



---

# Supervised and Zero-Shot Classification: Evaluating LLM Prompting Strategies for Mental Health Classification

---

Copenhagen Business School  
MSc in Business Administration and Data Science  
Natural Language Processing & Text analytics

*Project by:*

Mads Andersen - S167331

Oliver Beck - S137486

Benedicte Søvang Christiansen - S137322

Mikkel Pedersen - S167541

*Number of characters: 33,236*

*Number of pages: 15*

Submission date: 31-05-2024

# Contents

|   |           |
|---|-----------|
| <b>Introduction</b>                       | <b>2</b>  |
| <b>Related Work</b>                       | <b>2</b>  |
| <b>Data</b>                               | <b>3</b>  |
| Data Source                               | 3         |
| Data Description                          | 3         |
| Exploratory Data Analysis & Preprocessing | 3         |
| Data Splitting                            | 4         |
| <b>Methods</b>                            | <b>5</b>  |
| Evaluation Metrics                        | 5         |
| Macro Averaging Metrics                   | 6         |
| Models                                    | 6         |
| Training schemes                          | 6         |
| Multinomial Naive Bayes                   | 6         |
| Zero-Shot                                 | 6         |
| Few-Shot                                  | 7         |
| Zero-Shot Chain-of-Thought Prompting      | 7         |
| Output Processing                         | 8         |
| <b>Results</b>                            | <b>9</b>  |
| Confusion Matrices                        | 9         |
| <b>Discussion</b>                         | <b>10</b> |
| Data Quality                              | 10        |
| Naive Bayes versus LLMs                   | 12        |
| LLMs for Zero-Shot Classification         | 13        |
| Ethics                                    | 14        |
| <b>Conclusion</b>                         | <b>15</b> |
| <b>Bibliography</b>                       | <b>16</b> |
| <b>Appendix</b>                           | <b>18</b> |
| Appendix 1                                | 18        |
| Appendix 2                                | 20        |
| Few-shot classification                   | 20        |
| Appendix 3                                | 20        |
| Examples of mis-classification Fewshot    | 20        |
| Appendix 4                                | 21        |
| Appendix 5                                | 22        |

## Introduction

With the rise of the Covid-19 pandemic, a whole new level of concern for the invisible diseases was brought to life. According to WHO, the prevalence in anxiety and depression increased by 25% worldwide during the pandemic, putting pressure on mental health services and hindering care for those most in need. These gaps in mental health and life-saving services, such as suicide prevention, sometimes do not reach the vulnerable WHO (2022). Being able to identify different kinds of mental health issues in e.g. organizations can enable managers to take action where needed. The complexity of mental states can be hard to formulate through e.g. questionnaires and it can be difficult to ensure people's honesty which enables a valuable survey. Furthermore, it is not certain that individuals know whether they suffer from a mental diseases or not. When investigating text, however, these underlying structures can be more easily captured.

Two scenarios can be identified for organizations looking to utilize text analytics for mental health classification. One where an annotated data-set is readily available, and one where it is not. The first allows for utilizing a traditional supervised approach whereas the latter requires a transfer learning approach. Utilizing a conventional machine learning approach could possibly enable organizations with previously collected data to classify mental health issues inexpensively. Large language models (LLMs) such as Meta's Llama3 models Meta (2024) has shown to perform well in many sub text analytics tasks. LLMs rely on the general knowledge obtained from their training data, however the approach to utilizing this knowledge in classification tasks requires carefully consideration of the prompting techniques selected. These points motivates the following question:

**How do different prompting strategies influence the performance of LLMs as zero-shot classifiers in mental health tasks compared to traditional machine learning approaches?**

## Related Work

For the section on related work, a distinction between prior work concerning the overall topic and the methodological approach to the paper is made.

### Related to Topic

Prior research in the field of mental health disorders has been conducted by Olatunde et al. (2021), where they highlight the importance of considering mental health as equal to physical health, given numerous lives are lost due to mental illnesses worldwide. With their research of examining mental health status and detecting the risk of mental illnesses at an early stage, they developed a classification model for predicting mental health status and disorders for individuals, however focusing on those within the IT and Tech workplace for their research, and they identified the strongest factors in the determination of mental disorders, such as seeking clinical treatment or having past disorders. Olatunde et al. (2021) point

out that there exist many different mental disorders that come forth in various ways, but common for all/many is that they affect the individual's everyday life to a great extent, especially if they go untreated. In their paper, they suggest that this poses a possible mental health crisis and that it should be a public health priority. They likewise argue that the crisis could have severe implications for the global economy, by undeniably leading to a lack of productivity among others.

## Prior work on Zero-Shot learning with LLMs

Kojima et al. (2022) propose a chain of thought (CoT) prompting technique for zero- and few-shot learning, that helps the LLMs by splitting the task into step-by-step reasoning examples, as a replacement for the conventional question/answer examples used with LLMs. The approach is tested on logical datasets concerning arithmetic reasoning tasks. For the purposes of Zero-shot-CoT, two-stage prompting is utilized in which a prompt to extract reasoning of the text is applied first, followed by a prompt to extract the final answer. By using the CoT approach in comparison with the conventional approach as a baseline, the report finds that zero-shot-CoT outperforms a baseline Zero-shot approach in four out of six tasks. The two tasks in which it does not perform significantly better are simpler and do not require multi-step calculation, and therefore this result was expected prior to the tests. Wang et al. (2023) also investigates the use of LLM's for zero-shot learners to enable businesses to perform classifications even while a training data-set is not readily available.

## Data

This section of the paper describes from where the data has been accessed and retrieved as well as a description of the data's structure. Following will go into detail about how the data has been handled before the analysis.

### Data Source

The dataset used for this project has been accessed through kaggle.com Ghoshal (2023). It consists of text data scraped from different subreddits about mental health issues. The labels are therefore applied based on the subreddit from which each piece of text was collected.

### Data Description

The dataset contains 3 columns and 5957 rows. The first column "text" contains the text written for each observation, the second column "title" represents the titles for each piece of text, and in the final column "target", the labels are given from 0-4. Each label represents a mental health condition specified by a list (See Table 1).

## Exploratory Data Analysis & Preprocessing

For the EDA, the data was checked for NaN-values and it showed that only the text column contains 350 null observations. These null-values have been removed, leaving 5607 observations for the task. Then

the labels for each mental health issue are counted to check for class imbalance. As presented in Table 1, it is visible that the data is balanced with an evenly distributed number of observations for all labels. Therefore there is little to no risk of issues with class imbalance, even after removing observations. The first two rows for each mental health disorder have likewise been removed, as the corresponding text for these contains more of a welcome message and introduction to the mental health based conversations, rather than text written by users.

| Label | Condition            | Initial count | Final count |
|-------|----------------------|---------------|-------------|
| 0     | Stress               | 1181          | 1097        |
| 1     | Depression           | 1202          | 1200        |
| 2     | Bipolar Disorder     | 1185          | 1183        |
| 3     | Personality Disorder | 1201          | 1075        |
| 4     | Anxiety              | 1188          | 1142        |

Table 1: Observations of each mental health condition by label

The average length of the text instances is 836 characters which is around one-third of a normal page. This indicates that even though the text is written in a subreddit they are rather long as seen in Figure 4 in Appendix 5. An example of text written for each of the disorders is provided in the Appendix 4.

The title column has been disregarded as it does not yield any specific value for the purpose of this project. The columns kept for the purposes of this task are the *text* and *target* columns. Along with those, a column *label* has been created with corresponding mental health issues to each target value.

To create the baseline model, which is the multinomial naive bayes (MNB), the text data has been preprocessed further by defining a function that lowercases, removes punctuation and stopwords, while stemming and applying tokenization to the text.

Det er den finction i py der indeholder alle de trin der - Normalization i undervisingen

## Data Splitting

The data has been split into training and test sets, with a distribution of 85% for the training set and 15% for the test set. This leaves 840 observations for the test set, which should be sufficient. During the split, the data has been stratified, to ensure an even distribution of all labels in both training and test sets.

## Methods

This section of the paper covers the methods and models used for the project, along with the evaluation metrics utilized to assess their performances.

### Evaluation Metrics

The evaluation metrics used for the models are F1, Precision, Recall, and Accuracy, which are all macro averages of the individual scores. The metrics are calculated based upon True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP is where the model correctly predicts the specific class e.g. stress, whereas FP is the number of instances where the model predicts a class (e.g. stress) but the true label is a different class (e.g. depression).

#### Precision

Precision produces a single output indicating how many of the positive predictions made by the model were correct, and is calculated by [Géron \(2022\)](#) as follows:

$$Precision = \frac{TP}{TP + FP}$$

#### Recall

Recall is the number of TP from all positive samples in the dataset, and it is a measure for indicating the models' ability to detect positives. Recall is calculated by [Géron \(2022\)](#) as:

$$Recall = \frac{TP}{TP + FN}$$

#### F1

The F1 score calculates the number of times a model correctly predicted a sample through testing. The metric is calculated as the harmonic mean of precision and recall [Géron \(2022\)](#):

$$F1-Score = \frac{2Precision * Recall}{Precision + Recall}$$

#### Accuracy

The accuracy measure is used to obtain a single output metric that quantifies the classifier's performance. The accuracy formula is calculated by [Grandini et al. \(2020\)](#) as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

## Macro Averaging Metrics

Macro averaging provides a way to summarize these metrics across all classes. The macro avg. is a valuable performance metric for multi-label classification due to the way that it calculates the average of each metric across all classes.

$$Metric^{\text{macro}} = \frac{1}{N} \sum_{i=1}^N Metric_i$$

Where “Metric” is the metric to be macro averaged, e.g. F1, recall or precision.

## Models

Overall, two versions of the Llama3 model are utilized, the 8B and 70B parameter models. The models are decoder-only models trained on 15T tokens, with a vocabulary size of 128K tokens, with a context length of 8K tokens [Meta \(2024\)](#).

## Training schemes

This section encompasses the approaches used to classify text examples with a label. Three approaches for k-shot classification are tested, and an MNB model is trained for the purposes of comparison.

### Multinomial Naive Bayes

The MNB classifier builds upon bayes theorem, but assumes a multinomial distribution for the likelihood. The classification rule for a naive bayes classifier is given as in [Pedregosa et al. \(2011\)](#); [Kibriya et al. \(2005\)](#):

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y)$$

where  $\hat{y}$  is the predicted class, and  $P(y)$  is the estimated probability for class y. The likelihood is defined as:

$$P(x_i | y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Where  $N_{yi}$  describes the number of times word “i” appears for class y (in the training data), and  $N_y$  is the amount of samples with label y, lastly  $\alpha$  is a smoothing parameter, that diminishes the impact of likelihoods being zero.

The number of n-grams to include in the vocabulary and the  $\alpha$  smoothing parameter is found through a grid search with the bounds:  $n - \text{grams} \in [1, 5]$ , and  $\alpha \in [0, 0.5]$  with a sampling frequency of 0.05. The vocabulary size was: 12,716 unique tokens.

### Zero-Shot

Zero-shot learning refers to the approach in which a model is used directly on the test data without any example or demonstration from training. The model is provided with a natural language description of

the task to be completed and a prompt to define the specific task [Brown et al. \(2020\)](#), this prompt can be seen below.

#### Zero-Shot Prompt template

#### System Role

System Role: You are a mental illness classifier that can classify text into one of the following:  
Stress, Depression, Bipolar disorder, Personality disorder, Anxiety

Output format: Format prediction as >> X << where X is the prediction.

#### User

Now, classify the mental illness described in this text:

Text: "{input\_text}"

Mental Illness:

As indicated by the template above, the model is given instructions about its role and how to format its output. The text collected from the test set is reflected as 'input text' and the model will output the label for the text with >> 'label' <<.

#### Few-Shot

For few-shot (or k-shot) classification, a specified number of examples, defined by k, are given before predicting on the test example [Brown et al. \(2020\)](#). For this project, k has been set to five. The examples are sampled in a stratified fashion such that an example from each class is sampled at random from the training set. The prompt created for few-shot took a format similar to that of zero-shot and can be found in the appendix (Appendix 2). The few-shot model is provided with a few "examples" of the training data. This gives the model the possibility to include the examples into the context, for which it has to classify the example.

#### Zero-Shot Chain-of-Thought Prompting

Heavily inspired by the approach suggested in [Kojima et al. \(2022\)](#), a zero-shot-CoT approach is adopted, and slightly altered to filter the data and problem at hand. The structure of the prompting scheme is presented in Figure [1](#).



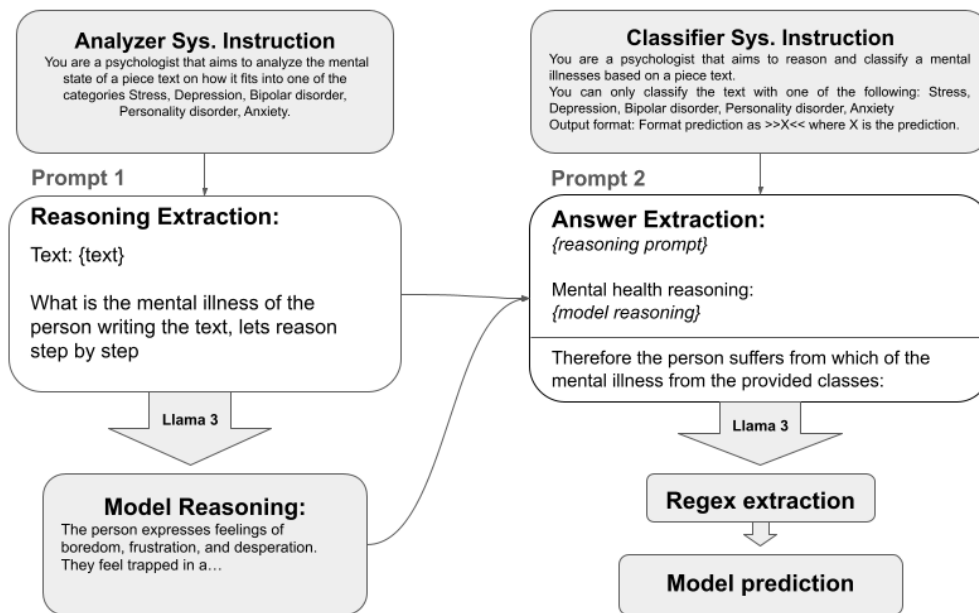


Figure 1: The prompting approach adopted from Kojima et al. (2022), adapted to reasoning about mental health for zero-shot classification.

The zero-shot prompting is divided into two prompts. One for logical reasoning, the other for extracting a label, and two different system instructions are utilized to give the model different roles. The first is for analyzing, and the other is for acting as a classifier.

## Output Processing

For finding the classification for the output, the model is instructed to output with the format presented in the previous section. This is inspired by Wang et al. (2023), where the output format is originally in json format. A format more easily identifiable by a regular expression has been chosen for this project. The regular expression pattern is defined as:

```
pattern = \> \> (.*)\< \<
```

The regex utilized for LLM outputs is a useful tool as it works as a generic representation of both singular or multiple strings. It is utilized to extract the predicted classification from the models' output and allows for only extracting the necessary part of the text without any additional characters [Kaur \(2014\)](#). It looks for the exact characters >> and <<, and between these, it is told to find any text (.\*), where \* refers to zero or more of the characters, and to only capture as little as possible (?). This is also known as non-greedy matching.

The inference made using prompting techniques was done with the freely available Groq API (Groq, 2024). This allowed for free and fast prompting of Llama3 model, however, this also meant that no

altering of the classification was available, as well as access to logits for tokens in the vocabulary. For that reason the following assumptions were made for the classification tasks:

1. The model is prompted to output one of the provided classes of the data. If it predicts close to one of the labels the data will be assigned to that label. Example: Social Anxiety will be treated as an anxiety prediction.
2. If the model predicts something completely outside the label space this prediction will be disregarded, but still be reported in the classification results. Example: Model predicts schizophrenia but the true label is Personality Disorder.

## Results

All models presented in the results have random states initialized with seed 42. The MNB was trained on  $n\_grams = (1, 2)$  alpha = 0.35. All LLMs have the temperature set to 0. The best results obtained by the LLMs have been marked with bold.

| Inference Scheme           | F1          | Precision   | Recall      | Accuracy    | $\hat{y} \notin Y$ |
|----------------------------|-------------|-------------|-------------|-------------|--------------------|
| Multinomial NB             | 0.78        | 0.79        | 0.78        | 0.78        | -                  |
| Zero-Shot                  | <b>0.59</b> | 0.73        | 0.58        | <b>0.59</b> | 27                 |
| Zero-Shot <sup>†</sup>     | 0.58        | <b>0.74</b> | 0.58        | <b>0.59</b> | 17                 |
| FS (K=5)                   | <b>0.59</b> | 0.65        | <b>0.59</b> | <b>0.59</b> | 11                 |
| Zero-shot-CoT              | 0.53        | 0.68        | 0.53        | 0.53        | 0                  |
| Zero-shot-CoT <sup>†</sup> | 0.56        | 0.72        | 0.57        | 0.58        | 28                 |

Table 2: Evaluation metrics of the presented techniques for learning mental illness classification. All of the metrics are macro averages of the scores on the individual labels.  $\hat{y} \notin Y$  refers to the predictions made by the model outside the defined label space. <sup>†</sup> refers to inference on the Llama-70B model.

As seen in Table 2, the best performing model is the MNB baseline which obtains an accuracy score of 0.78. Most of the LLM based classification show very similar performances across metrics, with F1 scores around 0.59.

## Confusion Matrices

The confusion matrix (cm) illustrated in Figure 2, shows high values along the diagonal. The highest classification rates appear when predicting anxiety but it is stress, and when predicting depression but it is bipolar disorder. Both have classification rates of 21.

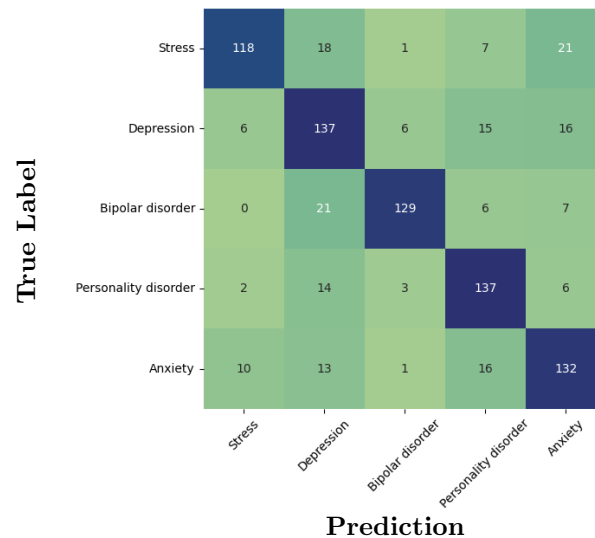


Figure 2: Confusion matrix for MNB.

The cm in Figure 3 for the zero-shot approaches shows higher levels of confusion compared to the MNB model. For the zero-shot classifier, the highest misclassification rate is found for predicting Anxiety but it is Personality disorder. For the the zero-shot-CoT classifier, it is for predicting anxiety but it is stress.

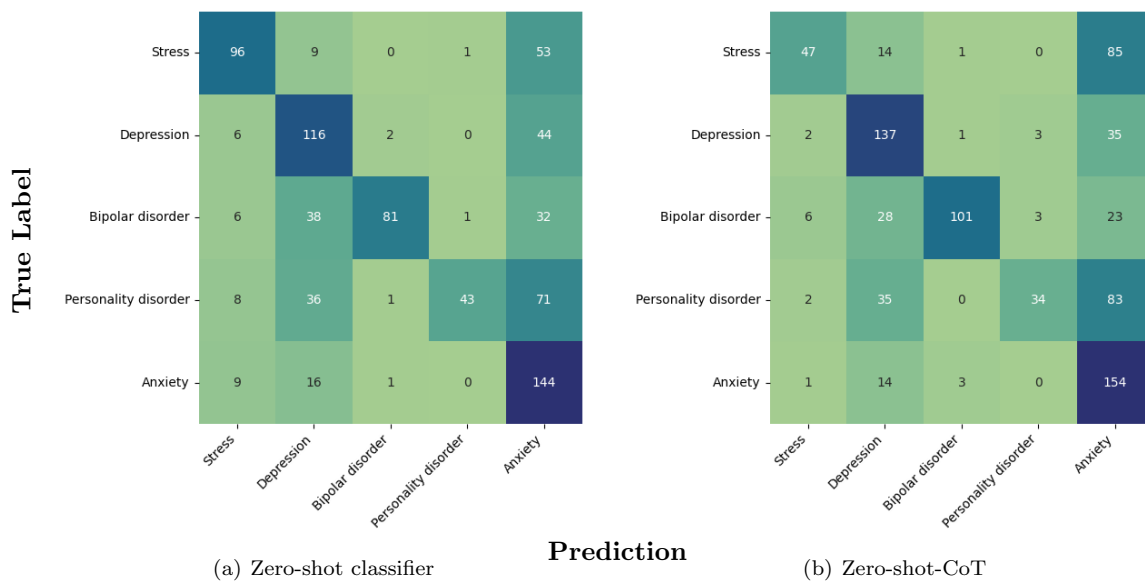


Figure 3: Confusion matrices of the zero-shot (a) and zero-shot-CoT (b) for the Llama-8B models.

## Discussion

### Data Quality

The data utilized for this project is scraped from various subreddits under the assumption that the individuals who wrote the text do indeed suffer from one of the five diagnoses. One of the main concerns of collecting data like this is that the data misses a clinical validation from field experts. An improvement

of the data could be the **validation by a psychiatrist**, which could either confirm or deny the diagnosis. That said, it would be fairly expensive for a psychiatrist to individually classify the text written by the individuals from Reddit. However, **applying domain knowledge to the dataset could arguably make the data more reliable**, since the labels in the current data might be the reasons why the models are e.g. classifying stress as anxiety or vice versa, as can also be seen in cm Figure **3**.

The way that the text might be worded or phrased can additionally influence how the model classifies. Stress, depression, and anxiety are disorders where the word in itself has become more common in everyday speech. **It is seen that people have started to use the terms more loosely about how they feel on a daily basis to describe e.g. a certain state of sadness or when feeling nervous about something, and not necessarily because they suffer from said disorders** **Xiao et al. (2023)**. It is important to acknowledge the distinction between perhaps feeling a type of anxiety or stress compared to actually suffering from the disorders. **It could be argued that this might be a reason why the model predicts these particular disorders more than the rest**. Moreover, some of the disorders are also connected as anxiety can be a response to stress, thus it can be hard to also distinguish these disorders. To exemplify, the following text can also be found in Appendix 4, *"I'm interested in CBT, but am too broke/anxious to see an actual therapist."*, is of someone that uses the word anxious to describe how he/she feels about something while suffering from personality disorder.

## Data Splits

**Dividing the dataset into 85% training data and 15% means that the number of test instances might not suffice for true estimates, but the choice to do so, reflects a limitation on time and computing power.** Having access to more computing power and less time restraint, the data could have been split to provide a larger test set without hurting the training of the model, as zero- and few-shot classification relies on a pre-trained LLM and not on training on the dataset (with the exception of the k number of examples for few-shot classification).

## Reconsidering the Problem

The focus of this paper has been to classify a piece of text as one mental health disorder. By doing so, the model only classifies the most likely disorder for each instance, but what if there were more disorders coupled to each instance? **This could be an individual who suffers from stress which leads to depression. The model will only predict the single class with the highest probability, which would be stress, thus neglecting that the individual also suffers from depression. This could help explain why all of the classifiers seems to struggle with classifying stress as anxiety as is evident in Figures **2** and **3**.**

The mentioned challenges likewise relate to the limitations of the models and how they are not meant for multi-label classification (e.g. an individual with more than one disorder). **Another aspect is that the data does not represent individuals with no disorder**. One could argue that it could have been beneficial

to include a class for when no mental health disorder is present, to prevent false positives in the binary case of a mental disorder or none. For the model to generalize to all real scenarios a class like the no mental health disorder should be included to handle instances where the individual is not mentally ill.

Being able to perform multi-class classification in which a threshold could be set, would necessitate further access to the Llama3 model to study its output probabilities. This way the model can be controlled in making predictions with a focus on thresholding to adjust between precision and recall. Say the preference is to detect all possible mental health disorders based on the text before the individual sees a psychiatrist, a high recall and a low threshold could be useful.

## Naive Bayes versus LLMs

The case of diagnosing mental illness on textual data is a rather difficult task and one could assume that the state of art ML should be used. However, MNB despite the simplicity of the model, can in this case and with this specific data outperform the bigger and more complex LLM. MNB assumes that the features are independent [Peng and Schuurmans (2003)]. The assumption of feature independence is often disrupted by real-world data, especially text. However, the created MNB model is still performing relatively well with an accuracy of 0.78 as seen in table 2. A strength of the MNB model is the low computational requirements, which makes the model efficient to use while still obtaining an overall satisfactory result.

On this particular dataset, the Llama3 is performing quite poorly relative to the MNB model as understood from the result in the table 2. The Llama3 pre-trained model is used with a zero-shot, and k=5 few-shot approach, and underperforms in both cases. The performance of the MNB model and the relative underperformance of the LLM could raise the question; in regard to performance and efficiency, is it necessary to utilize the complex LLM over a simpler machine learning model like the MNB? Despite the above arguments, it should be mentioned that the MNB has an advantage over the pre-trained LLM in the case of zero-shot-learning, due to the MNB's training, whereas the LLM is only used on the test data. However, the few-shot approach where the pre-trained model has seen some examples of the training data did not obtain any further advantage over the zero-shot, still underperforms the MNB as seen in table 2.

While the non-deep learning approach is underperforming in this case, [Kamath et al. (2018)] finds that DL models are usually outperforming traditional ML models on text data when trained on the same data. For this reason, the results presented in Table 2, do not discredit the use-case of DL models for text classification. It does however show that the pre-trained models cannot compete with even simple models fitted on a training set. A reason for the DL models' excellence in text classification, is due to their ability to model dependencies between words. This ability is imitated for the MNB classifier by including n-grams. This is done by combining words, e.g. bigrams and trigrams such that the model might catch things such as "feel anxious" instead of "feel" and "anxious" from the original preprocessed text. However, fitting the model with n-grams increases the computational requirements due to the

number of features being dramatically increased. Another con of adding n-grams is the risk of the model losing the ability to generalize to new data because the model has learned specific phrases in the training data [Peng and Schuurmans (2003)].

## LLMs for Zero-Shot Classification

As described by [Kojima et al. (2022)] zero- and few-shot classification has been identified to work well for especially models of +100B parameters in size. However, [Meta (2024)] presented superior results for the smaller models to perform on zero-shot tasks. For this reason, it did seem plausible that the lower parameter models could be sufficient, for achieving good results. As evident from the results in Table 2, zero-shot learning with the Llama3-8B model did yield decent results with an F1 score of 0.59. The same experiment was conducted with the 70B model, however this did not seem to make any improvements with an F1 of 0.58. This is consistent with the findings of [Kojima et al. (2022)], where the scale of models was not found to increase the ability for zero-shot learning. Given that [Rae et al. (2021)] finds performance in humanity tasks increases with model scale, it was speculated that this could transfer to increased performance in mental health classification. This was however not found to be the case, and so, considerations towards the quality of the presented data should be kept in mind. Beyond this, [Rae et al. (2021)] compares a < 7.1B parameter model to a 280B parameter model. For this research, the comparison was between an 8B parameter model and a 70B parameter model, and it could be speculated that this increase in parameters is not sufficient to see increases in performance for the task of mental health classification.

When imposing the zero-shot-CoT prompting scheme, some effect of using a larger model was detected, which can be seen in Table 2. Here, scaling the model to 70B parameters from 8B, yielded an increase from 0.53 to 0.56 in the F1-score, and increases in both precision and recall. This also fosters the speculation as to whether the performance could be increased even further when utilizing a zero-shot-CoT technique with an even larger model. Here, [Kojima et al. (2022)] finds that zero-shot vs. zero-shot-CoT shows very enhanced zero-shot-CoT capabilities when going from model size L to XL for the Instruct GPT3 and from 62B parameters to 540B parameters using the PaLM model [Chowdhery et al. (2022)]. For this reason, it is speculated that zero-shot-CoT could outperform the regular zero-shot classification technique on the mental health dataset if model scale is increased sufficiently. While it was beyond the scope and resources of this research to establish if there was significant difference (*e.g. using a t-test*), it is most likely not the case. While studies such as [Kojima et al. (2022)], find that providing examples yields better results, the nature of the task and the dataset might not make for very good examples for reasoning. This is since the examples provided are only providing a text and a label with no further elaboration as to why and how the result was reached. This approach is contrary to the logical reasoning tasks exposed to the LLMs by [Kojima et al. (2022)], where the thought pattern that leads to the result is included in the prompt. This could very likely be the reason for the K=5 shot learning not providing any gains in performance over the zero-shot methods.

## Towards Deterministic LLMs for Mental Health Classification

Upon initialization of the Llama 3 model, the temperature was set to zero, aligned with the approach of Wang et al. (2023) where temperature is set to 0.01 for zero-shot classification. This was done in order to provide more deterministic results, as it can be argued that creativity and variance in the output is not desirable when predicting on health data in general. The effect of temperature on the model outputs is discussed by Touvron et al. (2023) where a higher temperature is associated with heightened creativity and a lower temperature makes models perform better in factual tasks. Initial trials also showcased that setting the temperature to zero displays a better ability to predict within the label space, thus diminishing the number of labels that will have to be discarded. The temperatures effect on the ability to reason about mental health and for classification could be subject for further research.

## Design of Prompt Templates

As shown by Kojima et al. (2022), selecting the correct formulation for creating prompts can have serious effects upon the results of LLMs. Upon settling for the design of the prompt templates, several considerations mainly inspired by Kojima et al. (2022) were incorporated. First off, when creating the prompting structure as presented in Figure 1, it aims to stay as true as possible to the one proposed by Kojima et al. (2022). Here, the phrase: *“Let’s think step by step”* was included due to the performance effects found by incorporating this exact formulation. In order to make comparisons as fair as possible between all of the prompting techniques, the formulations were also aimed to be as comparable as possible. This was done to minimize the amount of noise that could potentially skew the models to a worse or better performance independent of the prompting technique selected. An example of this is the prompting templates provided for zero-shot and few-shot classification.

## Ethics

Ethics is an especially important aspect to consider when dealing with a sensitive topic like mental health. Whilst research on the topic of mental health is essential, it does come with ethical complexities. Some ethical issues and considerations for this project include the privacy of the individuals as well as the potential impact on said individuals.

It is crucial to protect users’ privacy and anonymize data to ensure that any personal information is not revealed without informed consent. Regarding the impact it can have on individuals, making sure that the model is accurate is essential to prevent misclassifications, as this can lead to potential harm. Misclassifying someone with a mental health disease that they do not have can possibly cause stress in itself. Hence, the mental and emotional well-being has to be considered of those individuals whose data is used and analyzed. Furthermore, it is possible that this is a rather sensitive subject for some, thus being cautious and considerate is crucial, to not cause further harm. However, this study aims to provide transparency for the model limitations, such as providing the number of classifications outside the label space, but also comparing the LLMs to baselines and exploring non-SOTA techniques to provide a

balanced overview of the capabilities. Beyond this, the approach with LLMs and especially the technique of using zero-shot-CoT can provide the patients with an explainable reasoning for the classification in conformity with the EU ethics guidelines for ethical AI [EU commission](#) (2019).

Another aspect to consider when using artificial intelligence, such as machine learning and deep learning concerns the computational requirements and the amount of energy consumed when running such models. The deep learning approach is generally more computationally heavy and in most cases has a longer computing time compared to relatively simple models like MNB. Thus, one could argue that the favorable model would be the MNB, yet the deep learning approach might be more suitable for general application or if the complexity of the task was higher.

## Conclusion

This paper set out to analyze how the use of different prompting techniques in LLMs for zero-shot and few-shot classification can be utilized in assessing mental health, and how these approaches compare to a traditional machine learning model.

The application of different prompting techniques shows that the use of traditional zero-shot and few-shot prompting slightly outperforms CoT-prompting, despite prior research by [Kojima et al.](#) (2022) reaching the opposite conclusion, of the CoT outperforming zero-shot prompts. In addition, the zero-shot-CoT method shows increases in performance when scaling from 8B (F1 of 0.53) to 70B (F1 of 0.56), meaning this approach has the possibility of improving with scale, and can be a point for further research.

Based on the findings of the report, it can be concluded that none of the LLM classifiers perform as well as the MNB baseline model when tested. The best LMM approach obtained a F1-score of 0.59, compared to the MNB of 0.78. This showcases that the traditional approach of establishing an annotated dataset is preferable when considering overall performance. All of these results should be considered in the light of the discussed data quality, which was also identified to contain several underlying problems.

In the light of the findings, the overall recommendation for organizations looking to implement text classification for mental health analysis is to consider the availability of data and resources. If performance and accuracy are of high importance the traditional approach is desirable. The use-case for LLMs would be under circumstances where a low cost, low effort approach is needed and accuracy is less crucial.



## Bibliography

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). Palm: Scaling language modeling with pathways.
- EU commision, I. H.-L. E. G. O. A. I. (2019). Ethics guidelines for trustworthy ai.
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."
- Ghoshal, N. (2023). Reddit mental health data. *kaggle*.
- Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.
- Groq.com (2024). Groq api. <https://groq.com/>. Accessed Between: 01-05-2024 and 28-05-2024.
- Kamath, C. N., Bukhari, S. S., and Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–11.
- Kaur, G. (2014). Usage of regular expressions in nlp. *International Journal of Research in Engineering and Technology IJERT*, 3(01):7.
- Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. (2005). Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pages 488–499. Springer.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Meta (2024). Introducing meta llama 3: The most capable openly available llm to date.

- Olatunde, O., Tipu, A., and Falola, B. (2021). Classification of mental health disorders. *Department of System Science and Industrial Engineering. State University of New York, Binghamton*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Peng, F. and Schuurmans, D. (2003). Combining naive bayes and n-gram language models for text classification. In *European Conference on Information Retrieval*, pages 335–350. Springer.
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, Z., Pang, Y., and Lin, Y. (2023). Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.
- WHO (2022). Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. *World Health Organization*.
- Xiao, Y., Baes, N., Vylomova, E., and Haslam, N. (2023). Have the concepts of ‘anxiety’ and ‘depression’ been normalized or pathologized? a corpus study of historical semantic change. *Plos one*, 18(6):e0288027.

## Appendix

### Appendix 1

The box below displays a COT prompt from the Answer extraction phase.

**COT prompt**

Text: I fear for how boring my future life is. I see how my life is planned out and I hate it. I see myself going to college and then graduate school, then meeting my husband through my parents, marrying someone who is part of my culture, having 2+ kids, and raising them in the suburbs near my parents, constantly working to make money so I can give my kids a nice life and taking my salary and putting it into a college fund, just so this life can be replicated by my kids. It's boring and I hate it. I want more. I want to be famous and make something of myself and do something. I don't want to be boring. The only problem is I only care about success and I only want to do something I love if I will be successful because if I'm not I feel I'm still living the same boring life. I need help because this is keeping me up all night of everyday and I think about it almost every second except when I eat, sleep, and watch TV.

What is the mental illness of the person writing the text, lets reason step by step

Mental health reasoning:

Based on the text, I will analyze the mental state of the person writing it and categorize it into one of the categories: Stress, Depression, Bipolar disorder, Personality disorder, or Anxiety.

Step 1: Identify the dominant emotions and thoughts

The person expresses feelings of boredom, frustration, and desperation. They feel trapped in a predetermined life plan and desire more excitement and success. They also struggle with the idea of settling for a life that is not fulfilling.

Step 2: Assess the level of distress

The person reports being kept awake at night, thinking about their situation almost constantly, except during meals, sleep, and TV time. This suggests a high level of distress and anxiety.

Step 3: Evaluate the presence of negative thought patterns

The person's thoughts are dominated by negative self-talk, self-doubt, and a fear of failure. They feel trapped and desperate for a change.

Step 4: Consider the presence of avoidance behaviors

The person does not explicitly mention avoidance behaviors, but their desire for success and fear of failure suggest that they may be avoiding certain situations or opportunities due to fear of failure.

Step 5: Assess the presence of emotional dysregulation

The person's emotions are intense and overwhelming, causing them to be kept awake at night and constantly preoccupied with their situation. This suggests a lack of emotional regulation.

Step 6: Evaluate the presence of cognitive distortions

The person's thoughts are dominated by all-or-nothing thinking ("I'll only do something I love if I'll be successful"), which is a common cognitive distortion.

Step 7: Consider the presence of interpersonal difficulties

The person does not explicitly mention interpersonal difficulties, but their desire for fame and success suggests that they may be struggling with feelings of inadequacy or low self-esteem.

Conclusion:

Based on the analysis, I would categorize the person's mental state as:

Anxiety Disorder (specifically, Generalized Anxiety Disorder): The person's constant worrying, fear of failure, and intense emotional distress are characteristic of an anxiety disorder. The presence of negative thought patterns, avoidance behaviors, and emotional dysregulation also support this diagnosis. While the person's thoughts and feelings may also be influenced by underlying personality traits or depressive symptoms, the dominant feature of their mental state is anxiety.

Classify the text with one of the provided classes:

## Appendix 2

### Few-shot classification

#### Few shot Prompt template

#### System Role

System Role: You are a mental illness classifier that can classify text into one of the following:

Stress, Depression, Bipolar disorder, Personality disorder, Anxiety

Output format: Format prediction as >> X << where X is the prediction.

#### User Given these examples:

{examples}

Now, classify the mental illness described in this text:

Text: "{input\_text}"

Mental Illness:

## Appendix 3

### Examples of mis-classification Fewshot

Examples of misclassification by k=5 shot learner:

| Text   | $\hat{y}$        | $y$                  |
|--|------------------|----------------------|
| <b>Misclassification</b>   |                  |                      |
| "I was going to go to a meet up group that eats different food around the city and then I canceled. I was so excited and hopeful about making friends, but then I realized how since I got divorced I look so much older and so much chubbier.... I felt I would be able to sense by their reaction to me that I am not good enough and I don't belong. So I canceled. I feel sad about it. Also I never have money and I had a little extra cash so this is one of the few times I was going to be able to afford this group. I just don't have anyone to talk to about this and I feel sad and silly." | Stress           | Personality disorder |
| <b>Prediction outside label space</b>  |                  |                      |
| "i used to be cool and popular and girls would even msg me out of the blue on fb, 10 years later...i have no family, 0 friends, 0 relationships and can't hold a job. im gonna be 24 this year for christ sake and i can't do anything. i want to end my life but im scared of the pain, i wish i had a gun or atleast a for sure method.im not scared of dying im scared of the pain moments before"  | Suicide Ideation | Anxiety              |

Table 3: Examples of a misclassifications by the k=5 shot classifier and an example of the k=5 shot classifier predicting outside the defined label space.

## Appendix 4

### Class 0 - stress

'I'm stressed about my living situation. It's hard. Being in a new city is harder. I'm not a self motivator.'

### Class 1 - depression

"Was feeling very low , not feeling like working and sleeping most of the time , feeling worth less and very lonely... wasn't able to deal with it. i went for a Tantra massage for an hour , felt touched, someone is making effort to make me happy... Touch of a skin on my body was triggering happy hormones... Finally i was on receiving end , getting naked in front of someone is kinda of excepting who and what you're, no body shaming.'s not an optimal or ethical solution but for time being... I am feeling happy and high spirit. want to know/ check with my fallow group members... Have you been to this situation any time and what you did ? that a right or wrong move? haven't done it in past . No kind of sex was involved in it or any orgasms/ ejaculation."

### Class 2 - bipolar disorder

'Last night as I laid down to sleep and had my eyes closed, I had something that felt like a dream, but it was something that happened just before I feel asleep. I lucidly thinking about the experience as it was happening.'ll start off with what I think it was. I think it was a new kind of suicidal ideation.'m not religious at all, but this dreamlike feeling was full of images that felt like a type of beautiful afterlife, like a heaven or something. It was sweeping land and seascapes that looked like something off of an Asia album cover. Everything was iridescent. And I was flying high above it all loftily and effortlessly. It felt like the scenes from the Lost Boys when you see the vampires' POV as they fly. looked so real, and so perfect and peaceful, and I felt like it was calling me to it. I began to speak to it. I said, "please.. if this is it, please take me." "Please take me away from this." if I have SI, it's miserable and darkly depressing. felt like the SI was being cunning and alluring this time.'

### Class 3 - personality disorder

"I'm interested in CBT, but am too broke/anxious to see an actual therapist."

### Class 4 - anxiety

"I always crack my neck and have been doing it for years, but now i read that multiple young people below 28 died of strokes because they tore a vein when they cracked their neck :( just cracked it before i read that and now i have a really bad headache on the right side and in the articles they say it took a week for the actual stroke to happen after the artery got damaged, i am so scared of strokes literally the first sign of my anxiety disorder were stroke symptoms that i went to the hospital with hate it i wish i never read about this, now i feel everything in my neck and i keep checking for blue or red spots in the mirror, and i keep touching the veins to feel if there is a bump or anything. i don't want to die like this i only cracked my neck, people do that all the time right?"

## Appendix 5

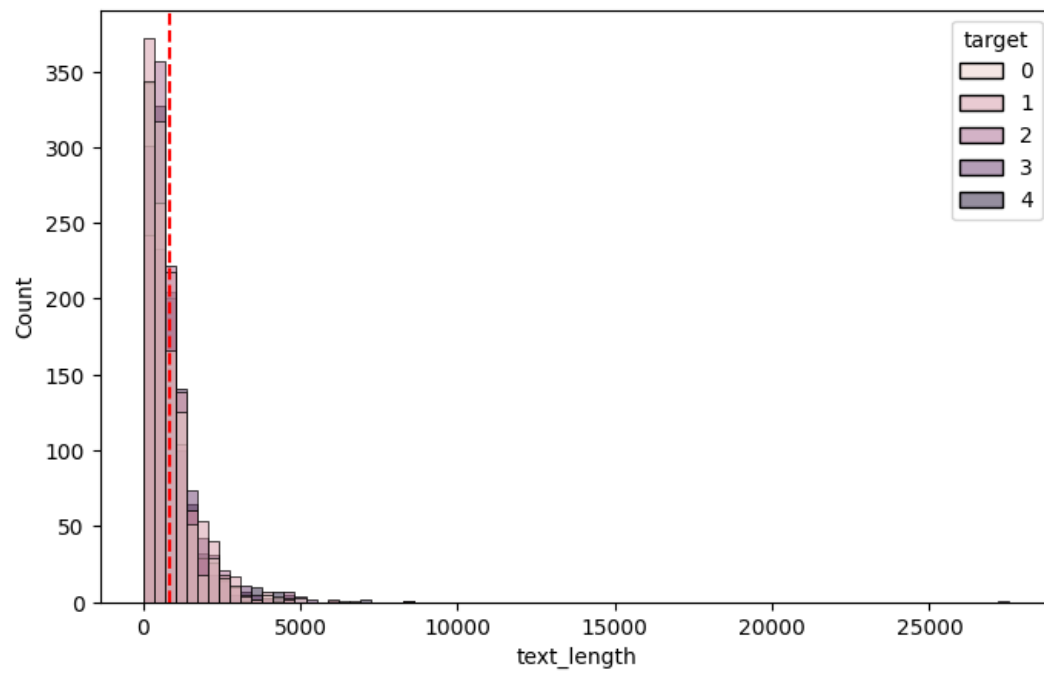


Figure 4: Text length with mean