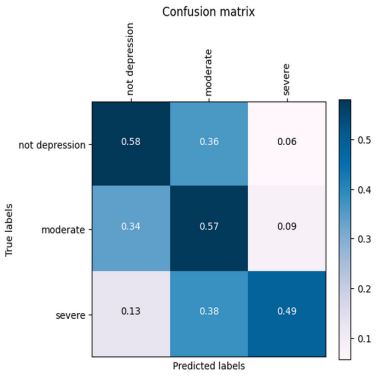

Figure 5: Confusion Matrix - Baseline Model

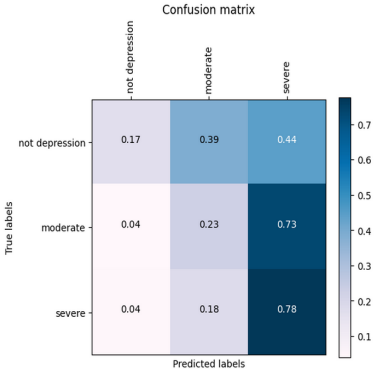

Figure 6: Confusion Matrix - Zero-Shot Learning with ChatGPT

yielded comparable low F1 scores even with data augmentation, we chose to investigate the dataset's annotation quality. As we lack expertise in clinical psychology or medicine, we refrained from verifying the "moderate depression" and "severe depression" labels. Instead, we scrutinized the "no depression" labels, searching for keywords such as "suicide" and its derivatives, "depress" and its derivatives, "harm myself", "anxiety", and so forth.

We identified approximately 450 texts in the training set and 165 texts in the development set that, to the best of our understanding and judgment, likely described some form of depression, as authors contemplated suicide or vividly discussed their depression. Several of these texts were so disheartening that we could not complete reading them. Training a baseline model without such data points, and testing it on the dev set that was pruned in a similar way, resulted in a 3% increase in the macro F1 score. Models in Table 1, trained without these data points, are designated as having "cleaned labels".

4.4 Model Comparison

The official competition metric for depression detection is the macro F1 score. Table 1 lists the macro and micro F1 scores for our models. All the scores in Table 1 are for the dev set. The best performing model shown in the last line of Table 1 scored 0.419 (macro F1) on the final blind test set. See Fig. 8 for the confusion matrix corresponding to the best model.

It is worth noting that the zero-shot learning method was outperformed by few-shot learning, but both of these methods scored below the baseline model. Fine-tuning a GPT-3 model demonstrated the best results on the dev set followed by the baseline model.

5 Conclusions

Our observations suggest that ChatGPT exhibits a degree of unpredictability, complicating the task of identifying a consistently effective configuration due to its dynamic nature. Hence, it is unsurprising that detecting depression using zero-shot and few-shot techniques proved challenging even for these cutting-edge models. In contrast, the largest fine-tunable OpenAI model, DaVinci, which is older and smaller than ChatGPT and lacks instruction following training, demonstrated superior efficiency for this task.

The fine-tuning capability addressed both few-shot learning constraints which we discussed in subsection 4.1. The model, while being fine-tuned, sees all the training set examples, and during inference, you are only charged for the tokens in the single example to be classified.

Also, if our doubts about the annotation quality

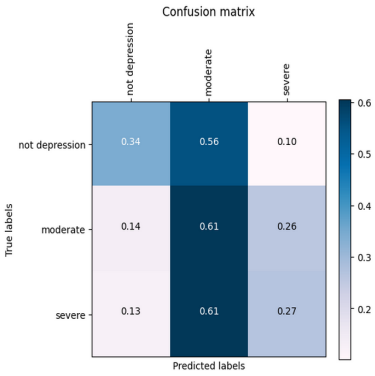

Figure 7: Confusion Matrix - Few-Shot Learning with ChatGPT

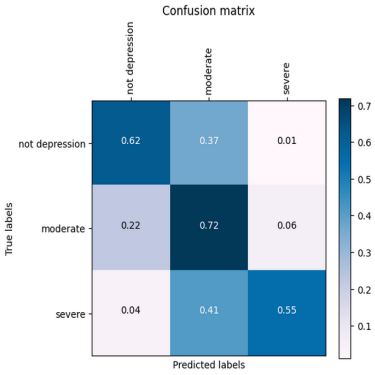

Figure 8: Confusion Matrix - GPT-3

are confirmed, then additional verification of the labels can significantly improve the classification results.

The exploration of non-textual features for depression detection warrants further study. Enhanced methods of aggregating numerical information from text could also contribute to improved classification outcomes.

References

Mohammed Al-Mosaiwi. 2018. People with depression use language differently – here's how to spot it.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6:529–542. PMID: 30886766.

MR Capecelatro, MD Sacchet, PF Hitchcock, SM Miller, and WB Britton. 2013. Major depression duration reduces appetitive word use: an elaborated verbal recall of emotional photographs. *Journal of psychiatric research*, 47:809–815.

Louisa Davis. 2020. How people with depression tend to speak differently. *The Mind's Journal*.

Jana M. Havigerová, Jirí Haviger, Dalibor Kucera, and Petra Hoffmannova. 2019. Text-based detection of the risk of depression. *Frontiers in Psychology*, 10.

S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and C Jerin Mahibha. 2022. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338. Association for Computational Linguistics.

Sampath Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, C Jerin Mahibha, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the second shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Atharva Kshirsagar, Shaily Desai, Aditi Sidnerlikar, Nikhil Khodake, and Manisha Marathe. 2022. Leveraging emotion-specific features to improve transformer performance for emotion classification. arXiv:2205.00283.

Ellen E. Newell, Shannon K. McCoy, Matthew L. Newman, Joseph D. Wellman, and Susan K. Gardner. 2018. You sound so down: Capturing depressed affect through depressed language. *Journal of Language and Social Psychology*, 37(4):451–474.

OpenAI. 2023. Gpt-4 technical report. arXiv:2303.08774.