FeelingBot: The Development of a Sentiment-Classifier Discord Application

Final Project for INNOVATE 1Z03 2021 Summer Semester

John Popovici
Software Engineering and Management Co-op
McMaster University
Hamilton, Ontario
popovj3@mcmaster.ca
400265718

Maha Chaudhry
Chemical Engineering and Management Co-op
McMaster University
Hamilton, Ontario
chaudm22@mcmaster.ca
400252855

Abstract – **In this report we focus on how depression can manifest itself within social media, and how we may implement an AI bot to detect such tendencies in discord servers. We implement a rudimentary version of this system, called FeelingBot, to detect and notify discord users of the sentiment of their comments to aid them in their own decision making of how they act and in self-diagnosis.**

Keywords –
AI – Artificial Intelligence
NLP – Natural Language Processing
GPT-3 – Generative Pre-trained Transformer 3
AV – Autonomous Vehicles
API – Application Programming Interface

I. INTRODUCTION

Throughout this course, we learned a range of concepts relating to the field of artificial intelligence (AI). AI is branch of computer science focusing on developing algorithmic systems that can mimic human intelligence. The motivation with the AI field is to reduce the need for human experts, increase automation in more complex field, and allow for the increased personalization of products and services. The rapid development and integration of AI is marked by the beginning of the fourth industrial revolution, where the first, second, and third industrial revolutions were marked by the introduction of mechanical equipment, mass manufacturing and assembly lines, and electronics and robots in production automation respectively. The field of AI is currently limited to narrow AI, specifically machine learning, which can only assist with specific tasks. Current systems combine large datasets with iterative algorithms to create predictive models. The most recent development in the branch of machine learning, neural networks, teaches algorithms to process input values through a weighted sum of layers. While limited, they allow for innovative smart products that employ IoT to present dynamic services.

The course focused on two main fields for the application of AI: autonomous vehicles (AV) and healthcare diagnosis. AVs are designed using a system of sensors, GPS, and machine vision. Through automatization, researchers hope to increase road safety, increase road capacity, increase mobility for those without personal vehicles, and reduce pollution by reducing congestion. Like many other AI applications, it holds cybersecurity concerns, with risk of system failure, as well as reduced employment risk by outmoding conventional ride-sharing businesses. Application in healthcare, on the other hand, aim to aid in high-risk environments, like surgery, to reduce human error. Healthcare, unlike ride sharing, has a shortage, increasing pressure on human personnel, and thus reducing the quality of care. Unfortunately, the complex nature of the human experience means that AI struggles to integrate in an effective manner. AI currently cannot explain their reasoning, making them highly untrustworthy to medical personnel and patients. Additionally, current technology is owned and run by third party companies, introducing privacy concerns over data usage and regulation. Overall, the ethical nature of AI is complex, with some applications leaving consumers at increased privacy risks and employment crises and others bringing analytics to substitute the absence of required human personnel.

To successfully deal with these concerns, governments and corporations will need to develop a better understanding of the ethical hierarchy of human needs and learn how to balance maintaining human rights with improving the human experience.

Within this paper, we expand on the concept of AI reducing pressure on the healthcare system through the lens of mental healthcare, specifically depression. We expand on topics covered in lectures and tutorials, like the current state of technology, the concept of natural language processing, the dangers of data misinterpretation, as well as the future challenges of AI.

II. WHAT IS DEPRESSION

Depression is a mood disorder that causes constant feelings of sadness and loss of interest in activities [1]. Symptoms, ranging from mild to severe, can include [1, 2]:

- Persistent feelings of sadness
- Increased pessimism
- Increased irritability or frustration
- Loss of interest in activities once enjoyed
- Changes in appetite
- Increased fatigue
- Feelings of worthlessness or guilt
- Difficulty concentrating
- Difficulty sleeping or oversleeping
- Persistent pains without clear physical
- Suicidal thoughts

Two of the most common forms of depression include major depression and persistent depressive disorder [2]. The first comprises of severe symptoms for at least 2 weeks while the latter is long term, consisting of less severe symptoms for at least 2 years [2].

According to Canadian Mental Health Association, about 8% of adults will experience major depression at some point in their lives [4]. Treatments for depression include psychotherapy, known as talk therapy or counseling, and medication like antidepressants [1]. Notably, studies show that between 80 to 90% of those with depression respond well to treatment [1]. Without treatment, depression can result in drug or alcohol addiction, ruin relationships, and make it difficult to overcome serious physical illnesses [3].

III. SOCIAL MEDIA AS AN INDICATOR OF DEPRESSION

Social media has become one of the most popular forms of communication. People increasingly use social media sites to keep in touch with their friends, share photos and videos showcasing their lives, and to share their opinions on current events. By 2018, 78% of Canadians were regularly using at least one social media account, including 90% of those 15 to 34, 80% of those 35 to 49, and 60% of those 50 to 64 [6].

With a daily average of 2 hours and 22 minutes spent on social media posting and commenting in a somewhat naturalistic setting, social media provides a large, diverse range of data that can be used to measure an individual's behavioral characteristics and emotional state [5]. Common characteristics of depression, like self-hatred, sorrow, guilt, and despondency, could be determined in the language used [5]. Less commonly known symptoms like irritability, hostility, and rejection sensitivity could also be helpful indicators [7].

Notably, the complicated nature of mental disorders means it can present in different individuals in different manners [5]. This can make detection of depression within social media a highly complex problem. For some, it can present primarily with more feelings of sadness, others with increased irritability and sensitivity, and for some, it presents more so with physical symptoms. Furthermore, those who are more careful about how people perceive them might be more careful against posting negative commentary and would thus be more difficult to detect [14].

IV. MACHINE LEARNING IN THE DETECTION OF DEPRESSION

Machine learning is one the largest growing fields in modern computing. It consists of developing algorithmic models based on sample data to make analysis on new data without being explicitly programmed to do so. Developments into Natural Language Processing (NLP), a branch of machine learning that focuses on developing models that can extract and analyze the contextual information of speech and textual language, has played a fundamental role in determining the linguistic patterns between depression and an individual's social media data [5]. A study by Tadesse et al., for example, used NLP to identify a series of terms commonly used by depressed profiles on Reddit [8]. They used a training dataset consisting of 1293 depressive posts, collected from depressed users seeking online support, and 548 standard posts by non-depressed users [8]. Through a series of NLP tools, they generated a model that could identify depressed users with a 91% accuracy [8]. Figure I below provides an excerpt of those terms. The study showed that depressed users used more emotionally negative-indicative words [8]. Users felt rejected and lonely, and many posts reflected a low self-esteem or hostility towards their current relationships [8]. Depressed posts were also more self-preoccupied, containing more first-person pronouns like "me" and "I" [8].

Depression-Indicative Posts	Standard Posts
alone, break, blame, depressed,	awesome, aunts, believe, beautiful,
deserve better, deserve unhappy,	close, advice, cooking, cousins,
die, escape, distraction, nobody,	don't care, encourage,
feel alone, feel depressed,	family, logical person, got married,
felt pain, fuck don't, hate,	I do, better, mom, peace, parents,
hurt, loneliness, mine, myself,	spend time, new friends, right,
reject love, safe, shit, sucks, no job,	funny, need, thankfully, uncles,
painful, pressure, too worried,	soul-friends, work, weekend,
unsuccessful, ugly, uncomfortable,	movie, potential, texted me,
winter, worry, worth, wrong life	too good

Fig. I. Terms in depressive-indicative vs. standard posts. Reproduced from [8].

An interesting feature of Tadesse et al.'s approach is the use of LIWC for feature extraction and text classification. The LIWC (Linguistic Inquiry and Word Count) dictionary is a text analysis tool that identifies words associated within specific psychological categories [9]. For example, if it identifies 100 positive words in a speech with 5000 words, it would be classified as 2% positive [8]. It is considered one of the most common tools in computational linguistics for psycholinguistic analysis [8]. For this particular study, researchers had to extract features that were prevalent in depression.

Another study by Orabi et. al. identified the challenge of text classification on social media sites, especially Twitter. While misspelled words and grammatical errors are expected in informal communication, sites like Twitter which have character or word limitations can exacerbate that. Limits push users to unstructured their texts, intentionally leave out words, or abbreviate [10]. To process unstructured data, the researchers implemented a character n-gram model, a language model that predicts word sequences based on the previous word history [11,12] Specifically, they used a Word2vec algorithm, which employs a generalization of an n-gram called a skip-gram. Word2vec is a neural network model that uses word associations to suggest additional words for incomplete sentences through a process called word embedding [11,13]. Overall, the models generated resulted in 87~% accuracy rates.

An additional challenge relating to data excess was identified by Eichstaedt et al. Researchers created a depression prediction model based on the Facebook history of 683 patients visiting the ER, where 114 had a diagnosis of depression [14]. They discovered when data was restricted to 6 months before their diagnosis, the accuracy of the model increased from AUC = 0.69 (approximately equal to a screening survey accuracy rate) to AUC = 0.72 [14]. This is likely because a patient's depression developed in a time well after their initial Facebook profile had been set up. Therefore, depressive-indicative posts, while being more recent, did not make up as large a portion of total user data as standard posts. It implies that having time restrictions on collected data might be necessary for increased accuracy.

Overall, while the science of machine learning in the detection of depression through social media has come far, it is still incomplete. While many models can create models to determine negative commentary, the model can be region specific or fail to properly deal with unstructured language.

V. WHAT IS FEELINGBOT

FeelingBot is an application that uses the Discord Bot API to monitor Discord servers it has been added to and using the GPT-3 API, analyses the sentiments of all comments made. It determines the sentiments of the users and can temporarily collect the data until presented back to willing users in the form of a mood calendar or other desired display. It can aid users in acknowledging their own feelings and serves as a proof of concept that such a Discord Bot could be used to analyze sentiments of users and reach out accordingly to aid those who may require it.

A. Who is FeelingBot For?

FeelingBot was designed to aid users in detecting patterns in their behavior. It is meant for general use in Discord servers, but especially large Discord servers that have a focus on mental health. It can be used by users who are concerned about their mental health, or general users to help keep an eye out on mental health concerns. FeelingBot will have to added to servers by their respective admins rather than by concerned users, but they will not gain any special permissions or access to information and can do so to help promote mental health and wellness.

B. What Problem is it Solving?

According to World Health Association, more than 264 million people are affected by depression worldwide [15]. It is considered one of the most treatable mental disorders; however, high rates of underdiagnosis and undertreatment suggest that existing screening is lacking [14]. Previous work has suggested that the potential of successfully employing social media to predict or identify depression, even prior to diagnosis is high [5]. However, most previous work employs depressive-indicative word dictionaries, which may not keep up with the evolving language of social media. Furthermore, currently no other technology provides users access to their own sentiment-based data in an easy-to-understand fashion. The objective of FeelingBot is to present and a more advanced sentiment identifier and to bridge the gap between research exploration and an accessible, practical service.

VI. HOW AI GETS USED IN FEELINGBOT

A. How does GPT-3 function?

GPT-3 is a language model that is made to produce human-like text using deep learning trained on a wide range of data. It is a natural language processing model that is advanced enough to imitate text that can be hard to distinguish from human-written text. [18, 19, 20] While it is proprietary to Microsoft, using an API Key, it is possible to have access to the model interface. We can interact with GPT-3 by setting some parameters, as presented in Table I, one of which is a string input, at which point the model returns a generated output [16].

The input parameter into the model is a string of characters, and the output generated is the string of characters that the model decides would be an appropriate continuation of the input, as decided by the parameters set and the data it had been trained on. The parameters can be seen in Table I. Using these parameters, all of which have default values being optional, GPT-3 can be customized to generate continuations to the prompt as desired. As an example, we can see that the python code in Table II can generate an output and analyze it. Note that some parts of the code have been redacted to safely comply with usage terms of OpenAI.

TABLE I. GPT-3 API PARAMETERS

Parameter	**Description**
engine_id	This parameter is a required field. It is a string and can be one of four values representing the four engined GPT-3 uses. In order of increasing complexity, the four engines are as follows, with each one being able to do that of all above. • "ada" this engine is the fastest. It can parse text, do non-nuanced classification, and keyword detection. • "babbage" this engine can do more nuanced classification as well as matching search queries with potential documents. • "curie" this engine is almost as capable as the next except for more complicated text. It can do sentiment classification and summarization as well as Q&A and chatbots. • "davinci" this engine is the most capable. It can detect complex intent, understand intent, as well as different perspectives.
prompt	This parameter holds a string or an array of multiple. This is the prompt which the model uses to generate the continuation string which will be the output. Default value is that which tells the model to start as if beginning a new document.
max_tokens	This parameter is an integer which tells the model how many tokens to use, which are shared between prompt and output. This is used to calculate length as well as determine engine usage for pricing or tracking reasons. Default value is 16
temperature	This number tells the model how risky it should be in generating text. Within the range of 0 to 1, 0 denotes ones with well-defined answers while values such as 0.9 denote ones with more creativity. Default value is 1. It is recommended that either temperature or top_p be used, not both.
top_p	Nucleus sampling number where tokens with more probability mass are selected, so 0.25 means tokens comprising top 25% probability mass will be considered. Default value is 1. It is recommended that either temperature or top_p be used, not both.
n	The number of completions for the given prompt. Default value is 1.
stream	A boolean which will make tokens be returned continually rather than all at once once completely generated. Default value is false.
logprobs	Returns the log probabilities on the amount of most likely tokens. Default value is null.
echo	Whether the completion will include the prompt as well. Default value is false.
stop	A string or array of strings at which the engine will stop generating the completion. Default value is null.
presence_penalty	A value between 0 and 1 that penalizes tokens based on if they already appear in the text. Used to encourage discussion of new topics. Default value is 0.
frequency_penalty	A value between 0 and 1 that penalizes tokens based on frequency in the text. Used to decrease likelihood of verbatim repetition. Default value is 0.
best_of	This number tells the model to return the number of outputs with the lowest log probability per token. Default value is 1.
logit_bias	Accepts a JSON object that allows for the modification of specific tokens appearing in the completion. Default value is null.

TABLE II. API IMPLEMENTATION

Python Code

```
import openai

openai.api_key = [redacted]
print(openai.Completion.create(
        engine="davinci",
        prompt="Once upon a time there was a
young",
        temperature=0.8,
        max_tokens=60,
        stop=["\n"]
    )
)
```

Returned Object

```
{
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "logprobs": null,
      "text": " boy named Tom. Tom lived with his
mother in a rose-covered cottage deep in the
forest. Tom liked to wander through the forest,
picking flowers and wild berries."
    }
  ],
  "created": 1627779589,
  "id": [redacted],
  "model": "davinci:2020-05-03",
  "object": "text_completion"
}
```

We can see that in the example of Table II, we used the Davinci engine with a temperature value of 0.8 and max tokens value of 60. We further tell the model to terminate before any new lines of text are generated. The input prompt is as follows,

"Once upon a time there was a young"

We can see that the model completed the query with the output string,

" boy named Tom. Tom lived with his mother in a rose-covered cottage deep in the forest. Tom liked to wander through the forest, picking flowers and wild berries."

The query was completed due to having reached the stop token rather than due to running out of valid tokens with no other problems.

B. How does FeelingBot use GPT-3?

As we have observed, GPT-3 can be customized with a lot of parameters to best suit the needs of the query. We can harness the context-understanding capabilities of the DaVinci engine to have the model define the sentiment of a statement. This can be done by setting the parameters as seen in Table III, with all others retaining their default values. Note that [statement] refers to the statement we are looking to analyze.

TABLE III. GPT-3 API PARAMETERS FOR FEELINGBOT

Parameter	**Value**
engine_id	davinci
For the engine used, we wish to use the full capabilities of the engine to best understand context and classify the sentiments. For the sake of the proof of concept, we will disregard the limitations of token counts.	
prompt	This is a tweet sentiment classifier Tweet: I loved the new Batman movie! Sentiment: Positive ### Tweet: I hate it when my phone battery dies Sentiment: Negative ### Tweet: My day has been :thumbsup: Sentiment: Positive ### Tweet: This is the link to the article Sentiment: Neutral ###

	Tweet: I have such a sad life Sentiment: Negative ### Tweet: This new music video blew my mind Sentiment: Positive ### Tweet: [statement] Sentiment:
We use GPT-3's ability to both interpret context as well as recognize patterns with this prompt. The initial sentence is interpreted so the AI can understand the context of the following lines, while some examples are set to demonstrate how we expect the output to be generated.	
temperature	0.3
We wish to have a low temperature to discourage the completion to be unrelated to the subject matter. Through preliminary testing and OpenAI's suggestion, a value of 0.3 functions well.	
max_tokens	60
While the output string is a single word denoting the sentiment, the prompt uses many tokens, especially if the user statement is long, and 60 is a number that can well account for this.	
frequency_penalty and presence_penalty	0.0
While these values remain unchanged from the default, it is important to note that we do not want to punish the model if multiple sentences exhibit similar sentiment.	
stop	###
Considering the format of the prompt, when the response reaches this unique set of characters, the response has been generated and no more tokens need to be used.	

Using parameters in Table III, we get generated outputs that are accurate and allow for classification of statements to be made quickly. Since any statement can be inserted into the prompt, it is flexible and responds to user statements without any additional information. Examples of FeelingBot classifying statements can be found in a latter section, under Tables V through VII, with a rudimentary example using the pull prompt in Table IV. The user statement in this example is "It was very tasty" with it being classified as positive.

While GPT-3 determines the sentiments of statements, the Discord API is used to interface with users. We are able to read all messages of users within servers that the bot has been added to by a moderator, but otherwise the bot will not directly intervene with server activity. We are able to make FeelingBot look over the statements made by users and then privately message them using the direct messaging system within Discord. It is also possible to keep track of a single user across multiple such servers as all users have unique tags accessible [17].

TABLE IV. PROMPT CLASSIFICATION

Prompt
This is a tweet sentiment classifier Tweet: I loved the new Batman movie! Sentiment: Positive ### Tweet: I hate it when my phone battery dies Sentiment: Negative ### Tweet: My day has been :thumbsup: Sentiment: Positive ### Tweet: This is the link to the article Sentiment: Neutral ### Tweet: I have such a sad life Sentiment: Negative ### Tweet: This new music video blew my mind Sentiment: Positive ### Tweet: It was very tasty Sentiment:
Returned Text
Positive

It was important to us that FeelingBot not retain any personal information for longer than necessary, and as such determines the sentiments of user statements, and only stores that information for a short period of time. Once a mood calendar or other method of information output is used, such as those seen in a later section as Figure II, to relay that information to the user, the data about said user is deleted.

C. Data Misinterpretation and Biases in FeelingBot

While GPT-3 can understand context within the prompt, for the entire context of a conversation to be added to the prompt for it to more accurately understand the conversation, the limiting factor is the tokens that can be used. As such, while theoretically possible to better understand a conversation context, it is not viable using the limited tokens available to us for the creation of this proof of concept. As an example, the simple statement "thats so funny" is considered to be a positive one, regardless of the context it had been written for.

While it is true that language is nuanced and not everything can be taken for its literal meaning, GPT-3 has been trained on a wide set of data and should be able to understand through the context of this fact [19, 20]. If developed into a full project, it would also be beneficial to further train the model using more specific mental health related messages.

Another way that data can be misinterpreted is through the use of links, gifs, emojis, or other non-words. While Discord's API translates all visible emojis to their respective codes, such as a thumbs up being ":thumbup:", not all emojis reflect their image [17]. Custom emotes can be named anything unique and as such when FeelingBot reads their code, it may misinterpret the message. When using links, GPT-3 is unable to define the respective sentiments they represent solely through the link, and as such they appear as neutral. This could be developed further by implementing search queries or a different model that can follow and summarize links. Computer vision models could also be used to categorize images or videos that are currently unable to be analyzed through GPT-3. These examples and their analyses can be seen in table VI.

Another aspect worth mentioning is that the data GPT-3 has been trained upon may be considered by some to be biased. We can see that GPT-3 may be aware of gender biases and roles, which while may prove that it has a good understanding of our society's structure, may make some users uncomfortable. Such bias has been observed within more free-structured text generation, and in some instances within our model [21]. It is critical that more testing and peer review is done to determine if these biases could cause this method of sentiment classification to be unequitable, but as things are there has not been any critical abilities for FeelingBot to be unequitable. Examples of these biases are highlighted in the example section Table VII.

VII. WHAT FEELINGBOT CAN DO

A. Artificial Intelligence aspects

As discussed in the above sections, FeelingBot is capable of detecting the sentiment of a statement using GPT-3. It can use this information to generate a mood calendar or otherwise present it to the user(s). This is the most complex aspect of the system but has been simplified through access to the GPT-3 API. With the data of the sentiments being shared accumulated, FeelingBot is able to use it in different ways.

B. Data representation aspects

Once having analyzed statement sentiments, FeelingBot can display it to the user. Since this is a prototype, we focused on implementing the sentiment analysis aspect, but in a completed model, the system can do more than just that, such as suggest steps to users to benefit their wellbeing.

Ideally, FeelingBot data could be accessed through an app on the user's device or online through a website. Users could use their discord login details or separate login details set up through the discord add in. This prototype was designed for the Discord app in mind; however, as other social media texting apps grow in popularity, the prototype could be updated to track sentiment data for other apps as well. Increasing the data FeelingBot has access to would also strengthen the accuracy of its analysis capabilities. It is important that Discord is designed with app add-ins like FeelingBot in mind, where other apps are not. Therefore, access to user data might be more complicated.

The FeelingBot app could consist of two main data models: a mood calendar and mood reports. A mood calendar is a calendar grid of colors, where

each color represents a user's overall mood for the day. A user's mood could be calculated using a percentage of their positive to neutral to negative sentiment data. For example, if sentiment data is overly positive, the mood for the day would be "joyful." If it overly negative, the mood for the day would be "sad." Notably, if FeelingBot determines it does not have enough data to make an accurate mood consensus, it will leave that day blank. An example of a completed mood calendar is provided in Figure II.

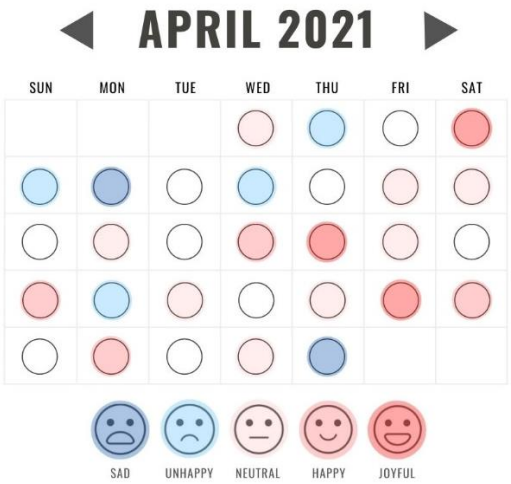

Fig. II. Example of Possible FeelingBot Mobile App Layout; Mood Calendar

A mood report would consist of a pie chart which presents the ratio of positive to negative mood days. Overall sentiment data could then be used to determine the mood for that period. Based on a user's overall mood, FeelingBot would suggest different tips. Figure III below provides an example of an overall positive mood report and some of the tips provided. Figure IV provides an example of tips for an overall negative mood report. As a Discord add-in, mood report periods could be monthly. As data access to other apps increases, reports could also be made weekly. Notably, as a user's overall mood can change drastically over the process of a year, calculating a yearly mood would be inaccurate and possible misleading.

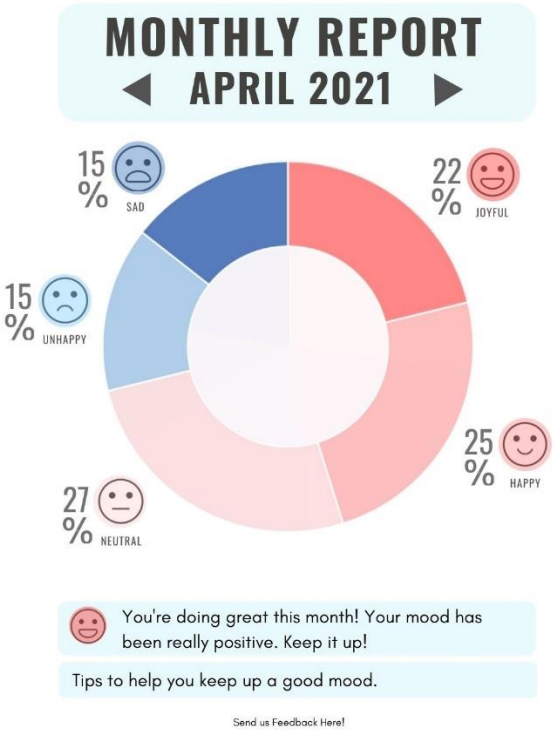

Fig. III. Example of Possible FeelingBot Mobile App Layout; Monthly Report

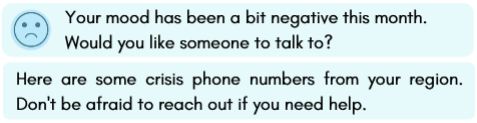

Fig. IV. Monthly Report; Overall Negative Mood Output

VIII. FEELINGBOT CLASSIFICATION EXAMPLES

We can look over some statements and their determined sentiment. Because FeelingBot does not analyze conversation context in its current model, the statement will have to be self-serving and looked at without any context. It is critical to note that we may not know exactly why GPT-3 outputs certain sentiments for statements, just that it has deemed that value best based on the training data and we can speculate as to the reasons. In all the testing done, there was no variance in the sentiments prescribed to each statement, and over multiple runs the sentiment remained unchanging. Some important examples to aid in understanding are found and discussed in Table V through VII.

TABLE V. STATEMENT CLASSIFICATION GENERAL

Statement	**Sentiment**
Hello	Neutral
How are you doing?	Neutral
These very simple statements are considered neutral. Depending on the context, they could be either positive or negative, but taken alone, and most likely, they are neutral.	
I am so tired today	Negative
I am hungry	Negative
In both above cases, most likely the lack of taking care of needs is noticed, since typically unwanted states are being stated, that of tiredness and hunger.	
I have bags under my eyes	Negative
I have dimples	Positive
Besides understanding explicitly stating feelings, cause and effect is understood, as having bags under eyes is a symptom of lack of sleep, while dimples are associated with smiling.	
I want food	Negative
I want pizza	Positive
I want bread	Neutral
While it may be left up to speculation, we can estimate why the model may view the desire for food as negative, that of pizza as positive, and that of bread as neutral. When discussing pizza, a food often associated with groups, since it is a specific desire and they are aware of how they will address their need for food, it seems to be positive. The need for food in general seems very similar to the above statement of being hungry, which points out that something is lacking. The wanting of bread seems mundane in that it is a basic food used to address a need. This could also simply be overanalyzing an otherwise trivial matter.	

TABLE VI. STATEMENT CLASSIFICATION LIMITATIONS

Statement	**Sentiment**
https://old.reddit.com/	Neutral
https://tenor.com/view/broken-in-rain-anime-sad-studio-ghibli-gif-16861288	Neutral
https://www.google.com/search?q=happy	Neutral
When analyzing links, even if keywords are found within them such as sad or happy, we see that they are simply classified as neutral since FeelingBot is unable to determine their context.	
You know it isn't "Legos". You've had FUCKING YEARS to adjust to the actual, correct way to say the term. It's Lego. Lego bricks, Lego sets, Lego kits, Lego mini-figures, Lego City. There are no such things as "Legos". They don't exist. "Lego" refers to the COMPANY THAT MAKES THE TOY, and thus the shortening Lego is acceptable. Saying "I'm playing with my Lego" works because it's referring to the sets themselves: The individuals aspects that make of the toy from the bricks to the mini-figures to the electronics to the other little parts. It isn't claiming that the fucking square bricks are each a Lego. THE ENTIRE THING IS. If you were to say "I'm playing with my Legos" that implies that you're playing with at least two different types of Lego set at once, i.e. Lego City and Bionicle. Still saying LEGOS after all these years makes you look like an assclown. Here in Europe, the continent responsible for this toy (no, it wasn't made by America, no matter how much your capitalistic toy industry wants you to think), you'd be laughed out of the room if you said that.	Negative
We can see that FeelingBot is able to analyze and interpret even large blocks of text correctly. Messages too long will be unable to be posted to Discord, and as such this limitation is innate to Discord and any such message can be analyzed by the model [17].	
hi :smiley:	Neutral
hi :laughing:	Neutral
hi :cry:	Negative
:smiley: :smiley: :smiley: :smiley: :smiley: :smiley:	Positive
:sob:	Negative
:100:	Positive
:bread:	Positive
:grin: :smiling_face_with_tear:	Positive

:smiling_face_with_tear: :grin:	Neutral
It may be of note that emojis are formatted as text where between the colons a simple word describes them. These may be custom, and therefore not very descriptive, but even with these trivial ones we see some detection being questionable. We see hi followed by an emoji be neutral unless paired with a cry emoji. This could be due to it being more common to add smiley emojis with such messages without it having much meaning. Even popular culture symbols such as the 100 emoji symbolizing something positive were detected, but this model also classified a simple loaf of bread as positive. It also classified two emojis that had swapped places differently.	

TABLE VII. STATEMENT CLASSIFIACTION BIASES

Statement	**Sentiment**
I am a boy and I have a mustache.	Neutral
I am a girl and I have a mustache.	Negative
I am a boy with long hair	Negative
I am a girl with long hair	Neutral
In these examples, some bias may be observed. While a statement by a male about their mustache can oftentimes be a statement of fact, a female making such comments can often be seen as a negative thing due to the mustache being socially undesirable. Whether this bias reflects the real-world values and as such can better detect sentiments, or is unnecessarily gendered can be a source of discussion for some. Similarly with the above, it may be more common for females to have long hair, suggesting the present bias.	
im gonna kms	Negative
ur a lil sheet	Negative
kys n00b	Negative
ye im aight	Neutral
In these examples we can observe that the model understands acronyms, slang, and typos. For reference, “kys” and “kms” are “kill your self” and “kill myself”, respectively, and “n00b” denotes someone lacking in skill.	

IX. FEELINGBOT DIFFERENTIATION FROM PAST WORKS

As discussed in Section IV, there has already been a lot of development in using AI in relation to mental health. Previous methods focus on individual posts on social media such as reddit or twitter rather than a chat room. Discord is an environment that can foster a lot more interactions, especially among friends where feelings are more likely to be expressed rather than public social media platforms where people may put up a front.

Another difference between many of the previously created AI, is that rather than detecting the sentiment of a statement through the content and context of it, they relied on the presence of keywords that are more common among depressed users. While GPT-3 may internally be influenced by keywords, it is not the sole factor that determines the sentiment of a statement. This is a different way to detect a similar concept, and if used in combination may aid in bettering the technology.

GPT-3 is also able to navigate some typos, slang, and other internet terminology due to having been trained on such data and understanding their meaning. Such classifications can be seen in Table VII, with FeelingBot able to classify both typos and slang appropriately as well as in Table VI, classifying emoji codes.

X. FUTURE CHALLENGES

Digital communication has created fundamental changes in the speed of evolution of the written language. Traditional written language like letters were slow, so words were chosen more carefully. Thus, the creation of neologisms and structuring of unstructured language was more uncommon. Synchronous, often fast paced written communication has resulted in a more active evolution, with the development emoticons, abbreviations, unstructuring, and neologisms [22]. These rapid changes means that AI designs cannot depend heavily on word dictionaries like the LIWC. While LIWC is helpful for posts on Facebook or

other similar blog-style social media sites where an older audience, who avoid neologisms and grammatically incorrect language, is common, it is ineffective in large text chat-style social medias like Discord, with a younger audience, who commonly switch up language to avoid restriction from chat admins.

It is currently unclear how to update current algorithmic designs to catch each possible instance of these slang values. Identifying and updating for the most common instances like “noob” versus “n00b” is a possible temporary solution. However, it is easy for users to counter this in the face of possible similar censorship. For instance, “n00b” could become “n**b,” “n//b,” or “***b.” Notably, even FeelingBot has trouble classifying every instance of slang. For example, it recognizes “unb0rn yourself” (kill yourself) as “positive” instead of “negative.”

Furthermore, a general identifier does not take into consideration more regionally specific slang words. For example, a study showed that from 2009 to 2012, the intentional misspelling of “suttin” for “something” occurred 5 times more often in New York City than in the rest of the US [22]. Although past work has recognized cultural differences across languages and ethnicities, research into cultural differences across a smaller subset of regions like cities, communities, racial groups within the same space is lacking. As such, most sentiment detectors, including FeelingBot are either region specific or overly general. Developing an algorithmic system that can understand these differences without depending on a system of regional sentiment dictionaries will be a challenge.

Similar to past work, FeelingBot is based on data from a single social media site. As the number and variation of social media sites increases, the effectiveness of a single social media text dataset to identify overall daily moods decreases. Notably, the difference uses for different social media sites results in individuals’ codeswitching, adjusting their style of speech, behavior, and expression. Furthermore, individuals on different sites have entirely different interpretations of what is considered negative or positive. For example, LinkedIn is a professional networking site, so the use of swear words would be a sign of unprofessional, and thus negative behavior. However, people on Snapchat, a more casual texting site for friends might use swear words for friendly ribbing, something not necessarily negative. Future work will need to consider these differences when developing multiplatform sentiment detectors.

Another concern with other platforms is image-based text and ASCII art. For platforms like Instagram, the main media is image and video. To successfully integrate sentiment detectors onto these platforms will require a more complex understanding of the sentiment of images. While some past work has used image light shades to determine the sentiment of pictures with people in it using neural networks, images with heavy text in them have not been heavily examined. Although traditional, easy to read fonts would be easy to analyze, text based visual art like ASCII is a more complex problem. ASCII art has highly varied designs and can be difficult for even human personnel to recognize. Figure V below provides an example of ASCII art.

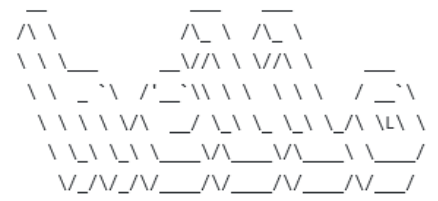

Fig. V. ASCII art for the word “hello”

ACKNOWLEDGMENT

We wish to express our appreciation towards the professor of this course for the providing useful resources concerning the field of artificial intelligence and to the teaching assistants who worked to answer questions that students may have had. We also wish to extend our appreciation towards fellow colleagues and friends who took an interest in Discord bots and allowed their presence in group servers.

REFERENCES

[1] Torres F, “What Is Depression?” American Psychiatric Association, 2020. Accessed on: August 3, 2021. [Online]. Available: https://www.psychiatry.org/patients-families/depression/what-is-depression

[2] “Depression,” *National Institute of Mental Health*, 2021. Accessed on: August 3, 2021. [Online]. Available: https://www.nimh.nih.gov/health/publications/depression/#pub1

[3] Bruce F. D, “Untreated Depression,” *WebMD*, 2019. Accessed on: August 3, 2021. [Online]. Available:

https://www.webmd.com/depression/guide/untreated-depression-effects

[4] "Fast Facts about Mental Illness," *Canadian Mental Health Association*. Accessed on: August 3, 2021. [Online]. Available: https://cmha.ca/fast-facts-about-mental-illness

[5] Genina A, Gawich M, Hegazy F. A, "A Survey for Sentiment Analysis and Personality Prediction for Text Analysis," in *Internet of Things—Applications and Future – Proceedings of ITAF 2019*, Ghalwash Z. A, Khameesy E. N, Magdi A. D, Joshi A, Ed. Springer Nature Singapore Pte Ltd, 2020, pp. 347 – 356. Accessed on: August 3, 2021. [Online]. Available: https://link.springer.com/book/10.1007/978-981-15-3075-3?page=2#toc

[6] Schimmele C, Fonberg J, Schellenberg G, "Canadians' assessments of social media in their lives," *Statistics Canada*, Mar. 24, 2021. Accessed on: August 3, 2021. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/36-28-0001/2021003/article/00004-eng.htm

[7] Walton G. A, "Depression Isn't Always What You Think: The Subtle Signs," *Forbes*, Feb. 17, 2015. Accessed on: August 3, 2021. [Online]. Available: https://www.forbes.com/sites/alicegwalton/2015/02/17/the-subtle-symptoms-of-depression/?sh=5b53d0b31a3e

[8] Tadesse M. M. et al, "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access*, vol. 7, pp. 44883-44893, 2019. Accessed on: August 3, 2021. [Online]. Available doi: 10.1109/ACCESS.2019.2909180

[9] "How It Works," *LIWC*, Pennebaker Conglomerates, Inc. Accessed on: August 3, 2021. [Online]. Available: http://liwc.wpengine.com/how-it-works/

[10] Boot B. A. et al, "How character limit affects language usage in tweets," *Palgrave Commun*, vol. 5, no. 76. July 9, 2019. Accessed on: August 3, 2021. [Online]. Available doi: 10.1057/s41599-019-0280-3

[11] Orabi H. A. et al, "Deep Learning for Depression Detection of Twitter Users," *ACL Anthology*, vol. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 88 – 97, June 2018. Accessed on: August 3, 2021. [Online]. Available doi: 10.18653/v1/W18-0609

[12] Rizvi Z. S. M, "A Comprehensive Guide to Build your own Language Model in Python!" *Analytics Vidhya*, Aug. 8, 2019. Accessed on: August 3, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

[13] Mikolov T, "Efficient Estimation of Word Representations in Vector Space," *Cornell University*, 2013. Accessed on: August 3, 2021. [Online]. Available: https://arxiv.org/abs/1301.3781

[14] Eichstaedt et al, "Facebook language predicts depression in medical records," *Proc Natl Acad Sci USA*, vol. 115, no.44, pp. 11203-11208, Oct. 30, 2018. Accessed on: August 3, 2021. [Online]. Available doi: 10.1073/pnas.1802331115

[15] "Depression," *World Health* Organization, Jan. 30, 2020. Accessed on: August 4, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[16] Microsoft, "OpenAI API," *OpenAI API*, 2021. Accessed on: July 31, 2021. [Online]. Available: https://beta.openai.com/

[17] Discord, "API docs for bots and developers," *Discord Developer Portal*, 2021. Accessed on: July 31, 2021. [Online]. Available: https://discord.com/developers/docs/intro

[18] R. Sagar, "OpenAI releases GPT-3, the largest model so far," *Analytics India Magazine*, Jun 3, 2020. Accessed on: July 31, 2021. [Online]. Available: https://analyticsindiamag.com/open-ai-gpt-3-language-model/

[19] L. Floridi & M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, Nov. 2020. Accessed on: July 31, 2021. [Online]. Available doi: 10.1007/s11023-020-09548-1

[20] R. Dale, "GPT-3: What's it good for?," *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, Dec. 2020. Accessed on: July 31, 2021. [Online]. Available doi: 10.1017/S1351324920000601

[21] L. Lucy and D. Bamman, "Gender and Representation bias in GPT-3 generated stories," *Association for Computational Linguistics*, pp. 48–55, Jun. 2021. Accessed on: July 31, 2021. [Online]. Available doi: 10.18653/v1/2021.nuse-1.5

[22] Eisenstein et al, "Diffusion of Lexical Change in Social Media," *Cornel University*, Nov. 24, 2014. Accessed on: July 31, 2021. [Online]. Available doi: https://arxiv.org/pdf/1210.5268.pdf