JMIR Data

Volume 5 (2024)   ISSN

Contents

Original Papers

Original Paper

Development of Depression Data Sets and a Language Model for Depression Detection: Mixed Methods Study

Faye Beatriz Tumaliuan[1*], MSc, MS; Lorelie Grepo[1*], MS; Eugene Rex Jalao[1*], MS, PhD

Department of Industrial Engineering and Operations Research, University of the Philippines Diliman, Quezon City, Philippines

[*]all authors contributed equally

Corresponding Author:
Faye Beatriz Tumaliuan, MSc, MS
Department of Industrial Engineering and Operations Research
University of the Philippines Diliman
Melchor Hall, Magsaysay Avenue
Quezon City, 1101
Philippines
Phone: 63 9176593613
Email: fayetumaliuan@gmail.com

Abstract

Background: Depression detection in social media has gained attention in recent years with the help of natural language processing (NLP) techniques. Because of the low-resource standing of Filipino depression data, valid data sets need to be created to aid various machine learning techniques in depression detection classification tasks.

Objective: The primary objective is to build a depression corpus of Philippine Twitter users who were clinically diagnosed with depression by mental health professionals and develop from this a corpus of depression symptoms that can later serve as a baseline for predicting depression symptoms in the Filipino and English languages.

Methods: The proposed process included the implementation of clinical screening methods with the help of clinical psychologists in the recruitment of study participants who were young adults aged 18 to 30 years. A total of 72 participants were assessed by clinical psychologists and provided their Twitter data: 60 with depression and 12 with no depression. Six participants provided 2 Twitter accounts each, making 78 Twitter accounts. A data set was developed consisting of depression symptom–annotated tweets with 13 depression categories. These were created through manual annotation in a process constructed, guided, and validated by clinical psychologists.

Results: Three annotators completed the process for approximately 79,614 tweets, resulting in a substantial interannotator agreement score of 0.735 using Fleiss κ and a 95.59% psychologist validation score. A word2vec language model was developed using Filipino and English data sets to create a 300-feature word embedding that can be used in various machine learning techniques for NLP.

Conclusions: This study contributes to depression research by constructing depression data sets from social media to aid NLP in the Philippine setting. These 2 validated data sets can be significant in user detection or tweet-level detection of depression in young adults in further studies.

(JMIR Data 2024;5:e53365) doi:10.2196/53365

KEYWORDS

depression data set; depression detection; social media; natural language processing; Filipino

Introduction

Background

Depression is a health condition involving changes in emotion, thinking, or behavior. It represents a substantial part of mental health discussions. It affects a wide range of the public, as it can be experienced by anyone to some degree regardless of nationality, culture, gender, age, financial status, or lifestyle. The World Health Organization reported that the Philippines has one of the highest numbers of persons with depression in Southeast Asia, affecting 3.3 million Filipinos [1]. In recent years, the country has become helpless from the negative impacts of the public health emergency and economic crisis from the COVID-19 pandemic. The Philippines' National Center for Mental Health reported an increase in suicide-related calls received on their hotlines, from an average daily call volume of 400 in 2019 to 700 from March to August 2020. The Philippine Statistics Authority in July 2021 also reported an increase of 57% in the suicide rate between 2019 and 2020, during the height of the pandemic [2]. The World Health Organization assessed in 2011 that, by 2030, depression will be the chief source of the worldwide disease burden.

Social Media and Natural Language Processing for Depression Detection

Although depression and other mental illnesses may lead to social withdrawal and isolation, it was found that social media platforms are indeed increasingly used by affected individuals to connect with others, share experiences, and support each other [3]. A similar study in the United States concluded that internet users with stigmatized illnesses such as depression are more likely to use online resources for health-related information and communication about their illness than people with another chronic illness [4]. It was found that depressed individuals perceived social media as a means of maintaining social awareness and consoling themselves, while nondepressed individuals perceived it as a means of information sharing and consumption [5]. Moreover, because of the amount of complex behavioral data generated by users in social media, it has been explored in mental health research using pattern recognition, predictive modeling, classification, and categorization techniques to predict the presence of mental disorders, such as depression, suicidality, eating disorders, schizophrenia, and anxiety [6] in social network platforms such as Facebook, Twitter, Reddit, Weibo, and Sina Microblog.

Most of the methods mentioned are techniques used in natural language processing (NLP), which is the ability of computers to understand written or spoken language. It requires an extensive amount of data to build such models for a specific language. A language is considered a high-resource language if many available data resources exist (dictionaries, lexicons, treebanks of syntactically annotated data, all types of digitized text data such as news reports, books, social media, and other web-based data and annotated data for training depending on the NLP objective), making it possible to train machine learning models with these amounts of data. The national language of the Philippines, which is called Filipino, is based on an Austronesian language called Tagalog [7]. English is also considered a second language in the Philippines, with both English and Filipino used in written or spoken conversation, and most of the time used together in informal conversation, which forms “Taglish” (the combination of the 2 languages). Although English is a high-resource language, Filipino is considered a low-resource language. Because Filipino is a low-resource language, most NLP tasks and studies on mental health research focus on corpus building (or building depression vocabularies) to be used for other studies that need language processing [8-10].

Objectives

In this work, we aim to construct depression data sets from social media data to improve Philippine resources on depression research to aid in NLP.

The main objective is to build a depression corpus of Philippine Twitter users who were clinically diagnosed with depression by mental health professionals. Related work on depression detection in social media builds data from web-based questionnaires via voluntary participation with psychologist help [11], builds data from self-declared individuals with depression on the web (clinically diagnosed or topics of self-harm) [12], or builds data from online forums with depression topics versus other nondepression-related forums [13]. These examples and other similar studies have attained successful results, but an extensive validation and cross-examination of the users of the data set with clinical processes and diagnosis is lacking.

We also aim to develop a corpus of depression symptoms that can later serve as a baseline for predicting depression symptoms in the Filipino and English languages. Because of the low-resource standing of Filipino depression data, this corpus of depression symptoms needs to be developed to perform and measure the predictive performance of various machine learning techniques in classification tasks.

Finally, we aim to develop a language model on the basis of Filipino and English social media data, which can be used to represent text data as vectors to aid in various downstream tasks in NLP.

Given the aforementioned objectives, we first constructed a data set of Twitter users validated by licensed clinical psychologists, and from these data, we constructed another data set of depression symptoms using a manual annotation process. We proposed a language model using the word2vec [14] algorithm using web-based texts in English and Filipino.

Related Work

Natural Language Processing

NLP is used by several studies to determine depression patterns or detect depression in social networks. Social media data are cleansed, segmented (separated into documents or sentences), tokenized (split into words or groups of words), and normalized (broken down into parts-of-speech, stemmed, or lemmatized to achieve root words or synsets for spelling correction and reducing complexity) in preparation for analysis. Sentiment analysis, a technique used to determine the sentiments of texts through data, is usually performed by classifying texts on the basis of polarity (positive and negative), valence or intensity, emotions, or subject dependency. Analysis may also be performed by categorizing data into ≥1 sets of “classes,” which are called “classifiers.” Such classifiers for machine learning used include regression, support vector machines, naive Bayes, and decision trees. Most recent methods convert words into embeddings or vectors (number representation of words) to perform deep learning analysis [11,13,15]. Deep learning uses networks that imitate the way neurons in the brain work to solve complex problems. Using deep learning methods can determine word similarity and analogy from well-trained corpora, which can be hard to achieve from the previously mentioned methods, such as accurately determining negation cues and word context and relationships. Results from social media studies that use these methods suggest that language use has been found to be a predictor of depression [16].

Word Embeddings

For text data to be inputted as features into machine learning or neural network models, text data will need to be translated or represented into numbered vectors. Doing so makes it easier to represent words in a vector space and enables computations to be performed. It also enables words or text to be presented in several dimensions.

Several word embedding algorithms are available, and each has its strengths and weaknesses. Most of these algorithms need considerable training data to learn and create a language model. A vocabulary is created from all words existing in the training data, words are one-hot encoded (represented as 0 and 1), and probabilities of co-occurrence are computed between all the words in the training sample. One example is the word2vec model [14], which uses Continuous Bag of Words (CBOW) and skip-gram techniques to train a language model. CBOW aims to predict a target word from a set of words, while skip-gram is the opposite, wherein it tries to predict the probability of a given word being present when an input word is present. While training this model, a hidden layer of n-dimensions is created, which act as features that represent a word and its relationship to other words, thereby enabling the model to learn the semantic relationship between words, such as synonyms, antonyms, superclass, and neighboring or similar words. Other algorithms are available, such as global vectors [17], efficient learning of word representations and sentence classification (FastText), and BERT (Bidirectional Encoder Representations from Transformers) tokenizers. These models are trained repeatedly through the number of epochs, in which 1 epoch is a complete pass of the training data through the algorithm.

Word embeddings can serve as an input to machine learning algorithms to be used in downstream tasks such as classification or prediction tasks for depression detection in text data.

Depression Detection in Social Media

The most common social networks used for studying depression include Twitter [12,18-21], Reddit [13,22,23], Facebook [11,15], Weibo [20,24,25], and Sina Microblog [16]. Owing to being high-resource languages, most of this research is for languages in English and Chinese. Detecting depressed users from these social networks is a common theme (user-level detection), with depressed users [11,12,15,16,18,20,24,25], self-harm [26], and suicidal risk [23]. Some explored the detection from depression comments or texts [13,27], which are not on a user level (tweet level or comment level). Some studies also focused on detecting the degree of depression (ie, mild, moderate, or severe [19,21,26]) and early detection [22,26,28].

User-Level Versus Tweet-Level Depression Symptom Detection

Overview

Linguistic patterns, such as the use of positive or negative words, and social media behavior patterns, such as posting and user interaction behavior, are shown to have positive results in detecting depression [11,16,20]. Although several studies focus on depression screening, most studies focus on screening the users (user level) and do not focus on the emergence of depression symptoms in social media language. The social media data that users create can help doctors have a glimpse of their mental state through daily activities [5]. For example, on a tweet-level basis, passive daily activities can show known signs of psychological depression symptoms, such as changes in mood, activity level, sleeping and eating patterns, and suicidal ideation. This level of granularity by looking at the symptom level can aid the depression symptom tracking process and can open up opportunities to closely study depression patterns. It can also eventually lead to detection on a user level, determine the severity of depression through tweet-level symptoms [21], or help with early detection studies.

Data Collection Methods

The data collection methods for social media use specific application programming interfaces (APIs), which are a set of programmable commands that allow software interaction with certain websites.

User-Declared or Self-Declared Depression

In this method, users are identified as depressed when the self-reported sentence pattern “I’m diagnosed with depression” is matched or other patterns such as “(I’m/I was/I am/I’ve been) diagnosed depression” [12,20,22,28]. For the control group, users are labeled as nondepressed if no tweets containing “depress” were published in the sampling period or by selecting random users.

Group or Topic Involvement or Keyword Search

Studies that made use of group involvement performed data collection for depressed users by crawling subtopic groups catered to each social media platform, for example, Reddit subreddit groups such as r/depression, r/SuicideWatch [23], and r/selfharm [26] and Weibo’s subtopic function “SuperTopic” [24,25]. The control group for nondepressed users is from other topics not related to depression or random users. Other studies use keyword search and apply it on Twitter and Sina Microblog using seed word generation methods from words extracted from each of the symptoms of the Patient Health Questionnaire-9 (PHQ-9), also with the help of psychologists [21].

Use of Depression Questionnaires

In the clinical setting, most initial screening tools make use of questionnaires for assessment by mental health experts or primary care physicians. A sample of the questionnaires from some of the previous studies that looked into depression detection are the Center for Epidemiologic Studies Depression Scale screening test [18,19], Beck Depression Inventory [26], Thai Mental Health Questionnaire [11], and PHQ-9 [21].

Manual Annotations

Looking at the previous data collection methods mentioned, it is crucial to note that each method has its limitations and biases. Some of the studies address this through manual checking or annotation of data by psychologists or trained personnel [21,22,24,25].

Interviews With Experts

From all the other data collection methods previously discussed, the "ground truth" method would be clinical interviews with mental health experts, as they serve as valid clinical assessments or diagnoses. Wang et al [16] used this method in their study, in which a group of psychologists diagnosed hundreds of volunteers using questionnaires and interviews. This method is the most costly and time consuming because mental health experts need to assess every participant in the study, but this removes the biases from the previous methods and makes sure that the 2 groups studied (with or without depression) are labeled correctly.

Combined Modalities

Overview

The studies above show that NLP and machine learning techniques can classify depression in users or depression in user comments on social media. It can be used to identify early risks and crises as well. Combining different modalities such as text, images, social media behavior, and results from psychological studies can improve detection capability. The accuracies of the techniques discussed in this review range from 0.69 [19] to 0.925 [13] using support vector machines, random forests, and other sentiment analysis techniques and have improved to up to 0.98 [23] in the last 2 to 3 years with the improvement of deep learning and word embedding techniques.

Philippine Studies for NLP and Depression

Initial studies for Filipino NLP act as an enabling resource for computational linguistics. This includes the Filipino wordnet project (FilWordNet) developed by Borra et al [29], which can serve as a basis for a stemmer, lemmatizer, or development of a named entity recognition system because of its manually created synsets and root words created in a dictionary format. Keen et al [9] developed FilCon, a generated subjective lexicon that contains subjectivity scores, positivity, negativity, and neutrality polarity values. Andrei [8] developed a Linguistic Inquiry and Word Count dictionary for positive and negative emotions using tweets, while Lapita et al [10] developed an emotion-annotated corpus of disaster-relevant tweets.

Bitsch et al [30] identified depression symptoms and categorized depression symptom categories to create a depression lexicon for pattern matching of the presence of depression symptoms. It is used as a mobile app solution. Nartia et al [31] developed a predictive model using a decision tree for identifying the mental health condition of university students, whereas Aliman et al [27] developed a bot that classifies potential mental health crisis tweets using logistic regression. Aperocho [32] used a qualitative analysis approach to understand the use of Philippine English in the depressive language by analyzing the discourses posted by netizens on Facebook.

Research Gaps

One of the challenges in depression detection in social media is the collection of training data. Social media offers a huge opportunity in data science as it offers a huge amount of data that can be easily collected through available APIs. The downside is that it suffers from ethical concerns, validity, and credibility. Studies must ensure that users are informed of the data collection process, in which case, most of the previous studies scraped data from free APIs and did not involve the consent of users in the process. Another issue is the validation of whether the users have experienced depression. Questionnaire methods and clinical interviews increase the validity and are most credible but are costly and challenging to implement, which is the reason why most studies can only focus on data science methods or involve psychological experts in improving the data science methodologies in these studies.

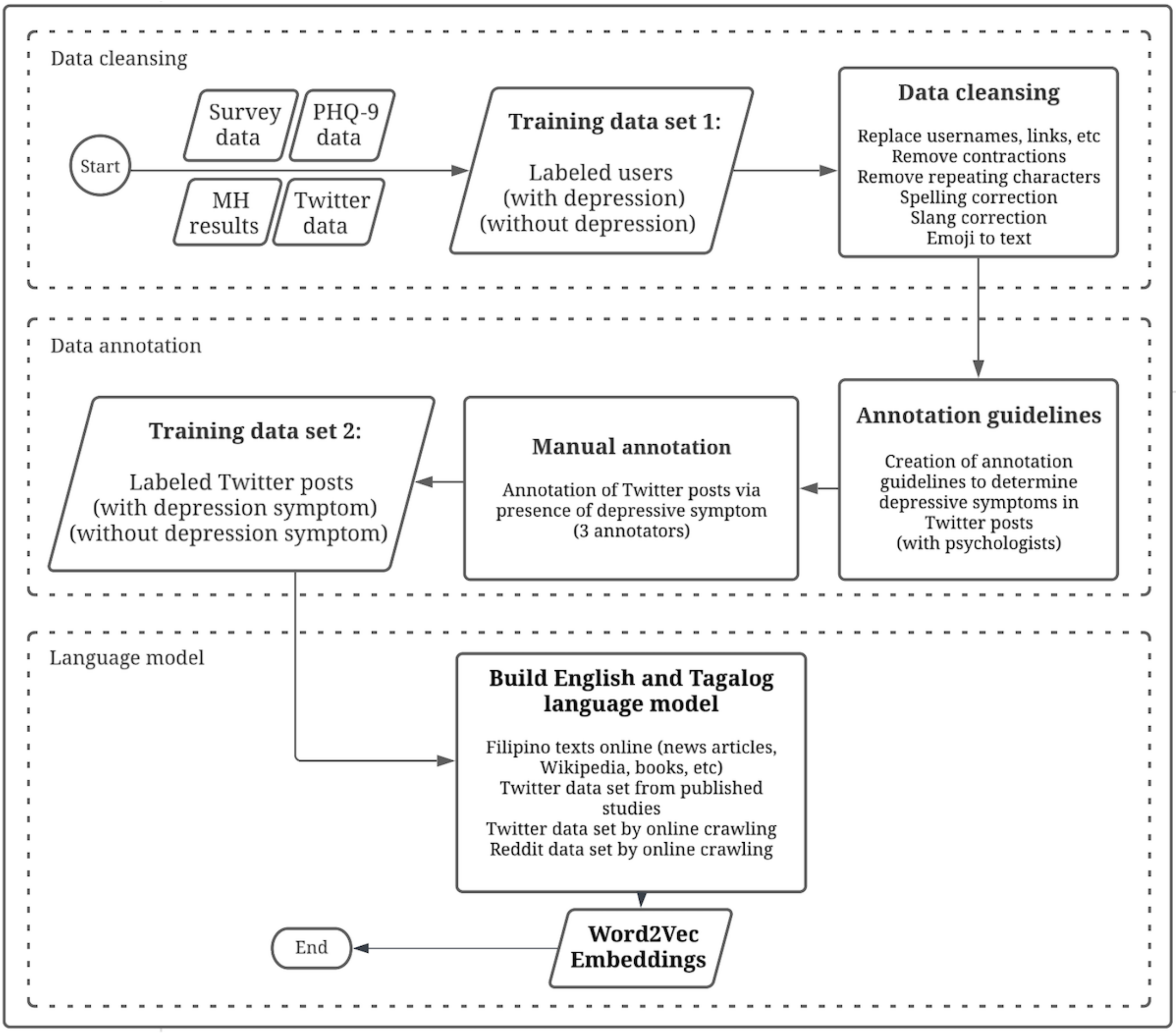
Most studies discussed are also focused on user-level detection. User-level detection not based on symptom tracking can mimic screening methods but is not efficient to implement because clinical diagnosis still needs expert help. However, a depression symptom detection model on a text or tweet-level detection that can identify symptoms over time can help identify symptoms that can complement clinical practice and improve its efficiency. Aside from this, the current Philippine studies discussed can aid depression research in NLP if the goal is to determine how polarity, emotion, or sentence structure in text data relates to depression. As per similar studies abroad that use social media to determine patterns of depression, there is a lack of data set of social media users with clinically validated depression cases and a lack of depression symptom text data sets in Filipino and English, which can act as a baseline for depression symptom detection.

Methods

Overview

This section describes the data collection and methods for data set creation for this study. Figure 1 illustrates the whole methodology framework for this study. It includes the criteria for participant collection, sampling design, annotation guidelines and processes, and Filipino and English language model creation.

Figure 1. Methodology—this figure shows the different steps performed in this methodology, from data cleansing, data annotation process, and language model creation. It shows the creation of data sets 1 and 2 and word2vec embeddings, which are the major outputs and objectives of this study. MH: mental health; PHQ-9: Patient Health Questionnaire-9.


Data cleansing
Start
Survey data
PHQ-9 data
MH results
Twitter data
Training data set 1:
Labeled users
(with depression)
(without depression)
Data cleansing
Replace usernames, links, etc
Remove contractions
Remove repeating characters
Spelling correction
Slang correction
Emoji to text
Data annotation
Annotation guidelines
Creation of annotation guidelines to determine depressive symptoms in Twitter posts (with psychologists)
Manual annotation
Annotation of Twitter posts via presence of depressive symptom (3 annotators)
Training data set 2:
Labeled Twitter posts
(with depression symptom)
(without depression symptom)
Language model
Build English and Tagalog language model
Filipino texts online (news articles, Wikipedia, books, etc)
Twitter data set from published studies
Twitter data set by online crawling
Reddit data set by online crawling
Word2Vec Embeddings
End

Study Design

This study is a 2-stage research that aims to detect depression symptom patterns in social media. It involves stage 1 (data generation) and stage 2 (depression symptom detection). This paper focuses on the first stage of generating depression data sets to be used for detecting depression symptoms in the second stage. The second stage is intended for future research in another paper.

Ethical Considerations

There are numerous ethical questions in mental health research, especially with regard to this study, which deals with social media and its various issues with privacy concerns. All participants signed an informed consent form, aiming participant understanding on what data was collected, and the privacy considerations and confidentiality measures that were applied, in addition to having complete control on what part of their data they wanted to share. Participant data were anonymized and any personal identifiable information were removed before any analysis was performed. Tweet text data were cleansed of participant names, usernames, links, and hashtags, which ensures no tweets can be identifiable and traced back to a particular user.

The UP Manila Research Ethics Board (UPMREB 2022-0135-01) have reviewed and approved all methodologies and recruitment materials before the start of participant recruitment.

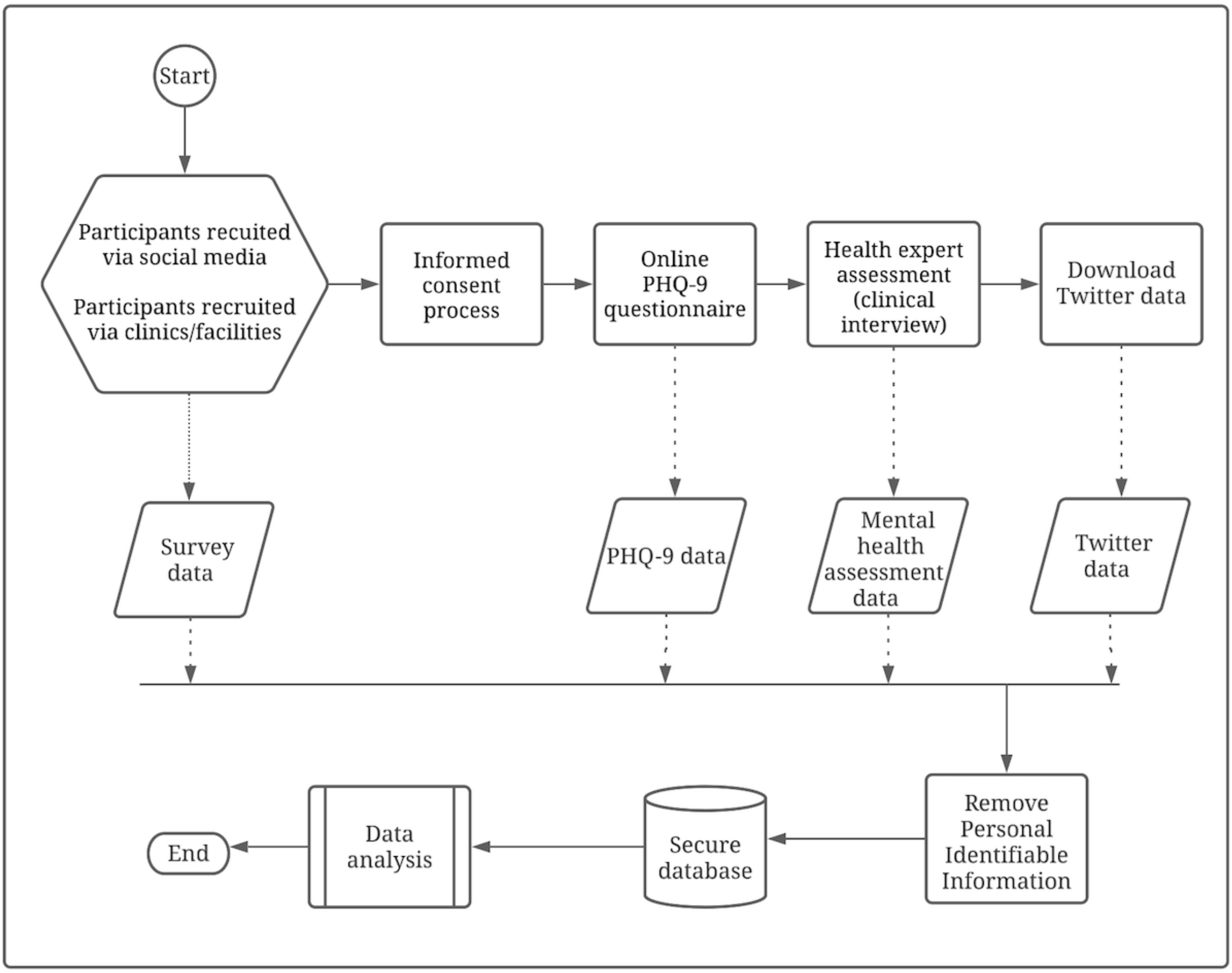
Recruitment

The inclusion criteria for participants in this study are Filipino Twitter users living in the Philippines who speak or write mainly in English or Filipino aged between 18 and 30 years. This age group is chosen for this study as the young adults' age group has been studied for mental health conditions and social media use (aged 18-24 years) [3], and the prevalence of suicide and depression in young adults in the Philippines (aged 15-29 and 18-24 years) [1,33].

Participants are recruited from clinics and social media platforms and are asked to answer a survey form and a web-based self-assessment form which is the PHQ-9 depression questionnaire. After the web-based self-assessment, all participants are assessed by a psychologist if experiencing or diagnosed with clinical depression (in any form or severity) or otherwise. The control group is the participants not assessed with depression or any other mental health conditions. Participants who were not assessed with depression but were assessed with other mental health conditions are excluded from this study. If assessed with any other mental health conditions (posttraumatic stress disorder, obsessive-compulsive disorder, bipolar disorder, schizophrenia, etc), they are included in the study as part of the depression group as long as the participant is experiencing depression because these are symptoms that are said to cause or can coexist with depression [34]. In short, the control group consists of participants who are assessed to have "usual" or "typical" mental health behaviors and the depressed group consists of participants assessed with depression.

Figure 2 illustrates the data collection workflow for this study. The participants were asked to sign a consent form for the collection of their social media data. This step informed the participants about what data will be collected and the privacy considerations and confidentiality measures that will be applied to their data. All participants were asked to share their data via the "Download an archive of your data" feature of Twitter in their account settings, which includes all historical data for an account. This step also makes sure that participants have complete control over what part of their data they want to share. Interactions with participants were performed by a mental health expert who was also on standby throughout the data collection phase should the participants experience distress and need support from a mental health professional.

Figure 2. Data collection workflow and generated data—each participant who completed the data collection process is expected to have 4 types of data collected. These are the survey data (which includes demographics and personal information), Patient Health Questionnaire-9 (PHQ-9) results, Mental Health Assessment results (includes symptoms, medical history, and depression assessment), and Twitter data archive (all historical Twitter data).


Start
Participants recuited via social media
Participants recruited via clinics/facilities
Informed consent process
Online PHQ-9 questionnaire
Health expert assessment (clinical interview)
Download Twitter data
Survey data
PHQ-9 data
Mental health assessment data
Twitter data
Remove Personal Identifiable Information
Secure database
Data analysis
End

Sampling Design

Nonvalidated study data sets on depression on social media (those which are either scraped from the internet via self-declared users or self-assessment questionnaires) vary from a minimum of 20 [11] to 137 users with depression data [26]. As there are no known studies on this topic that use a data set with validated users (those whose users were examined by clinics or experts), the greatest number of validated data sets is from the study by Wang et al [16] with a data set size of 122 users with depression.

Our data collection process included an assessment session with a psychologist, thus restricting the cost of the total number of participants. Puyat et al [33] estimated that 8.9% of young adults experience moderate to severe depression in the Philippines. To attain population-proportion sample sizes for the 2 groups (persons with depression and the control group) in the study, the aim was to achieve at least 10% of the depression group from the overall participant data in this study. We aimed to gather approximately 80 participants for this study; thus, we needed at least 10 participants in the depression group. For a user-level detection of depression, a balanced class distribution (40 participants for the depressed group and 40 participants for the control group) is ideal for most classification tasks. However, because the goal is to create a data set of depression symptoms in text data (tweet-level depression detection), the balanced class is not imposed as long as the minimum requirement of 10 participants in the depression group is met. The resulting data collected from this process was tagged as data set 1, which includes labeled data of users who were "assessed with depression" (mild, moderate, or severe) or "not assessed with depression."

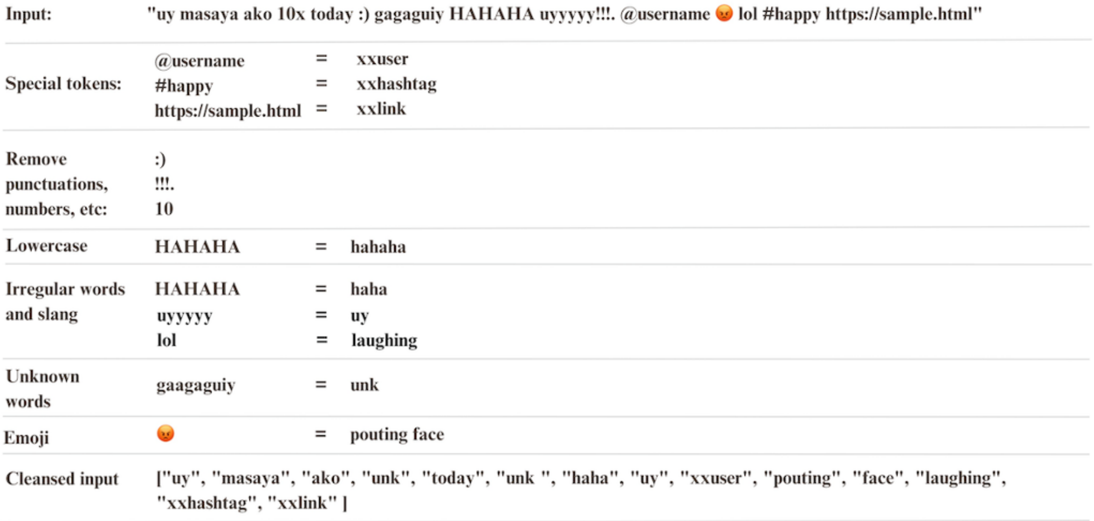
Data Preparation

Data Cleansing

Tweet text data were cleansed of usernames, links, and hashtags. Participants' names are sometimes included in the tweet texts when referencing themselves during tweets; thus, these were cleansed as well. Special tokens for these cases were created, namely "xxuser, xxlink, xxhashtag, xxname." All these types of texts were replaced via their corresponding tokens.

Some text data were removed, such as nonalphabet or nonemoji characters, punctuations, and numbers. All remaining texts were converted to lowercase. Tweet text data contain irregular words such as contractions, typographical errors, elongated words, and slang words. For English words, contractions are replaced using the Python "contraction" package. Using this package transforms the word "don't" into "do not." Words with repeating characters were normalized by removing the occurrence of consecutive single characters unless it is a valid English or Filipino word. Abbreviations, wrong spellings, and slang were also corrected by creating a list of words that usually occur in social media texts. Spelling correction examples are Filipino words such as "d" transformed to "hindi" or English words such as "some1" changed into "someone." Slang correction examples are "pov" into "point of view," "dasurb" into "deserve," and so on. Every word in a tweet is checked whether it exists in this list that was generated for spelling correction and slang correction and is replaced by the corresponding word equivalent using Python search. Finally, emojis are transformed into equivalent text using the Python "demoji" package and tokenized using the "TweetTokenizer" package, illustrated in Figure 3.

Figure 3. Data cleansing—this figure shows the data cleansing steps of collected tweets from the conversion of tokens, removal of text, transforming into lower case and irregular words, conversion of unknown words, and conversion of emojis.



Input:	"uy masaya ako 10x today :) gagaguiy HAHAHA uyyyyy!!!. @username 😡 lol #happy https://sample.html"		
Special tokens:	@username #happy https://sample.html	= = =	xxuser xxhashtag xxlink
Remove punctuations, numbers, etc:	:) !!!. 10		
Lowercase	HAHAHA	=	hahaha
Irregular words and slang	HAHAHA uyyyyy lol	= = =	haha uy laughing
Unknown words	gaagaguiy	=	unk
Emoji	😡	=	pouting face
Cleansed input	["uy", "masaya", "ako", "unk", "today", "unk ", "haha", "uy", "xxuser", "pouting", "face", "laughing", "xxhashtag", "xxlink"]		

Annotation Guidelines and Process

To create a depression symptom data set, a training data set 2 needed to be created in which individual tweets were manually annotated as having depression symptoms or no symptoms. Annotation guidelines were needed for annotators to be in sync for data to be well labeled and valid for the next steps. The creation of annotation guidelines was held with 2 clinical psychologists in a series of web-based sessions. The objective was to create depression symptom categories and to determine rules and guidelines for annotating tweets into the respective categories. The categories were not mutually exclusive, as they can simultaneously occur together in one given tweet.

The result of this exercise is the reviewed and finalized depression symptom categories with the psychologists. These categories are based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) criteria for depression (The Structured Clinical Interview for DSM-5), PHQ-9 results, and Mental Health Assessment results from this study. Textbox 1 shows the depression categories from the annotation guidelines developed in Textbox 2.

Textbox 1. Depression categories for annotation.

Thinking, concentration, and decisions

- Unable to handle minor problems and daily activities—procrastination and academics
- Diminished ability to think or concentrate, indecisiveness, unorganized thinking, forgetfulness, slow

Interest and motivation

- Loss of energy or motivation
- Loss of interest or enjoyment of activities, such as sex, hobbies or sports

Physical: sleep

- Insomnia or sleeping too much

Physical: fatigue

- Fatigue, tiredness, or lack of energy
- Emotional, mental, or physical fatigue
- Example: Filipino context—“Pagod na ako” (in English: “I am already tired”)

Physical: appetite

- Weight loss or gain, no appetite, or increased cravings

Physical: movement

- Moving or speaking slowly
- Feelings of restlessness or agitation
- Movement issues or wanting to stay in bed
- Example: Filipino context—“Hindi makakilos, gusto lang sa kama, hindi mapakali” (in English: “I cannot move, I want to stay in bed, I feel restless”)

Substance use

- Using recreational drugs or alcohol, etc (excluding coffee, melatonin, etc)
- If a substance is used as a coping mechanism
- If the deed is already done, not if intention only

Patient Health Questionnaire-9 (PHQ-9)—suicidal tendencies

- Recurrent thoughts of death, suicide attempts, or suicide
- Self-harm
- Overelaborate or strange ideas

PHQ-9—emotional: depressed and sadness

- Depressed mood or lonely
- Feelings of hopelessness, tearfulness, emptiness, or grief
- Breakdowns
- Emotions not induced by movies, events, or other outside triggers

PHQ-9—emotional: worthlessness and worry

- Feelings of worthlessness or no confidence, feeling misunderstood, self-doubt, or hypercritical toward self
- Feelings of guilt, fixating on past failures or self-blame, worry, discouragement, demotivation, or disappointment
- Stress, overthinking, or anxiety
- Negative thoughts and existential questions

Physical: pains

- Headache, back pains, stomachache or sexual dysfunction, tremors, or cold hands and feet
- Panic, choking, or numbness
- Pains triggered by stress (with the context of stress [eg, allergies, acid reflux, or gastroesophageal reflux disease])
- Exclude premenstrual syndrome signs or pains induced by known causes (accidents, etc), as said in the tweet

Social

- Social withdrawal or avoiding social interaction
- Sensitivity to criticism, irritability, agitation, or angry outbursts
- Detachment or isolation

Mental health–related issues

- Tweets not pertaining to any symptoms previously mentioned but denote that the user is associated with any mental health issues, such as talking to psychologists or being aware of mental health issues, or reminiscing traumas

No symptom

- The tweet has no symptom
- The tweet does not have enough context to be determined as a symptom

Textbox 2. Annotation guidelines.

Tweet annotations

- Tweet annotated as a whole, not by subsentences

Multiple symptoms

- A maximum of 3 symptoms can be annotated per tweet
- Example: "I cannot sleep nor move my body in this bed. Brb gonna die of starvation"
- Symptoms: Physical: sleep; Physical: movement; Physical: appetite

Speaker's view

- Refrain from viewing the tweet as an outside person

Cause and effect relationship

- If any sentence builds upon a cause and reflects an effect, the underlying context or symptom in that tweet should be judged on the basis of the effect, for example, "If he wins, I will kill myself"; this will be judged as a symptom regardless of the said cause

Quotes, song titles, lyrics, etc

- Users sometimes portray their emotions or thoughts using these methods. As such, if the text suggests enough context, these are annotated as symptoms

Jokes and sarcasm

- On the basis of the context, jokes and sarcasm are annotated as symptoms
- Jokes (eg, "I will kill myself. Kidding"); symptom
- Sarcasm (eg, "Kill me now I am so bored"); no symptom

Rants

- Treated as irritability symptoms

Politics

- These are treated as a symptom if a tweet suggests a strong emotional context that denotes a symptom, for example, "I am so angry at this government it makes me worry about my future"; social, PHQ-9—emotional: worthlessness and worry

Sadness over movies, events, "missing" people, places, or things

- These are not treated as symptoms because these are triggered by outside events

Metaphors or expressions

- These are examined depending on context; with enough context, they are treated as symptoms.
- Example: "Ang sakit mo sa ulo" (in English: "You are a headache"); no symptom
- Example: "Sakit ng ulo ko" (in English: "I have a headache"); symptom

Annotation Process

The process started with the selection of tweets to be annotated. All participants were included in this selection; included tweets are dated January 1, 2021, to November 30, 2022. Retweets and non-English and non-Filipino tweets were removed. Some tweets that were empty after data preparation and cleansing steps were discarded. The tool used for annotating tweets is Microsoft Excel. After tweet selection, these tweets were all annotated by the first-pass annotator. Tweets without depression symptoms are treated as "no symptom," and tweets that denote a depression symptom are passed on to the next 2 annotators.

These tweets checked by the first-pass annotator are tweets that significantly do not denote any depression symptoms. Examples of these texts were tweets with incomplete contexts, such as emoticon or emoji replies, responses to other tweets (random conversations in a thread), tweets with only photos, links, or other media, one-worded or vague tweets, and third-party application–related tweets (eg, Wordle).

Tweets were annotated anonymously, individually, and in a randomized order. Annotators had no knowledge of the users and of the tweets being annotated, and the annotations are based solely on professional experience, together with the guidelines provided to them. This process is repeated in 4 iterations with different subsets of the data sets in each iteration, and all annotators provided their annotations in 10 days. Tweet label categories were considered correct if at least 2 annotators selected the same symptom category. After each iteration, an annotation review is conducted within the annotators for tweets that do not have 2 agreeing annotations on the symptom category. A thorough discussion for each tweet is held until a consensus is reached among the annotators. Most tweets that are reviewed during this annotation review are due to human error, as the annotators expressed their sentiment that they intended to select a category but mistakenly selected the neighboring category. The remaining tweets needed some thorough discussions, and annotators decided together which category was most appropriate.

After the annotation process, an annotation validation is held wherein 4 to 5 random samples per category were chosen, a total of 68 tweets. These tweets were reviewed by the psychologists together with the annotators.

Measuring Annotation Reliability

The data set produced in the annotation step served as the baseline or "ground truth" when predicting depression symptoms in downstream tasks. The reliability of the annotation task can be measured by an interannotator agreement score, which measures how well multiple annotators can make the same decision for the given annotation labels or categories. This measure determines how well the annotation guidelines have been established, how well the annotation labels or categories have been identified, and how well the annotators have grasped the annotation rules. It can also measure the reproducibility of the annotation task at hand.

For this annotation task, the Fleiss κ measurement is used, as this can consider multiple annotators. This measurement also considers the possibility of the agreement occurring by chance. Raters having complete agreement will have an output of $\kappa=1$, and rates with no agreement will have $\kappa<0$. The following equation is used to compute the Fleiss κ score [35]:

$$= (\rho_o - \rho_e)/(1 - \rho_e)$$

Where ρ_o = observed agreement of raters and ρ_e = expected agreement of raters.

The annotation results were aggregated, and the interannotator Fleiss κ was computed using the Python package "statsmodels."

English and Filipino Word2vec Language Model

To train a language model, several text data gathered from various sources were used. A combination of English and Filipino texts is collected. A total of approximately 4.7 GB of data were collected and trained for this model. Textbox 3 summarizes the sources of the data used for this language model.

A word2vec model [14] was trained using the language model training data in the previous textbox as the training data set. The CBOW algorithm was used as a configuration. The model is trained with 300 features and 20 epochs with a window of 5. The window means the algorithm checks 5 words before and 5 words after the target word, which is predicted in every hidden node in the network. The algorithm outputs a word embedding, which is a 300-dimensional vector for every word included in the training data set vocabulary.

Textbox 3. Language model training data.

Filipino Texts online (News articles, Wikipedia, books, etc)

- PALITO corpus [36]
- Leipzig corpus [37]
- Isawika lexicon [38]

Twitter data set from published studies

- Yolanda data set [39]
- Gay language data set [40]
- Election data set [41]
- Hate speech data set [42]
- Emotion-annotated Tweets for Disaster Risk Assessment [EMOTERA] data set [10]

Twitter data set by web-based crawling (created by this study)

- Mental Health PH seed words
- Self-declared depression data set from 2019

Reddit data from the program which this study is a part of

- Mental health discussions on Reddit

Results

Overview

The first objective of this study is to construct a data set of Philippine Twitter users who are assessed by mental health experts. The second objective is to develop a data set of depression symptoms that can serve as a baseline for predicting depression symptoms in Filipino and English, and the third objective is to create a language model that can represent Filipino and English text through vector representations. The following sections discuss the data collected, constructed data sets for this study, and the language model results, which address the objectives of this study.

Data

A total of 75 participants were assessed by psychologists and had a complete set of data categorized at the user level, with 60 assessed with depression, 12 not assessed with depression, and 3 not assessed with depression but have other mental health conditions and are excluded from the study. Further statistics in this study will include only these 72 participants (60 assessed with depression, 12 not assessed with depression). Six of these participants assessed with depression provided 2 accounts for Twitter. In total, 78 Twitter users have 577,202 tweets; 433,029 liked tweets; and a total of 1,010,231 tweets. This is the training data set 1. Figures 4 and 5 show the data collection results including the distribution and statistics of the collected data.

Tables 1 to 3 show the count of Twitter users, considering the 6 dummy Twitter accounts from the participants. Table 1 shows the statistical summaries for the Twitter data between the depressed group and control group for tweet and retweet counts. Table 2 shows the statistical summaries between the depressed group and control group for account age and mentions count. Table 3 shows the statistical summaries between the depressed group and control group for follower and following count.

During the annotation task, an initial number of 438,718 tweets were selected and preprocessed. A total of 102,262 tweets were annotated by the first-pass annotator and from this data set, 11,335 tweets have been annotated by all 3 annotators. The final data set generated in this study is 79,614 tweets, with 11,163 labeled with depression symptoms (13 categories) and 68,451 labeled with no symptoms, and is named as data set 2.

During the validation for data set 2, there were only 3 tweets out of 68 randomly chosen validation tweets in which the psychologists disagreed with the category label, reaching 95.59% psychologist validation score. Figure 6 shows the total counts of symptom categories resulting from the annotation task.

Another validation step is used using the Fleiss κ to measure the interannotator score using the 11,335 tweets annotated by 3 annotators and 14 label categories. The agreement score is 0.735, interpreted as “substantial agreement” on the basis of Figure 7 [43] interpretation of the Fleiss κ measurement.

Figure 4. Data collection results (part 1). The distribution of depression assessment results shows the imbalance of data, as more users are assessed with depression. Regarding severity, 26 individuals were assessed with “severe” depression, whereas 20 were assessed with “mild” and 14 with “moderate” depression. The distribution of gender in our collected data is also imbalanced, with more female participants recruited for this study. For the distribution of age, most users were aged 18 to 25 years.

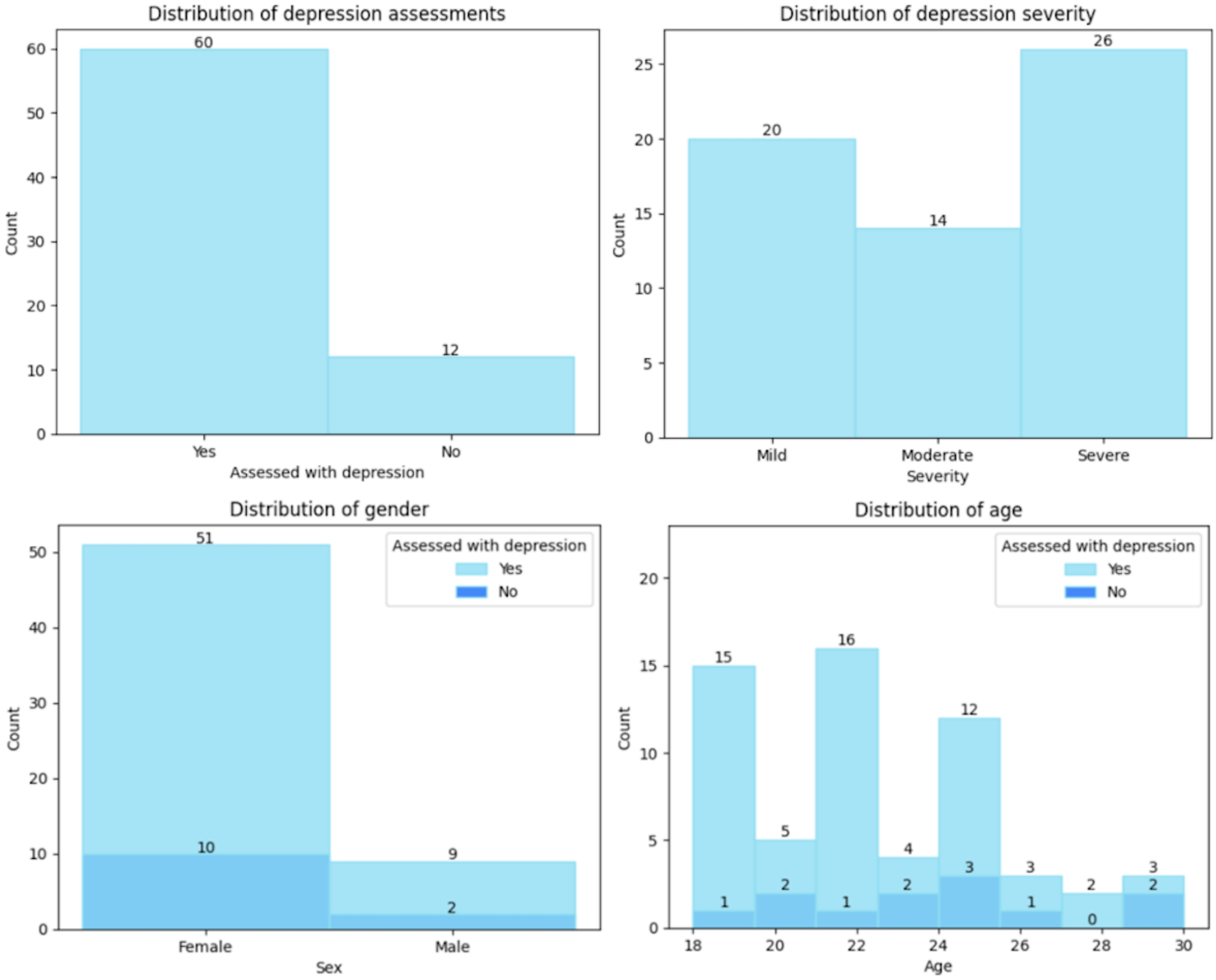


Figure 5. Data collection results (part 2). The distribution of regions shows most participants were from the National Capital Region (NCR), IV-A, and III regions and other parts of the Luzon area. In the distribution of education, most are college and high school graduates. Most participants are either employed or students as shown by the distribution of employment. CAR: Cordillera Administrative Region.

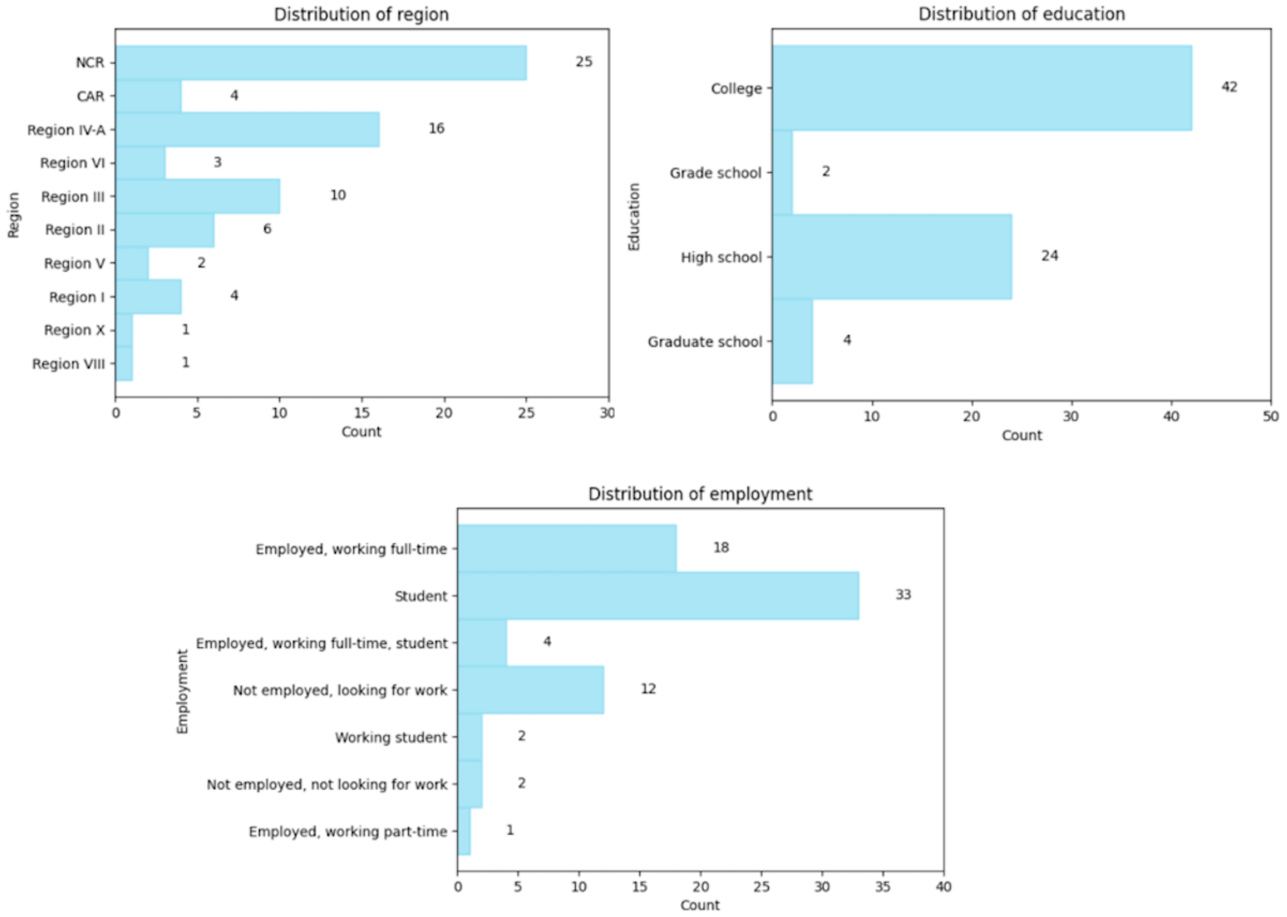


Table 1. Data collection results (part 3).

	tweet_count							retweet_count						
	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum
Assessed with depression														
No	12	3498.42 (3982.08)	80	461.5	1606.00	5477.25	10,843.00	12	313.75 (569.71)	0	5.25	29	405.75	1960.00
Yes	66	6146.03 (9198.76)	3	404.75	2313.50	6724.75	47,644.00	66	605.94 (1395.77)	0	1	30.5	232	6670.00

Table 2. Data collection results (part 4).

	tweet_age_days							mentions_count						
	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum
Assessed with depression														
No	12	2429.17 (1614.81)	599	904.5	2284.00	3564.75	5480.00	12	1674.25 (2049.69)	24	169.75	909	2448.50	6664.00
Yes	66	1780.61 (1389.65)	101	609.5	1470.50	2581.75	4860.00	66	2959.41 (5634.60)	1	145.25	853	2596.25	37,315.00

Table 3. Data collection results (part 5).

	Follower							Following						
	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum	Count	Mean (SD)	Minimum	25%	50%	75%	Maximum
Assessed with depression														
No	12	218.17 (175.73)	7	61	178	370	490	12	330 (364.07)	53	113.5	174	395.5	1,169.00
Yes	66	252.53 (380.18)	0	29	98.5	276.75	1,839.00	66	321.65 (393.44)	0	64	156	406.75	1,787.00

Figure 6. Annotation task results. Total counts per symptom category for all the 11,163 tweets labeled with depression symptoms and their corresponding symptom categories are shown here. Social, PHQ-9-emotional: worthlessness and worry, PHQ-9-emotional: depressed and sadness have the most number of tweets, whereas Physical: movement and Substance use and PHQ-9-suicidal tendencies have the least number of tweets. MH: mental health; PHQ-9: Patient Health Questionnaire-9.

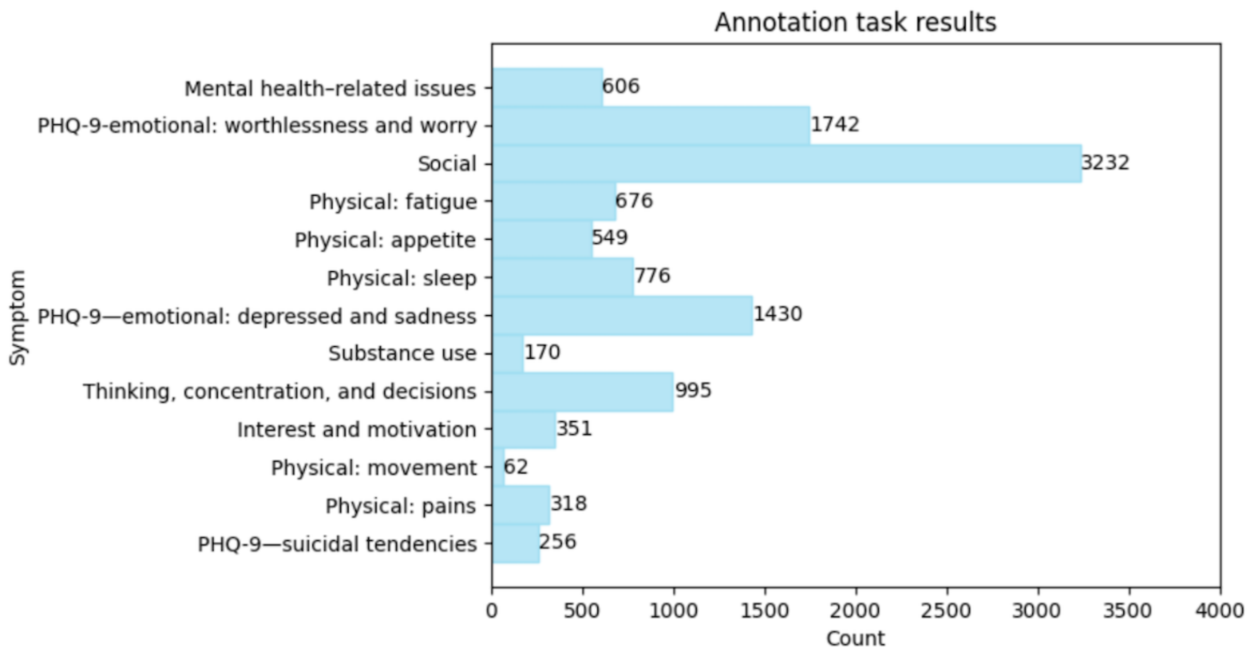
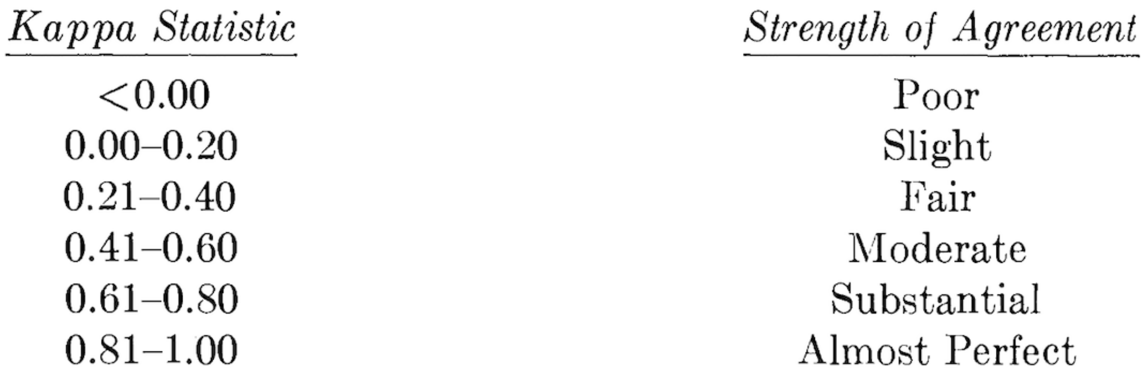


Figure 7. Interpretation of Fleiss κ. The resulting Fleiss κ is 0.735, denoting the annotation task agreement score as “substantial agreement”. Used with permission of John Wiley & Sons - Books, from “The measurement of observer agreement for categorical data. Landis JR, Koch GG, 33(1):159-174, Mar 1977” [43]; permission conveyed through Copyright Clearance Center, Inc.”.



Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Constructed Data Sets

From all 4 types of data collected in this study, 2 labeling mechanisms are applied: labeled data at a user level and tweet level. The data set of user-level annotated data is a combination of survey data, PHQ-9 questionnaire results, and Mental Health Assessment results (data set 1). The other tweet-level annotated data set was created from the previous annotation tasks from the Twitter archive data of participants (data set 2). Both data sets share the same number of Twitter users (78 users).

A complete list of data per data set is provided in Textbox 4, and the summary is shown in Table 4.

Data sets collected and processed in this study can greatly help both the mental health research and the Filipino language processing research in the country.

We have created a data set that includes demographic data, PHQ-9 questionnaire data, and Mental Health Assessment data. It is important to mention that all participants in this data collection were assessed by psychologists via clinical interviews. These data can be used to explore user-level depression detection of users in social media.

The annotated training data set created by this study is also a good starting point for exploring the development of depression symptom detection models because annotation rules were constructed with mental health experts and processed by 3 annotators, with guidance and validation from the experts as well. The interannotation agreement score achieved a Fleiss κ score of 0.735, which is interpreted as “substantial agreement.” This data set has also reached a 95.59% psychologist validation score on the validation data set.

Textbox 4. Data set information.

Survey data (data set 1)

- Age
- Sex
- Region
- Education
- Employment
- MentalHealth_SelfRating
- MentalHealth_Services

Patient Health Questionnaire-9 (PHQ-9) data (data set 1)

- PHQ-9 score

Mental health assessment data (data set 1)

- Symptom_ThinkingConcentrationDecisions
- Symptom_InterestMotivation
- Symptom_Sleep
- Symptom_Fatigue
- Symptom_Appetite
- Symptom_Movement
- Symptom_Substance
- Symptom_SuicidalTendencies
- Symptom_DepressedSadness
- Symptom_WorthlessnessWorry
- Symptoms_Pain
- Symptoms_Social
- Symptoms
- Medical history
- Psychiatric History (current)
- Psychiatric History (past)
- History of Mental Ilness in Family
- History of Family Illness
- Religion
- Sexual Orientation
- Substance Use History
- Other providers
- Self Harm
- Suicidal Behavior
- Final Assessment
- Assessed with Depression
- Severity
- Assessment time (mins)

Twitter data (data set 1)

- tweet_daily_mean
- tweet_daily_max tweet_min_dt

- tweet_max_dt
- is_retweet
- is_reply
- mentions_count
- tweet_count
- retweet_count
- favorite_count
- tweet_age_days
- isretweet_percentage
- isreply_percentage
- isowntweet
- isowntweet_percentage
- createdAt
- blocked
- follower
- following
- mute
- lists_created
- lists_member
- lists_subscribed
- protectedHistory
- twitterCircle

Twitter data (data set 2)

- id
- full_text_cleaned
- Final Symptom Category 1
- Final Symptom Category 2
- Final Symptom Category 3

Table 4. Data set summaries.

Raw data	Data set 1: User-level depressed vs nondepressed	Data set 2: Tweet-level depression symptom category
Survey data	✓	—[a]
PHQ-9[b]	✓	—
Mental health assessment	✓	—
Twitter data archive	✓ (summarized—means, averages, etc)	✓ (annotated)
Summary	72 users (78 Twitter users): 60 assessed with depression; 12 not assessed with depression	78 Twitter users (6 dummy accounts); 11,163 tweets (depression symptom); 68,451 tweets (no symptom)

[a] Not applicable.

[b] PHQ-9: Patient Health Questionnaire-9.

Discussion

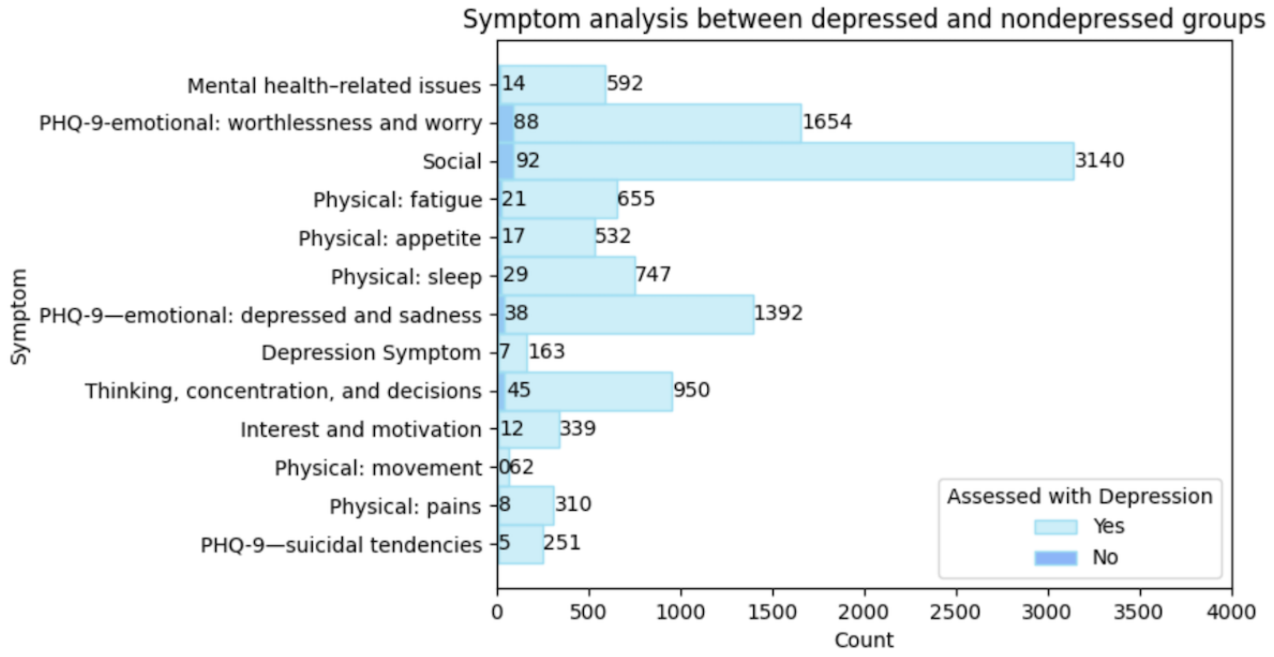
Depression Symptom Between Depressed and Control Groups

Data set 2 (tweet-level depression symptom category) was created by annotating depression symptoms of anonymized data from both the depressed and nondepressed groups. By doing this, we remove the bias that these symptoms can only be experienced by the depressed group. Depression still faces many challenges with the diagnosis because the symptoms also occur in other conditions, and these symptoms can be experienced by nondepressed users all the same. Every person experiences a lack of sleep or common emotions such as sadness, worry, disappointment, rage, and physical symptoms of headache and fatigue every now and then.

Figure 8 shows this comparison of the amount of depression symptoms recorded for depressed and nondepressed groups. Further studies can be done on analyzing these depression symptom patterns and comparing the differences in how both groups experience these symptoms and how these symptoms are revealed through language use.

It is interesting to note that some symptoms of PHQ-9-suicidal tendencies, are also detected in the nondepressed group. These are tweets containing talks about death, for example, “if i die i die i do not even care anymore” or “i just want to be dead face with tongue xxlink.”

Figure 8. Symptom analysis between depressed and nondepressed groups. This figure shows the number of symptoms recorded for both groups, as depression symptoms occur for both but differ in intensity and frequency. Social symptoms are the most recorded symptom being experienced by the nondepression group, which may include sensitivity to criticism, irritability, and anger. The next common symptom for the nondepression group was the PHQ-9-emotional: worthlessness and worry, which includes texts about daily worries such as school and work-related problems. MH: mental health; PHQ-9: Patient Health Questionnaire-9.



Word2vec Language Model Results

A language model trained with ample data can be a powerful tool if it can capture semantic relationships between words. Because word embedding features can capture a word’s representation in several dimensions, the more it can capture these features, the better results we can achieve when training our models on downstream tasks such as classification and prediction tasks. Figure 9 illustrates a sample word vector output from this algorithm.

The language model created is a combination of Wikipedia articles, news articles, books, and data sets specific to web-based language, such as Twitter and Reddit data sets. There are also depression-specific topics included in these data sets to better model the depression patterns in natural language in social media. Finally, English and Filipino are both included in the data sets, with purely English text, Filipino texts, and a combination of the two. This is to model both languages in a single vector space, because “Taglish” is the common language used on the web, with direct code switching between words in both languages.

To check the language model’s ability to capture both languages, the cluster embeddings of words and how they look in a 2D space are presented subsequently. To plot these words, each word is transformed into its 300-dimensional vector and is fit into a KMeans model of ≥2 clusters. In KMeans, data points are grouped together based on certain similarities. Centroids are determined, which are representations of the location of the center of each cluster, and observations are assigned to each cluster with the nearest mean. In this test case, we check if certain words belong to a cluster, specifically comparing if the language model can determine words from symptoms and cluster them correctly. Figure 10 illustrates our first test case to show word relationships.

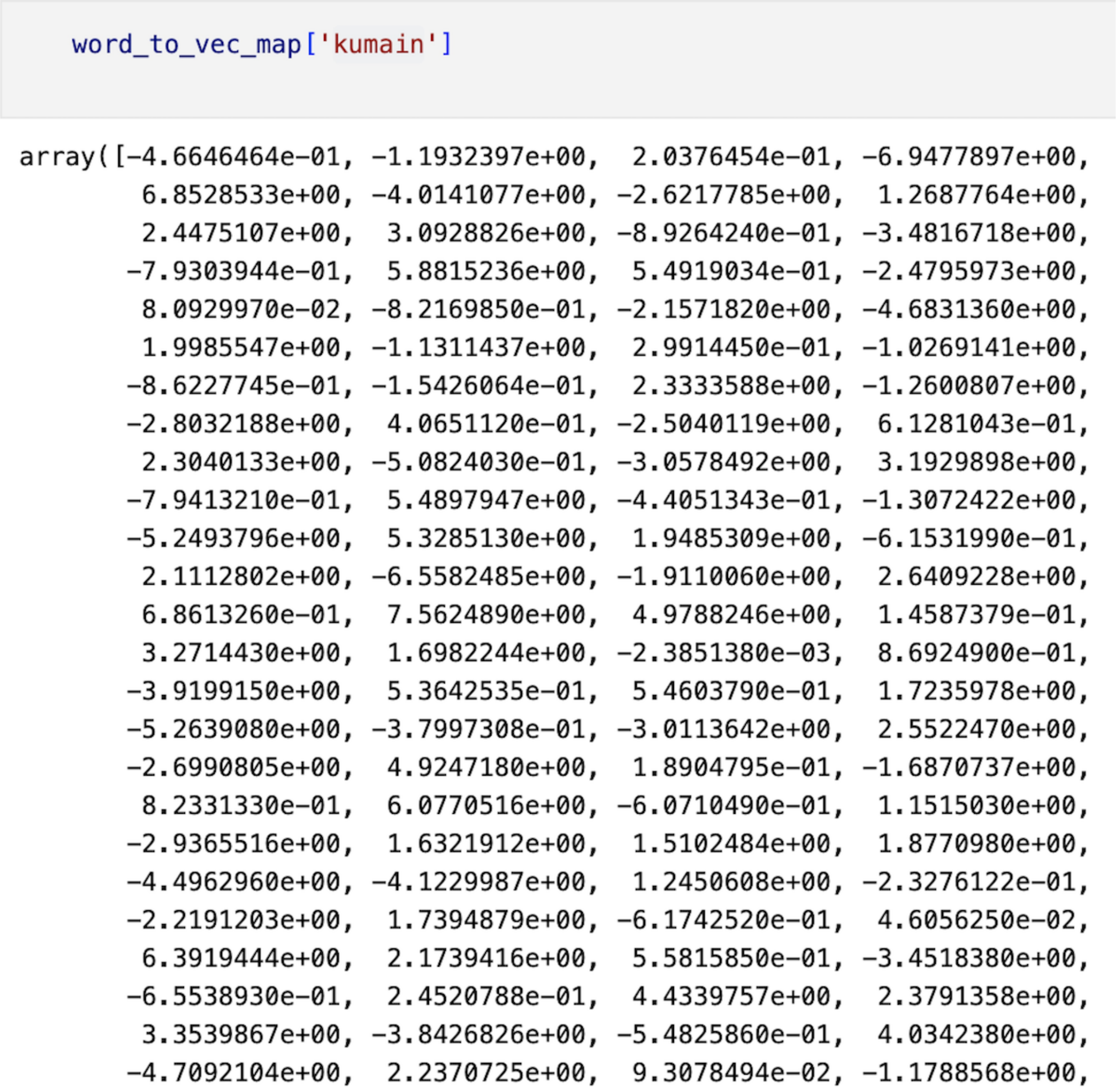
We also look at semantic relationships captured by the language model. Several examples below show word analogies using cosine similarities between word vectors. Neighboring words are words that are most similar to a target word in the vocabulary on the basis of this cosine similarity. Positive words contribute positively toward similarity, and negative words contribute negatively. Figure 11 illustrates our second test case to show similar words.

Looking at the analogies captured by the language model using cosine similarities to compute positive and negative word relationships through word vectors, Figure 12 shows examples of correctly captured relationships by the language model.

A language model can never fully capture a language as it can only learn what it has been trained on. Some aspects may not be captured, such as examples in Figure 13. There are also some issues with biases, for example, with gender bias in the training data, as shown in Figure 13.

The Filipino and English language models developed in this study can be used for research on other downstream tasks such as classification, labeling, and named entity recognition tasks. It can also be used on non–mental health research tasks and is best for tasks focusing on the Filipino and English languages in social media data.

Figure 9. Word2vec vector embeddings: the word vector output for the word “kumain” or in English “ate.” Each dimension represents an extracted feature of the text data that is extracted from this unsupervised learning algorithm.



```
word_to_vec_map['kumain']
```

```
array([-4.6646464e-01, -1.1932397e+00,  2.0376454e-01, -6.9477897e+00,
        6.8528533e+00, -4.0141077e+00, -2.6217785e+00,  1.2687764e+00,
        2.4475107e+00,  3.0928826e+00, -8.9264240e-01, -3.4816718e+00,
       -7.9303944e-01,  5.8815236e+00,  5.4919034e-01, -2.4795973e+00,
        8.0929970e-02, -8.2169850e-01, -2.1571820e+00, -4.6831360e+00,
        1.9985547e+00, -1.1311437e+00,  2.9914450e-01, -1.0269141e+00,
       -8.6227745e-01, -1.5426064e-01,  2.3333588e+00, -1.2600807e+00,
       -2.8032188e+00,  4.0651120e-01, -2.5040119e+00,  6.1281043e-01,
        2.3040133e+00, -5.0824030e-01, -3.0578492e+00,  3.1929898e+00,
       -7.9413210e-01,  5.4897947e+00, -4.4051343e-01, -1.3072422e+00,
       -5.2493796e+00,  5.3285130e+00,  1.9485309e+00, -6.1531990e-01,
        2.1112802e+00, -6.5582485e+00, -1.9110060e+00,  2.6409228e+00,
        6.8613260e-01,  7.5624890e+00,  4.9788246e+00,  1.4587379e-01,
        3.2714430e+00,  1.6982244e+00, -2.3851380e-03,  8.6924900e-01,
       -3.9199150e+00,  5.3642535e-01,  5.4603790e-01,  1.7235978e+00,
       -5.2639080e+00, -3.7997308e-01, -3.0113642e+00,  2.5522470e+00,
       -2.6990805e+00,  4.9247180e+00,  1.8904795e-01, -1.6870737e+00,
        8.2331330e-01,  6.0770516e+00, -6.0710490e-01,  1.1515030e+00,
       -2.9365516e+00,  1.6321912e+00,  1.5102484e+00,  1.8770980e+00,
       -4.4962960e+00, -4.1229987e+00,  1.2450608e+00, -2.3276122e-01,
       -2.2191203e+00,  1.7394879e+00, -6.1742520e-01,  4.6056250e-02,
        6.3919444e+00,  2.1739416e+00,  5.5815850e-01, -3.4518380e+00,
       -6.5538930e-01,  2.4520788e-01,  4.4339757e+00,  2.3791358e+00,
        3.3539867e+00, -3.8426826e+00, -5.4825860e-01,  4.0342380e+00,
       -4.7092104e+00,  2.2370725e+00,  9.3078494e-02, -1.1788568e+00,
```

Figure 10. Language model word symptom relationships. The first figure shows the appetite versus sadness symptoms, with test cases “craving, kain (eat), breakfast, lungkot (sadness), haist (expression of sigh), sad,” which are combinations of English and Filipino words. We notice here that they are also a combination of expressions (haist), emotions (craving, lungkot, and sad), or action words (kain). These examples show that these are correctly clustered into appetite and sadness clusters. The second figure shows suicidal tendency–related words “kill, suicide, death, patayin (kill)” are clustered together versus sadness words “sad, lungkot, lugmok (sunken).” In addition, the third figure shows mental health issue–related words “psychologist, therapy, and trauma” are clustered together versus movement-related words “tamad (lazy), galaw (move), kama (bed).”.

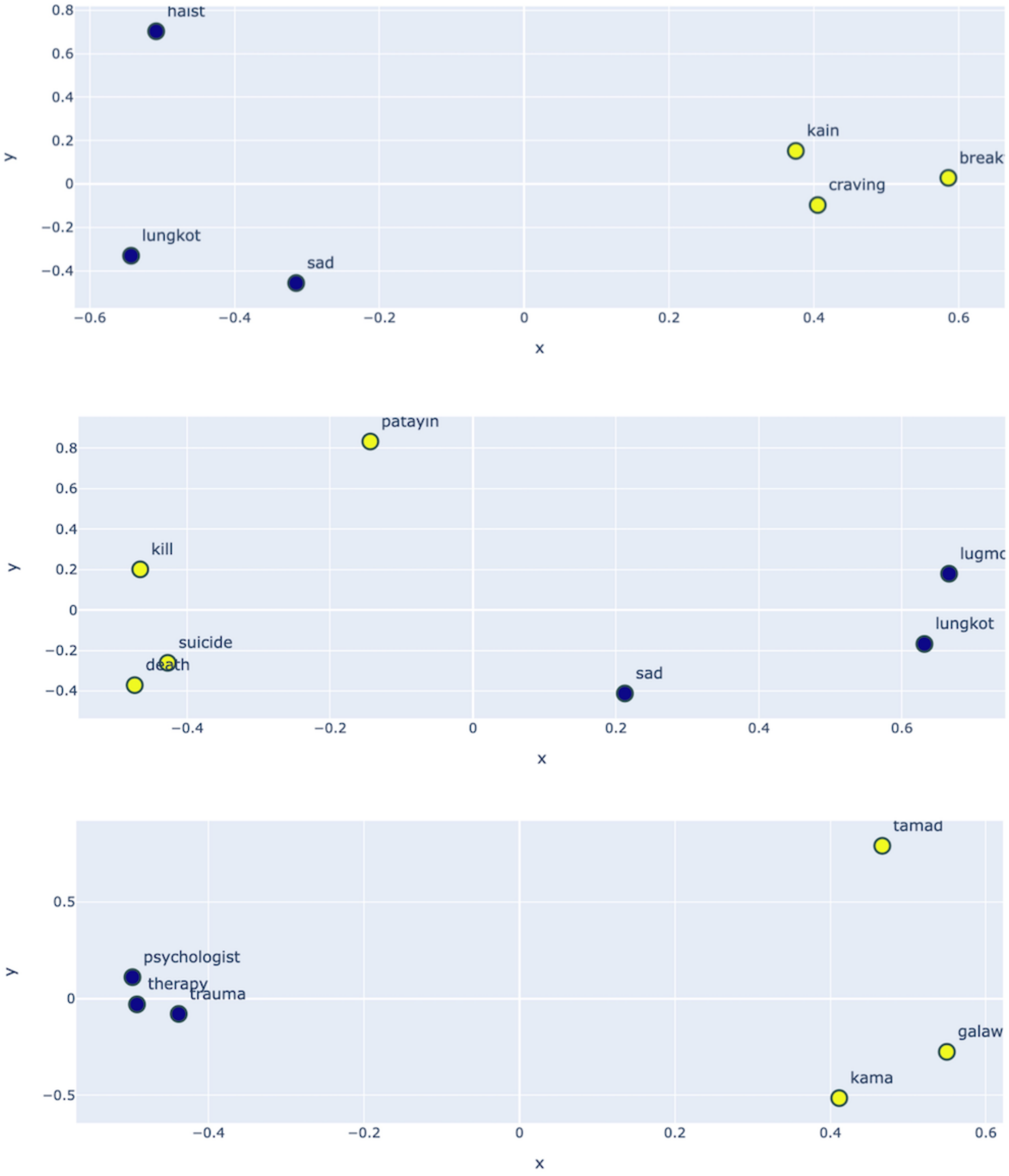
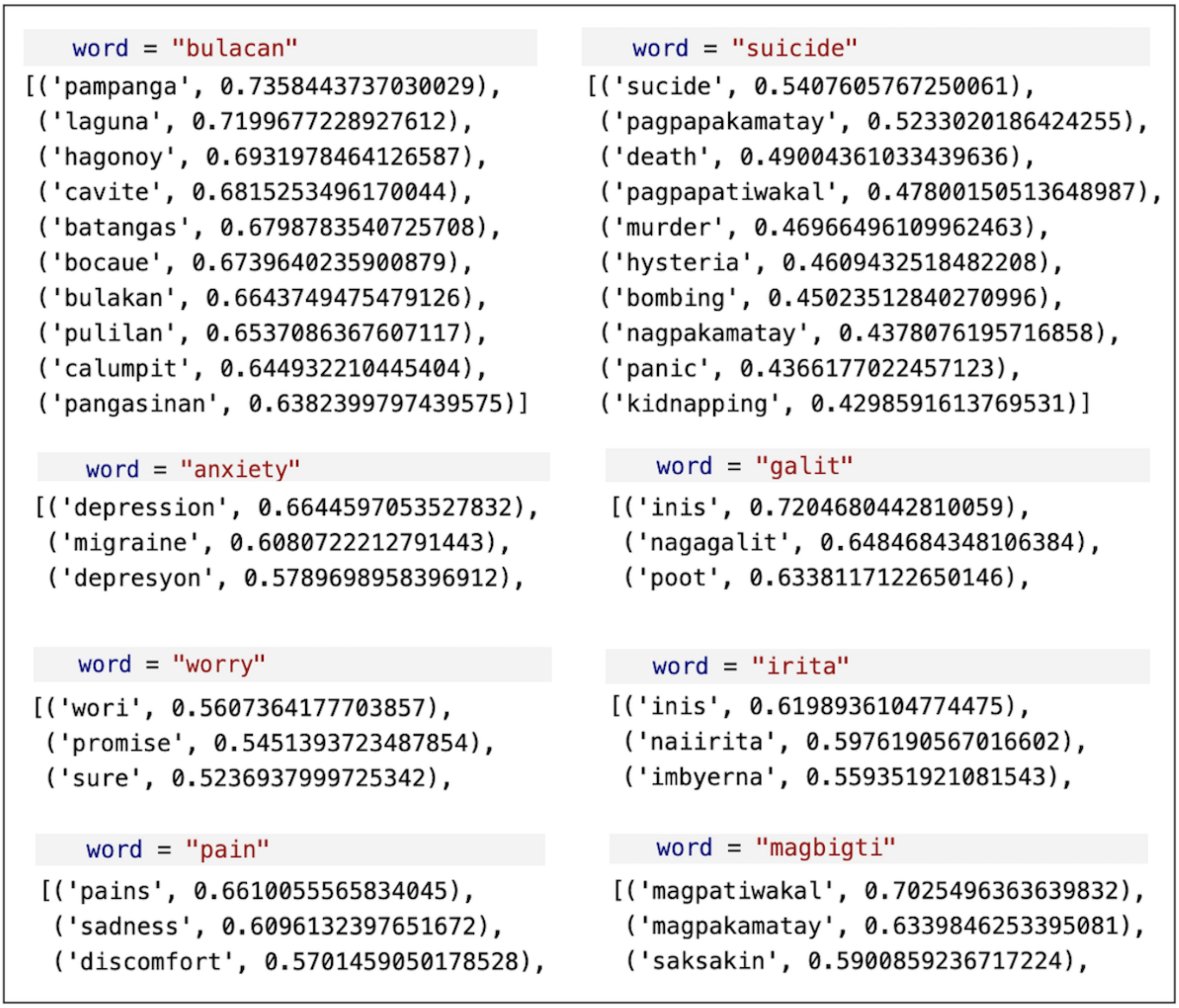


Figure 11. Language models for similar words. From this figure, we can see that the most similar words for “Bulacan,” which is a province in the central Luzon region in the Philippines, has similar words “Pampanga, Laguna, Hagonoy, Calumpit, etc” which are also places (provinces, municipalities), Pampanga, Hagonoy, and Calumpit being geographically in close boundaries with Bulacan, and “Bocaue, Pulilan, Hangonoy” being towns and municipalities in Bulacan. We also see the word neighbors for “suicide,” which includes “sucide,” which is a common wrong spelling for “suicide,” “magpakamatay,” which is the Filipino translation, and other “suicide” related words in both languages. We can also see “bombing and kidnapping,” which are common words seen with suicide as per news articles such as “suicide bombing.” We also see the words mostly associated with symptoms such as “anxiety, worry, pain, galit (mad), irita (irritated), and magbigti (hang).”.



```
   word = "bulacan"
[('pampanga', 0.7358443737030029),
 ('laguna', 0.7199677228927612),
 ('hagonoy', 0.6931978464126587),
 ('cavite', 0.6815253496170044),
 ('batangas', 0.6798783540725708),
 ('bocaue', 0.6739640235900879),
 ('bulakan', 0.6643749475479126),
 ('pulilan', 0.6537086367607117),
 ('calumpit', 0.644932210445404),
 ('pangasinan', 0.6382399797439575)]
```

```
   word = "anxiety"
[('depression', 0.6644597053527832),
 ('migraine', 0.6080722212791443),
 ('depresyon', 0.5789698958396912),
```

```
   word = "worry"
[('wori', 0.5607364177703857),
 ('promise', 0.5451393723487854),
 ('sure', 0.5236937999725342),
```

```
   word = "pain"
[('pains', 0.6610055565834045),
 ('sadness', 0.6096132397651672),
 ('discomfort', 0.5701459050178528),
```

```
   word = "suicide"
[('sucide', 0.5407605767250061),
 ('pagpapakamatay', 0.5233020186424255),
 ('death', 0.4900436103343963 6),
 ('pagpapatiwakal', 0.47800150513648987),
 ('murder', 0.46966496109962463),
 ('hysteria', 0.4609432518482208),
 ('bombing', 0.45023512840270996),
 ('nagpakamatay', 0.4378076195716858),
 ('panic', 0.4366177022457123),
 ('kidnapping', 0.4298591613769531)]
```

```
   word = "galit"
[('inis', 0.7204680442810059),
 ('nagagalit', 0.6484684348106384),
 ('poot', 0.6338117122650146),
```

```
   word = "irita"
[('inis', 0.6198936104774475),
 ('naiirita', 0.5976190567016602),
 ('imbyerna', 0.559351921081543),
```

```
   word = "magbigti"
[('magpatiwakal', 0.7025496363639832),
 ('magpakamatay', 0.6339846253395081),
 ('saksakin', 0.5900859236717224),
```

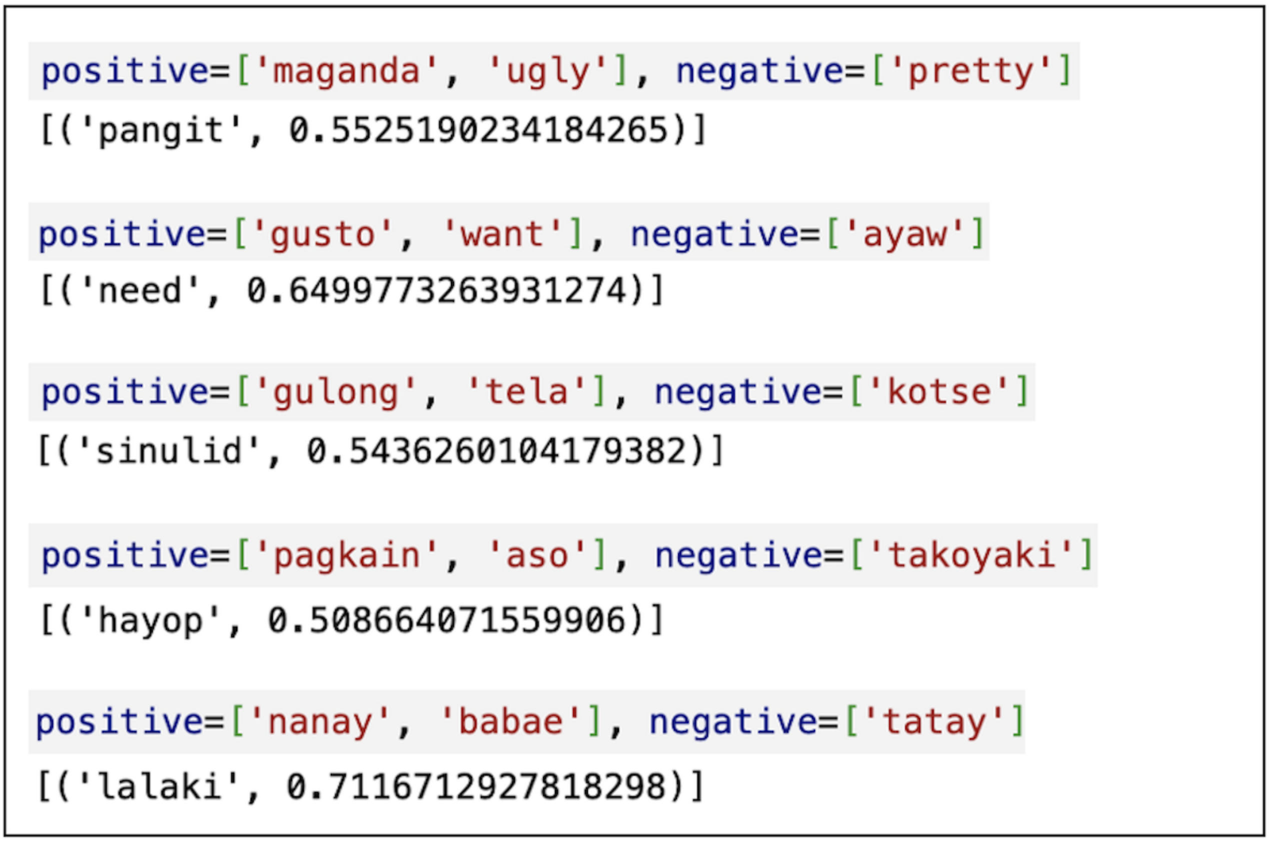


Figure 12. Language model analogies. Synonyms were captured, such as “maganda” (beautiful) is to “pretty” and “ugly” is to “pangit” (ugly). This is another example of the model capturing word similarities in Filipino and English code switching. Antonyms were also captured, “gusto” (want) is to “ayaw” (do not want) and “want” is to “need.” In addition, part-whole and superclass were captured, “gulong” (tire) is to “tela” (fabric) and “kotse” (car) is to “sinulid” (thread). Other examples captured: “pagkain” (food) is to “aso” (dog), “takoyaki” (octopus balls) is to “hayop” (animal), “Nanay” (mother) is to “babae” (woman), and “tatay” (father) is to “lalaki” (man).

```
positive=['maganda', 'ugly'], negative=['pretty']
[('pangit', 0.5525190234184265)]

positive=['gusto', 'want'], negative=['ayaw']
[('need', 0.6499773263931274)]

positive=['gulong', 'tela'], negative=['kotse']
[('sinulid', 0.5436260104179382)]

positive=['pagkain', 'aso'], negative=['takoyaki']
[('hayop', 0.508664071559906)]

positive=['nanay', 'babae'], negative=['tatay']
[('lalaki', 0.7116712927818298)]
```

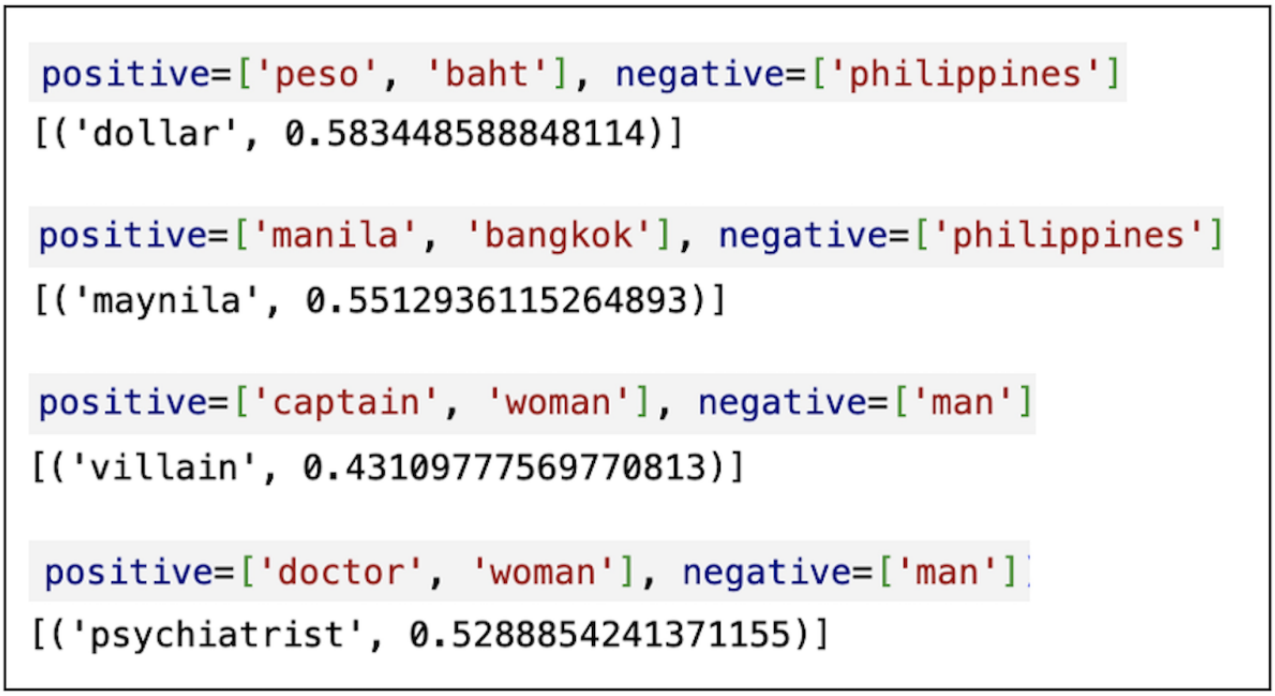


Figure 13. Language model faults and biases: the first 2 examples show some faults with currency and geography, which gave incorrect results. Gender biases are also shown in the last 2 examples; if “captain” is to “man,” the model results in a “woman” as a “villain.” In addition, if “doctor” is to “man,” the model outputs “woman” is to “psychiatrist.”.

```
positive=['peso', 'baht'], negative=['philippines']
[('dollar', 0.583448588848114)]

positive=['manila', 'bangkok'], negative=['philippines']
[('maynila', 0.5512936115264893)]

positive=['captain', 'woman'], negative=['man']
[('villain', 0.4310977569770813)]

positive=['doctor', 'woman'], negative=['man']
[('psychiatrist', 0.5288854241371155)]
```

Data Set Limitations

Age Group Limitation

The population included in this study focuses on the young adult population (aged 18-30 years). The created data set may not be generalizable to other population samples as the language and social media use of individuals may be different within different age demographics.

Tweet Symptom Annotation Limit

In this study, a maximum limit of 3 symptoms per tweet is enforced as part of the annotation guideline. This is aimed at simplicity because of the complex nature of the annotation task. Aside from the number of tweets to be annotated, several depression categories are also considered, which proved quite challenging for the annotators. For binary label classification tasks (depression symptom vs nondepression symptom), this limitation may be ignored as the presence of at least one symptom is considered a depression symptom label. Multilabel classification tasks (14 symptom depression categories) should be considered as a limitation of the created data set, as some symptoms may not be included in some tweets.

Data Imbalance

This study ensures the validity of depression screening for each participant, requiring multiple data collection workflow steps and the participation of both participants and mental health experts. Because of the limited slots for expert assessment, the data collected shows an imbalance between the participants for the depressed and control groups. This data set imbalance might be worth noting for user-level depression detection studies, but because the goal of this study is to create a tweet-level annotation of depression and not on a user level, the data imbalance on the user level is insignificant. However, it is useful that all participants were assessed by experts, so the results of all tweet-level analyses can be routed back or compared with user-level analyses, which unfortunately is not a scope of this paper but can be explored in the future.

Another class imbalance in this data set is the distribution of sex. In this data set, 85% are female participants (61 female and 11 male participants). This may be attributed to the higher prevalence of depression in female individuals among young adults in the Philippines [33], while some similar studies on web-based mental health intervention show similar patterns of gender imbalance where participants are mostly female [44]. These data set limitations should be noted when doing data analysis on the resulting data set.

Conclusions and Future Work

Valid data sets need to be constructed to develop solutions to identify depression patterns through NLP and machine learning. This study aimed to help in the area of depression research through the construction of depression data sets from social media to aid NLP in the Philippine setting. The proposed process included interdisciplinary methods between psychology and data science methods, implementing clinical screening methods with the help of psychologists. A total of 72 participants were assessed by psychologists and provided their Twitter data, with 60 assessed with depression and 12 not assessed with depression. A baseline data set of depression symptoms in tweets was created by manual annotation in a process constructed, guided, and validated by psychologists. In total, 13 depression categories and 1 no symptom category were identified during this process. This annotation process was done by 3 annotators, accomplishing a substantial interannotator agreement score of Fleiss κ of 0.735 and a 95.59% psychologist validation score. From this task, a total of 79,614 depression symptom–annotated tweets are created. A language model using the word2vec algorithm is also created, which can represent text into numbered vectors and can be used in various machine learning techniques for NLP. This study created several validated data sets that can be used further to enhance depression research in the Philippines.

For future research, 4 directions can be explored. (1) This study suggests future research to apply the methodologies performed in this study to improve the class imbalance and size of the data sets. The data sets can be further expanded to collect more users and Twitter data, and the annotation process of detection symptoms can be expanded into more tweets. (2) The data sets in this study can be used to explore the creation of depression symptom detection models in social media data using machine learning techniques on a tweet-level detection. (3) The data sets in this study can also be used to explore the creation of user-level depression detection using social media behavior data, PHQ-9 data, and demographic data. (4) Finally, depression symptom patterns can be explored further in depth between the depression and nondepression groups in this data set, as both groups contribute to the depression symptoms data set, and both show some level of depression symptoms.

Acknowledgments

The authors are grateful to all participants for their interest and contribution, who believe this research will open doors to mental health research opportunities in the country. Furthermore, the authors would like to thank reviewers for their support and influence in shaping our methods and results and #MentalHealthPH and other clinics and organizations for helping them in participant recruitment. This work was funded by the University of the Philippines (UP) System Emerging Inter-Disciplinary Research Program (OVPAA-EIDR-C09-07).

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to the sensitive nature of the respondents in this study, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017 Jan 3. URL: https://www.who.int/publications/i/item/depression-global-health-estimates [accessed 2024-07-28]
2. Pandemic year sees 57% rise in suicide rate in Philippines. Philstar Global. 2021 Jul 6. URL: https://www.philstar.com/headlines/2021/07/06/2110596/pandemic-year-sees-57-rise-suicide-rate-philippines [accessed 2024-07-28]
3. Gowen K, Deschaine M, Gruttadara D, Markey D. Young adults with mental health conditions and social networking websites: seeking tools to build community. Psychiatr Rehabil J 2012;35(3):245-250. [doi: 10.2975/35.3.2012.245.250] [Medline: 22246123]
4. Berger M, Wagner TH, Baker LC. Internet use and stigmatized illness. Soc Sci Med 2005 Oct;61(8):1821-1827. [doi: 10.1016/j.socscimed.2005.03.025] [Medline: 16029778]
5. Park M, McDonald D, Cha M. Perception differences between the depressed and non-depressed users in Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media. 2013 Presented at: ICWSM-13; July 8-11, 2013; Cambridge, MA. [doi: 10.1609/icwsm.v7i1.14425]
6. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 2017 Dec;18:43-49 [FREE Full text] [doi: 10.1016/j.cobeha.2017.07.005]
7. Eberhard DM, Simons GF, Fennig CD. Ethnologue: languages of the world. Twenty-seventh edition. SIL International. 2024. URL: https://www.ethnologue.com/ethnoblog/welcome-27th-edition/ [accessed 2024-08-03]
8. Andrei AL. Development and evaluation of Tagalog linguistic inquiry and word count (LIWC) dictionaries for negative and positive emotion. The MITRE Corporation. 2014 Dec 23. URL: https://www.mitre.org/sites/default/files/publications/pr_14-3858-development-evaluation-of-tagalog-linguistic-inquiry.pdf [accessed 2024-07-29]
9. Keen D, King NM, Lopez JL, Mondares A, Ponay C. FilCon: Filipino sentiment lexicon generation using word level-annotated dictionary-based and corpus-based cross lingual approach. In: Proceedings of the 9th International Conference of Asian Association of Lexicography. 2015 Presented at: ASIALEX 2015; June 25-27, 2015; Hong Kong, China.
10. Lapita FR, Batista-Navarro RT, Albacea E. Crowdsourcing-based annotation of emotions in Filipino and English tweets. In: Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing. 2016 Presented at: WSSANLP2016; December 11, 2016; Osaka, Japan.
11. Katchapakirin K, Wongpatikaseree K, Yomabootand P, Kaewpitakkun Y. Facebook social media for depression detection in the Thai community. In: Proceedings of the 15th International Joint Conference on Computer Science and Software Engineering. 2018 Presented at: JCSSE 2018; July 11-13, 2018; Nakhonpathom, Thailand. [doi: 10.1109/jcsse.2018.8457362]
12. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, et al. Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017 Presented at: IJCAI'17; August 19-25, 2017; Melbourne, Australia. [doi: 10.24963/ijcai.2017/536]
13. Cornn K. Identifying depression on social media. Stanford University. URL: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15712307.pdf [accessed 2024-07-29]
14. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, NV.
15. Rosa RL, Schwartz GM, Ruggiero WV, Rodriguez DZ. A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Trans Industr Inform 2019 Apr;15(4):2124-2135. [doi: 10.1109/tii.2018.2867174]
16. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In: Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2013 Presented at: PAKDD 2013; April 14-17, 2013; Gold Coast, Australia. [doi: 10.1007/978-3-642-40319-4_18]
17. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP 2014; October 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
18. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. 2013 Presented at: ICWSM-13; July 8–11, 2013; Cambridge, MA. [doi: 10.1609/icwsm.v7i1.14432]
19. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from Twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702280]
20. Shen T, Jia J, Shen G, Feng F, He X, Luan H, et al. Cross-domain depression detection via harvesting social media. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018 Presented at: IJCAI'18; July 13-19, 2018; Stockholm, Sweden. [doi: 10.24963/ijcai.2018/223]

21. Kabir M, Ahmed T, Hasan MB, Laskar MT, Joarder TK, Mahmud H, et al. DEPTWEET: a typology for social media texts to detect depression severities. Comput Hum Behav 2023 Feb;139:107503 [FREE Full text] [doi: 10.1016/j.chb.2022.107503]
22. Losada DE, Crestani F, Parapar J. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: Proceedings of the 8th International Conference of the CLEF Association. 2017 Presented at: CLEF 2017; September 11-14, 2017; Dublin, Ireland. [doi: 10.1007/978-3-319-65813-1_30]
23. Adarsh V, Arun Kumar P, Lavanya V, Gangadharan GR. Fair and explainable depression detection in social media. Inf Process Manag 2023 Jan;60(1):103168 [FREE Full text] [doi: 10.1016/j.ipm.2022.103168]
24. Wang Y, Wang Z, Li C, Zhang Y, Wang H. Online social network individual depression detection using a multitask heterogenous modality fusion approach. Inf Sci 2022 Sep;609:727-749 [FREE Full text] [doi: 10.1016/j.ins.2022.07.109]
25. Li Z, An Z, Cheng W, Zhou J, Zheng F, Hu B. MHA: a multimodal hierarchical attention model for depression detection in social media. Health Inf Sci Syst 2023 Dec 18;11(1):6 [FREE Full text] [doi: 10.1007/s13755-022-00197-5] [Medline: 36660408]
26. Losada DE, Crestani F, Parapar J. Overview of eRisk 2020: early risk prediction on the internet. In: Proceedings of the Conference and Labs of the Evaluation Forum. 2020 Presented at: CLEF 2020; September 22-25, 2020; Thessaloniki, Greece. [doi: 10.1007/978-3-030-58219-7_20]
27. Aliman GB, Nivera TF, Olazo JC, Ramos DJ, Sanchez CD, Amado TM, et al. Sentiment analysis using logistic regression. J Comput Innov Eng Appl 2022 Jul:35-40 [FREE Full text]
28. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk 2022: early risk prediction on the internet. In: Proceedings of the Conference and Labs of the Evaluation Forum. 2022 Presented at: CLEF 2022; September 5-8, 2022; Bologna, Italy. [doi: 10.1007/978-3-031-13643-6_18]
29. Borra A, Pease A, Edita R, Roxas O, Dita S. Introducing Filipino wordnet. In: Proceedings of the 5th Global WordNet Conference. 2010 Presented at: GWC2010; January 31-February 4, 2010; Mumbai, India.
30. Bitsch J, Ramos R, Ix T, Ferrer-Cheng PG, Wehrle K. Psychologist in a pocket: towards depression screening on mobile phones. Stud Health Technol Inform 2015;211:153-159. [Medline: 25980862]
31. Nartia EJ, Paragas JR, Pascual N. Detection of students' mental health status: a decision support system. In: Proceedings of the 3rd International Conference on Research and Academic Community Services. 2021 Presented at: ICRACOS 2021; October 9-10, 2021; Surabaya, Indonesia. [doi: 10.1109/icracos53680.2021.9701996]
32. Aperocho MD. Philippine English in online depressive language. Psychol Educ Multidiscip J 2022;4(3):1-12. [doi: 10.5281/zenodo.7065807]
33. Puyat JH, Gastardo-Conaco MC, Natividad J, Banal MA. Depressive symptoms among young adults in the Philippines: results from a nationwide cross-sectional survey. J Affect Disord Rep 2021 Jan;3:100073 [FREE Full text] [doi: 10.1016/j.jadr.2020.100073]
34. Hofmeijer-Sevink MK, Batelaan NM, van Megen HJ, Penninx BW, Cath DC, van den Hout MA, et al. Clinical relevance of comorbidity in anxiety disorders: a report from the Netherlands Study of Depression and Anxiety (NESDA). J Affect Disord 2012 Mar;137(1-3):106-112 [FREE Full text] [doi: 10.1016/j.jad.2011.12.008] [Medline: 22240085]
35. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382 [FREE Full text] [doi: 10.1037/h0031619]
36. Dita SN, Roxas RE, Inventado P. Building online corpora of Philippine languages. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation. 2009 Presented at: PACLIC 2009; December 3-5, 2009; Hong Kong, China.
37. Leipzig corpora collection - Tagalog. Leipzig University. 2017. URL: https://corpora.wortschatz-leipzig.de/en?corpusId=tgl_community_2017 [accessed 2024-07-29]
38. Roxas RE, Borra A. Computational linguistics research on Philippine languages. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. 2000 Presented at: ACL '00; October 3-6, 2000; Hong Kong, China. [doi: 10.3115/1075218.1075292]
39. Soriano CR, Roldan MD, Cheng C, Oco N. Social media and civic engagement during calamities: the case of Twitter use during typhoon Yolanda. Philipp Polit Sci J 2016 Feb 26;37(1):6-25. [doi: 10.1080/01154451.2016.1146486]
40. Oco N, Fajutagana R, Lim CM, Miñon JD, Morano JA, Tinoco RC. Witchebelles anata magcharot kay mudra na nagsusuba si akech: developing a rule-based unidirectional beki lingo to Filipino translator. J Sci Technol Arts Res 2015;1(1):29-37 [FREE Full text]
41. Pablo ZC, Oco N, Roldan MD, Cheng C, Roxas RE. Toward an enriched understanding of factors influencing Filipino behavior during elections through the analysis of Twitter data. Philipp Polit Sci J 2014 Nov 27;35(2):203-224. [doi: 10.1080/01154451.2014.964794]
42. Cabasag NV, Chan VP, Lim SC, Gonzales ME, Cheng CK. Hate speech in Philippine election-related tweets: automatic detection and classification using natural language processing. Philipp Comput J 2019 Aug;XIV(1):1-14.
43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]

44. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014 May 16;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]

Abbreviations

API: application programming interface
CBOW: Continuous Bag of Words
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
NLP: natural language processing
PHQ-9: Patient Health Questionnaire-9

Edited by A Mavragani; submitted 10.10.23; peer-reviewed by J Magadia, M Remolana, AJ Li, C Entwistle, X Ren; comments to author 15.03.24; revised version received 29.04.24; accepted 02.07.24; published 04.09.24.

Please cite as:
Tumaliuan FB, Grepo L, Jalao ER
Development of Depression Data Sets and a Language Model for Depression Detection: Mixed Methods Study
JMIR Data 2024;5:e53365
URL: https://data.jmir.org/2024/1/e53365
doi:10.2196/53365
PMID:

Original Paper

A Database of Randomized Trials on the HIV Care Cascade (CASCADE Database): Descriptive Study

Lawrence Mbuagbaw[1,2,3,4], MPH, MD, PhD; Diya Jhuti[5], BHSc; Gohar Zakaryan[5], BHSc; Hussein El-Kechen[1], MSc; Nadia Rehman[1], BDentSc; Mark Youssef[6], MD; Michael Cristian Garcia[1], MSc; Vaibhav Arora[5], BHSc; Babalwa Zani[7], MSc; Alvin Leenus[8], MMASc; Michael Wu[9], BSc; Oluwatoni Makanjuola[5], BHSc

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

[2]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare, Hamilton, ON, Canada

[3]Centre for Development of Best Practices in Health, Yaoundé Central Hospital, Yaoundé, Cameroon

[4]Department of Global Health, Stellenbosch University, Cape Town, South Africa

[5]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[6]Department of Medicine, University of Toronto, Toronto, ON, Canada

[7]Nyasha Consulting, Cape Town, South Africa

[8]Faculty of Science, Western University, London, ON, Canada

[9]Michael DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Lawrence Mbuagbaw, MPH, MD, PhD
Department of Health Research Methods, Evidence and Impact
McMaster University
1280 Main Street West
Hamilton, ON, L8S3L8
Canada
Phone: 1 905 962 2477
Email: mbuagblc@mcmaster.ca

Abstract

Background: The Joint United Nations Programme on HIV/AIDS has set targets for 2025 regarding people living with HIV. For these targets to be met, 95% of people with HIV would need to know their HIV status, 95% of people with HIV would need to be receiving antiretroviral therapy, and 95% of people on antiretroviral therapy would need to be virally suppressed. Some countries are on track to meet these targets. However, within and across countries, several vulnerable populations may not meet these targets. This is in part because several approaches to improving the cascade of care after an HIV diagnosis are not tailored to and are not appropriate for vulnerable populations, such as men who have sex with men, sex workers, people who inject drugs, Black people, people in prisons, women, and youth. To inform research, policy, and practice, there is a need for curated data on HIV care cascade research.

Objective: The CASCADE database is a repository of randomized clinical HIV trials. It was designed to inform, support, and improve HIV care cascade research, systematic reviews, and evidence maps.

Methods: PubMed, Embase, CINAHL, PsycINFO, Web of Science, and the Cochrane Library were searched to obtain randomized trials that were designed to address at least one of the following care cascade outcomes: the initiation of therapy, adherence to medication, retention in care, and engagement in care. Data were screened and extracted in duplicate using DistillerSR software (Evidence Partners Incorporated) and were cataloged based on the following features: year, income level, setting, the delivery of the intervention, the population receiving the intervention, intervention type, and the level of pragmatism of the intervention.

Results: A total of 298 HIV clinical trials are included in the CASCADE database, of which 162 (54.4%) were conducted in high-income countries, and 116 (38.9%) targeted vulnerable populations. Adherence to antiretroviral therapy was the most investigated HIV care cascade outcome (216/298, 72.5%), followed by retention in care (34/298, 11.4%). CASCADE has a user-friendly interface with simple and advanced search features. The CASCADE database has inspired 2 methodological papers and 13,567 website visits from over 10 countries.

Conclusions: CASCADE is the first database dedicated to trials that focus on the HIV care cascade, and it can be used for systematic reviews, evidence maps, and methodological research. It is freely accessible, and the data can be downloaded in CSV format.

(JMIR Data 2022;3(1):e36874) doi:10.2196/36874

KEYWORDS

data set; database; cascade; HIV; antiretroviral therapy; adherence; retention; pragmatic; database; randomized controlled trial; randomized; clinical trial; review; evidence map; methodological research; open access

Introduction

From 2020 to 2021, about 37.7 million people were living with HIV worldwide, of which 28.2 million were accessing antiretroviral therapy (ART) [1]. Deaths from HIV-related causes have reduced by about 47% in the last decade [1]. People at higher risk of HIV infection, who are also referred to as *key populations* (eg, sex workers, men who have sex with men, people who inject drugs, and transgender people), account for 65% of all HIV infections [1].

Success in curbing morbidity and mortality due to HIV is measured based on the proportion of people achieving key steps in the HIV care cascade. The Joint United Nations Programme on HIV/AIDS (UNAIDS) has set targets for 2025 regarding people living with HIV. It is expected that by 2025, a total of 95% of people living with HIV will know their status, 95% of people who know their status will be on ART, and 95% of people on ART will achieve viral suppression [1]. For countries to meet these goals, targeted policies and interventions are required to enhance the uptake of testing for HIV; engagement in care; adherence to medication; and, ultimately, viral suppression. With close to 40 years of HIV research and innovation, enormous strides have been made toward reaching these targets in many countries [2-4]. However, there are concerns that even if countries meet these targets, stigma, discrimination, and structural barriers may prevent key subpopulations from meeting these targets. For instance, many countries have punitive laws and policies against people living with HIV, and gender inequalities and violence compromise access to care for many [5]. Further, successful research may never be implemented in the real world because the interventions are too complex, resource intensive, or insufficiently tailored [2].

As we strive to ensure that all of the UNAIDS targets are met globally (ie, across countries and within subpopulations), we must understand what needs to be done to ensure that all people living with HIV can meet these targets. Evidence syntheses can help with identifying knowledge gaps and shortcomings from the existing evidence base and highlighting areas where further research is needed [6]. To this end, and as part of an overview of systematic reviews to improve treatment initiation, adherence to ART, and retention in care for people living with HIV [5], we built the CASCADE database.

The CASCADE database [7] was born from efforts to create evidence maps on the HIV care cascade. This database is a "one-stop shop" for randomized trials of interventions that improve the initiation of treatment, adherence to medication, and retention and engagement in care. The database can be used to write literature reviews, conduct systematic reviews, conduct methodological research, create evidence maps, or simply learn about HIV care cascade research.

Methods

Database Implementation

The trials were identified by using a standard evidence synthesis methodology. The first set of trials were identified from an overview of systematic reviews on strategies for improving the HIV care cascade [8,9]. In brief, we searched PubMed, Embase, CINAHL, PsycINFO, Web of Science, and the Cochrane Library for systematic reviews published from 1995 to 2018, including at least 1 randomized trial of an intervention that was designed to improve the initiation of ART, adherence to ART, or retention in care. The second and subsequent sets of trials were identified through targeted searches for trials on the HIV care cascade that were published in the same databases from 2018 to August 2021. Only trials that were published as full texts and had participants who were people living with HIV that reported at least 1 care cascade outcome were eligible. All retrieved searches were deduplicated. Screening and data extraction were conducted in duplicate using DistillerSR software (Evidence Partners Incorporated).

The database is composed of 1 back-end table, which includes the following data: year, income level, the vulnerable population of focus, who delivered the intervention, the setting in which the intervention was delivered, the intervention type, and the level of pragmatism of the intervention (ie, how close it is to real-world care). The vulnerable population, who delivered the intervention, the setting of the intervention, and the intervention type are not mutually exclusive (ie, trials may belong to more than 1 category).

Database Functionalities

Simple Search

The simple search function identifies any key words in the title or abstract, including author and journal names.

Advanced Search

The advanced search function allows users to filter the results of a search based on the following features:

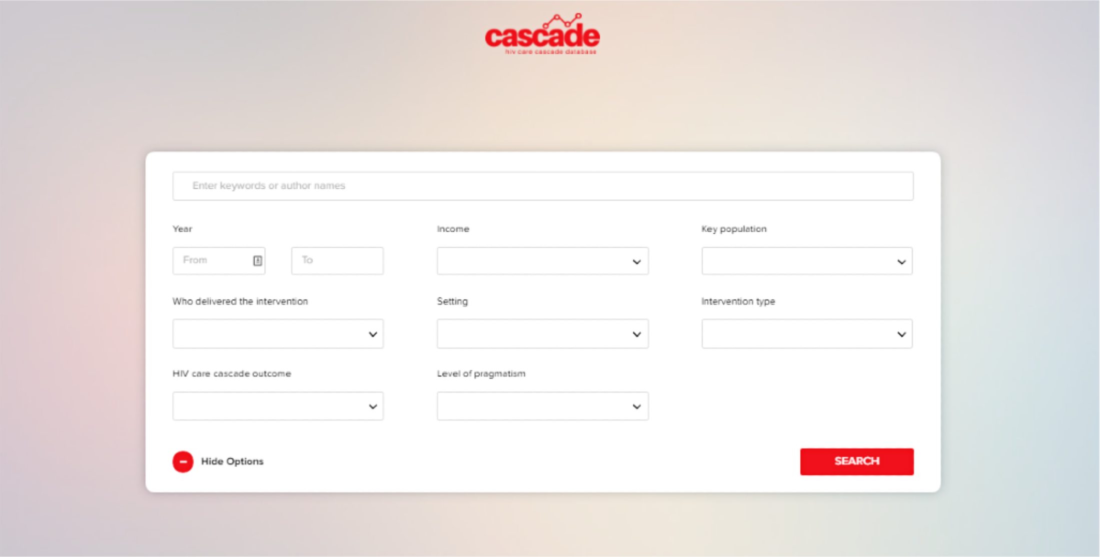
- Year: The platform allows the user to specify a start year and end year for the search and is based on the year of publication of the trial.
- Income level: This drop-down menu allows the user to select trials from countries with different income levels, as

defined by the World Bank, and includes high-income, upper-middle–income, low-middle–income, and low-income countries [10]. There is also a mixed category for studies conducted in multiple countries with different income levels.

- Key population: Trials are organized according to whether they primarily included people from any key population (ie, African, Caribbean, or Black people; commercial sex workers; men who have sex with men; prisoners; people who inject drugs; women; and youth).
- The person who delivered the intervention: The database distinguishes among trials of interventions delivered by clinicians, lay workers, or peers.
- The setting: Trials may be for interventions delivered in a clinic or community-based setting.
- Intervention type: A variety of intervention types are included, such as changes in health care delivery, counseling, education, electronic interventions, incentives, mobile health, outreach, peer navigation or support, and psychotherapy.
- Care cascade outcome: Users may select 1 of 4 care cascade outcomes—the initiation of treatment, adherence to ART, retention in care, and engagement in care.
- Level of pragmatism: The level of pragmatism—a measure of how well a trial matches usual care—was determined for each trial by using the RITES (Rating of Included Trials on the Efficacy-Effectiveness Spectrum) tool [11]. Trials may be situated along a continuum ranging from having a strong emphasis on efficacy to having a strong emphasis on effectiveness.

The layout of the advanced search page is shown in Figure 1.

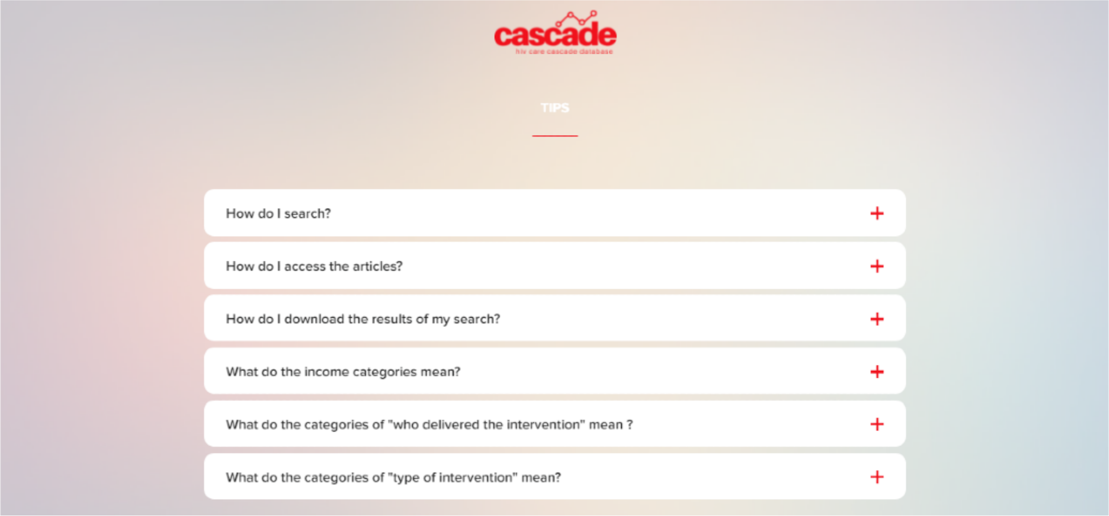
Figure 1. Image of the advanced search layout on the CASCADE website.



Tips

On the tips page, we display information on how to use the database and the types of trials that are included in each category. The layout of the tips page is shown in Figure 2.

Figure 2. Image of available tips on the CASCADE website.



How do I search? +

How do I access the articles? +

How do I download the results of my search? +

What do the income categories mean? +

What do the categories of "who delivered the intervention" mean ? +

What do the categories of "type of intervention" mean? +

User Interaction

All of the trials have hyperlinks that take users to the PubMed page or journal page of the identified articles. Search results can also be downloaded as a CSV file. Users are encouraged to suggest studies that could be added by emailing hccd@mcmaster.ca.

Results

As of December 30, 2021, CASCADE includes 298 HIV clinical trials that were published between 2015 and 2021. About half of these trials were conducted in high income countries (n=162, 54.4%), and 116 (38.9%) focused on a vulnerable population. About half of the interventions were delivered exclusively by clinicians (158/298, 53%) and were clinic-based interventions (173/298, 58.1%). The most frequent types of interventions were combinations (136/298, 45.6%), and adherence was the most common HIV care cascade outcome (216/298, 72.5%). About half of the trials (136/298, 45.6%) had a balanced emphasis on both efficacy and effectiveness. The key features of the trials in the database are outlined in Table 1.

In addition to the protocol for the overview of systematic reviews and the published overview [7,8], the CASCADE database has inspired 2 methodological papers. The first is an assessment of how HIV pilot studies are conducted [12], and the second evaluates the virological measures used in HIV clinical trials [13].

In the last year, as of June 24, 2022, the database has had 13,567 visits, with most visitors accessing the database from the United States, Canada, Singapore, Germany, the United Kingdom, China, Ireland, France, Lithuania, and the Netherlands.

Table 1. Characteristics of trials in the CASCADE database (N=298).

Variable	Trials, n (%)
Income level	
High	162 (54.4)
Upper middle	51 (17.1)
Low middle	39 (13.1)
Low	31 (10.4)
Mixed	15 (5)
Population	
General population of people with HIV	185 (61.1)
Vulnerable population	116 (38.9)
The people who delivered the intervention	
Clinicians exclusively	158 (53)
Laypersons exclusively	45 (15.1)
Peers exclusively	25 (8.4)
Combinations	70 (23.4)
Setting of intervention	
Clinic-based intervention	173 (58.1)
Community- and clinic-based intervention	44 (14.8)
Community-based intervention	46 (15.4)
Other	35 (11.7)
Type of intervention	
Changes in health care delivery	40 (13.4)
Counseling	36 (12.1)
Education	13 (4.4)
Incentives	9 (3)
Mobile health	34 (11.4)
Outreach	4 (1.3)
Peer navigation or support	9 (3)
Psychotherapy	17 (5.7)
Combinations of any types of interventions	136 (45.6)
Care cascade outcome	
Adherence	216 (72.5)
Retention	34 (11.4)
Engagement	8 (2.7)
Initiation	14 (4.6)
More than 1 cascade outcome	26 (8.7)
Level of pragmatism	
Strong emphasis on efficacy	1 (0.3)
Rather strong emphasis on efficacy	35 (11.7)
Balanced emphasis on both efficacy and effectiveness	136 (45.6)
Rather strong emphasis on effectiveness	80 (26.8)
Strong emphasis on effectiveness	46 (15.4)

Discussion

Principal Findings

CASCADE is a database of 298 HIV clinical trials that were published between 2015 and 2021. With 13,567 visits as of June 24, 2022, it is a widely used resource for clinical trials on the HIV care cascade. We plan to update the database on a semiannual basis, and we will continue to develop it based on user feedback. Updates will include formal searches for new trials and previously unpublished trials, and we will contact the authors of the included trials to suggest additional trials. Additional information that could be added for each trial include a detailed description of the intervention based on the TiDieR (Template for Intervention Description and Replication) checklist [14], risk of bias assessments, the indication of whether the study is a pilot study or a full trial, graphical displays, and other features. We will also validate the completion of the database against future systematic reviews. Further development is contingent on user interest and resources.

Comparison to Prior Work

To the best of our knowledge, there is no other database dedicated to randomized trials of the HIV care cascade. Other existing HIV databases either store data on the prevalence and incidence of HIV [15] or are dedicated to storing HIV genetic sequences [16,17]. Within the Cochrane Library, the Cochrane Controlled Register of Trials can be used to identify completed published and unpublished randomized trials, and ongoing trials. Specific filters can be applied to identify HIV trials. However, these trials are not categorized by the following: income level, the vulnerable population of focus, who delivered the intervention, the setting in which the intervention was delivered, the intervention type, and the level of pragmatism of the intervention [18].

Limitations

We faced challenges in categorizing trials that were not accessible as full texts, either because they were only published as conference abstracts or because the full texts were not available through our institutional libraries and the interlibrary loan system. We will continue to seek these full-text articles to include and curate these trials. Further, as this database is focused on the HIV care cascade, it does not include research on preventive interventions, such as pre-exposure prophylaxis and postexposure prophylaxis.

Conclusion

CASCADE is the first database dedicated to trials on the HIV care cascade. It includes information that can be used to learn about interventions, supplement systematic reviews, create evidence maps, and conduct methodological research. By looking at the bigger picture, this database can help researchers explore the similarities and differences among HIV clinical trials that may help explain why some interventions are more successful than others, highlight knowledge gaps, and inform future research.

Acknowledgments

Data extraction was completed by Anisa Hajizadeh, Annie Wang, Rita Morassut, and Jessica Bartoszko. The database and website were built by the Webcapitan team [19]. This work was supported by The Ontario HIV Treatment Network (grant EFP-1096-Junior Inv) and the Canadian Institutes of Health Research Canadian HIV Trials Network.

Authors' Contributions

LM and DJ wrote the first draft. All other authors (GZ, HE-K, NR, MY, MCG, VA, BZ, AL, MW, and OM) reviewed subsequent drafts and approved the final version.

Conflicts of Interest

None declared.

References

1. Fact sheet 2022. UNAIDS. URL: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf [accessed 2022-07-20]
2. 2025 AIDS targets. UNAIDS. URL: https://aidstargets2025.unaids.org/ [accessed 2022-11-26]
3. Lima V, St-Jean M, Rozada I, Shoveller JA, Nosyk B, Hogg RS, et al. Progress towards the United Nations 90-90-90 and 95-95-95 targets: the experience in British Columbia, Canada. J Int AIDS Soc 2017 Nov;20(3):e25011 [FREE Full text] [doi: 10.1002/jia2.25011] [Medline: 29130644]
4. Marinda E, Simbayi L, Zuma K, Zungu N, Moyo S, Kondlo L, et al. Towards achieving the 90-90-90 HIV targets: results from the south African 2017 national HIV survey. BMC Public Health 2020 Sep 09;20(1):1375 [FREE Full text] [doi: 10.1186/s12889-020-09457-z] [Medline: 32907565]
5. Lebelonyane R, Bachanas P, Block L, Ussery F, Alwano MG, Marukutira T, et al. To achieve 95-95-95 targets we must reach men and youth: High level of knowledge of HIV status, ART coverage, and viral suppression in the Botswana Combination Prevention Project through universal test and treat approach. PLoS One 2021 Aug 10;16(8):e0255227 [FREE Full text] [doi: 10.1371/journal.pone.0255227] [Medline: 34375343]

6. O'Leary BC, Woodcock P, Kaiser MJ, Pullin AS. Evidence maps and evidence gaps: evidence review mapping as a method for collating and appraising evidence reviews to inform research and policy. Environ Evid 2017 Jul 17;6(19):1-9 [FREE Full text] [doi: 10.1186/s13750-017-0096-9]
7. HIV care cascade database. CASCADE. URL: https://hivcarecascade.com/ [accessed 2022-01-02]
8. Mbuagbaw L, Hajizadeh A, Wang A, Mertz D, Lawson DO, Smieja M, et al. Overview of systematic reviews on strategies to improve treatment initiation, adherence to antiretroviral therapy and retention in care for people living with HIV: part 1. BMJ Open 2020 Sep 23;10(9):e034793 [FREE Full text] [doi: 10.1136/bmjopen-2019-034793] [Medline: 32967868]
9. Mbuagbaw L, Mertz D, Lawson DO, Smieja M, Benoit AC, Alvarez E, et al. Strategies to improve adherence to antiretroviral therapy and retention in care for people living with HIV in high-income countries: a protocol for an overview of systematic reviews. BMJ Open 2018 Sep 11;8(9):e022982 [FREE Full text] [doi: 10.1136/bmjopen-2018-022982] [Medline: 30206089]
10. World Bank open data. The World Bank. URL: https://data.worldbank.org/ [accessed 2019-06-21]
11. Wieland LS, Berman BM, Altman DG, Barth J, Bouter LM, D'Adamo CR, et al. Rating of included trials on the efficacy-effectiveness spectrum: development of a new tool for systematic reviews. J Clin Epidemiol 2017 Apr;84:95-104 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.01.010] [Medline: 28188898]
12. El-Khechen HA, Khan MIU, Leenus S, Olaiya O, Durrani Z, Masood Z, et al. Design, analysis, and reporting of pilot studies in HIV: a systematic review and methodological study. Pilot Feasibility Stud 2021 Nov 30;7(1):211 [FREE Full text] [doi: 10.1186/s40814-021-00934-9] [Medline: 34847957]
13. Youssef M, Zani B, Olaiya O, Soliman M, Mbuagbaw L. Virological measures and factors associated with outcomes, and missing outcome data in HIV clinical trials: a methodological study. BMJ Open 2021 Oct 25;11(10):e039462 [FREE Full text] [doi: 10.1136/bmjopen-2020-039462] [Medline: 34697107]
14. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
15. HIV/AIDS surveillance data base. United States Census Bureau. URL: https://www.census.gov/programs-surveys/international-programs/about/hiv.html [accessed 2022-07-20]
16. Foley BT, Korber BTM, Leitner TK, Apetrei C, Hahn B, Mizrachi I, et al. HIV sequence compendium 2018. Los Alamos National Laboratory. 2018. URL: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-18-25673 [accessed 2022-07-20]
17. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5(1):52-61 [FREE Full text] [Medline: 12875108]
18. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. Eval Health Prof 2002 Mar;25(1):38-64. [doi: 10.1177/016327870202500104] [Medline: 11868444]
19. Webcapitan. Webcapitan. URL: https://webcapitan.com/en/ [accessed 2022-07-20]

Abbreviations

ART: antiretroviral therapy
RITES: Rating of Included Trials on the Efficacy-Effectiveness Spectrum
TiDieR: Template for Intervention Description and Replication
UNAIDS: Joint United Nations Programme on HIV/AIDS

Edited by A Mavragani; submitted 29.01.22; peer-reviewed by FM Calisto, CMJ Wong; comments to author 19.05.22; revised version received 29.06.22; accepted 18.07.22; published 11.08.22.

Please cite as:
Mbuagbaw L, Jhuti D, Zakaryan G, El-Kechen H, Rehman N, Youssef M, Garcia MC, Arora V, Zani B, Leenus A, Wu M, Makanjuola O
A Database of Randomized Trials on the HIV Care Cascade (CASCADE Database): Descriptive Study
JMIR Data 2022;3(1):e36874
URL: https://data.jmir.org/2022/1/e36874
doi:10.2196/36874
PMID:

Original Paper

Conflicts of Interest Publication Disclosures: Descriptive Study

S Scott Graham[1], PhD; Jade Shiva[1], MA; Nandini Sharma[2], PhD; Joshua B Barbour[3], PhD; Zoltan P Majdik[4], PhD; Justin F Rousseau[5], MD

[1]The Department of Rhetoric & Writing, The University of Texas at Austin, Austin, TX, United States

[2]The Moody College of Communication, The University of Texas at Austin, Austin, TX, United States

[3]Department of Communication, The University of Illinois Urbana-Champaign, Champaign, IL, United States

[4]Department of Communication, North Dakota State University, Fargo, ND, United States

[5]Department of Neurology & O'Donnell Brain Institute, The University of Texas Southwestern Medical Center, Dallas, TX, United States

Corresponding Author:
S Scott Graham, PhD
The Department of Rhetoric & Writing
The University of Texas at Austin
Parlin Hall 29
Mail Code: B5500
Austin, TX, 78712
United States
Phone: 1 5124759507
Email: ssg@utexas.edu

Abstract

Background: Multiple lines of previous research have documented that author conflicts of interest (COI) can compromise the integrity of the biomedical research enterprise. However, continuing research that would investigate why, how, and in what circumstances COI is most risky is stymied by the difficulty in accessing disclosure statements, which are not widely represented in available databases.

Objective: In this study, we describe a new open access dataset of COI disclosures extracted from published biomedical journal papers.

Methods: To develop the dataset, we used ClinCalc’s Top 300 drugs lists for 2017 and 2018 to identify 319 of the most commonly used drugs. Search strategies for each product were developed using the National Library of Medicine’s and MeSH (Medical Subject Headings) browser and deployed using the eUtilities application programming interface in April 2021. We identified the 150 most relevant papers for each product and extracted COI disclosure statements from PubMed, PubMed Central, or retrieved papers as necessary.

Results: Conflicts of Interest Publication Disclosures (COIPonD) is a new dataset that captures author-reported COI disclosures for biomedical research papers published in a wide range of journals and subspecialties. COIPonD captures author-reported disclosure information (including lack of disclosure) for over 38,000 PubMed-indexed papers published between 1949 and 2022. The collected papers are indexed by discussed drug products with a focus on the 319 most commonly used drugs in the United States.

Conclusions: COIPonD should accelerate research efforts to understand the effects of COI on the biomedical research enterprise. In particular, this dataset should facilitate new studies of COI effects across disciplines and subspecialties.

(JMIR Data 2024;5:e57779) doi:10.2196/57779

KEYWORDS

conflicts of interest; biomedical publishing; research integrity; dataset; COI; ethical; ethics; publishing; drugs; pharmacies; pharmacology; pharmacotherapy; pharmaceuticals; medication; disclosure; information science; library science; open data

Introduction

Multiple lines of research have documented the effects that author conflicts of interest (COI) can have on the biomedical research enterprise [1-5]. Author COI have been shown to increase the likelihood of positive findings [1] and influence study design [6-9], and they may be associated with diminished product safety [5,10]. To mitigate the effects of COI, journals, universities, professional societies, and academic medical centers have adopted disclosure requirements and policies designed to support transparency around COI [2,11]. Although such transparency endeavors are essential for continued efforts to understand the nature, prevalence, and effects of author COI, research in this area remains limited by the lack of a comprehensive centralized repository of author COI disclosure data for a few reasons [12-15]. First, the vast majority of COI disclosures are not readily available for analysis. Although PubMed has included COI disclosure statements as an available data category since 2017 [16], most journals do not participate at all or do not participate in all papers [17-19]. Related research in scientometrics also indicates that these data are frequently incomplete, incorrect, or both in commercial publication databases such as the Web of Science [13-15]. In the COIPonD dataset, for example, journals deposited disclosure statements for an average of 6.5% (SD 0.21%) of papers, and 86% (3242/3769) of journals deposited no disclosure statements for any paper sought for retrieval. Similarly, the US Centers for Medicare and Medicaid Services Open Payments Database can support research into COI, but not all COIs are reflected in this resource. Second, the data that do exist may be difficult to match with individual journal authors, and third, they include only US health care providers, greatly limiting the applicability of the dataset to nonproviders, researchers, and those working in other jurisdictions. In sum, in its current state, much of the published research on COI is grounded in datasets limited to specific subspecialties, disciplines, or publication venues [20,21] due, in part, to the difficulty in accessing disclosure statements.

To address these issues, the Conflicts of Interest Publication Disclosures (COIPonD) dataset provides a comprehensive database of author-reported COI collected from published research pertaining to the most commonly used drug products in the United States. Specifically, the dataset provides author-reported COI data on over 38,000 individual papers published between 1949 and 2022. The entries include specific disclosures and, where relevant, information about the absence of disclosures. The data come from over 3500 English-language journals published across medical subspecialties. COIPonD offers a unique, extensive, and otherwise nearly impossible to obtain a collection of disclosed COIs.

The dataset offers numerous advantages for continued and rigorous research on COI. Researchers and policy makers can use these data to improve and expand efforts to assess variability in disclosure requirements and thresholds across multiple publication domains and timelines. In addition, this dataset was developed intentionally to provide a comprehensive view of PubMed-indexed literature. Therefore, COIPonD includes, by design, not only rigorous studies published in high-quality journals but also non–peer-reviewed content and opinion pieces, which have been shown to be a common vector for COI-induced biases [22], and publications in predatory journals, which increasingly influence the state of science through consultation and citation [23]. The dataset spans a 70-year publication period and includes over 38,000 papers, both broad and long-term in scale for this search context. Furthermore, an additional advantage of this dataset is that it extends beyond the currently common focus on individual disciplines or subspecialties by focusing instead on a diverse range of drug products. It can also support future research into more effective policy solutions for addressing the risks of COI. For example, existing research suggests that different types of COI involve different risks of bias, but has not yet specified the nature of those risks or how they might be mitigated in part because of the absence of a dataset such as this one [2]. Larger datasets, like this one, are necessary for appropriately powered subsample analyses. Understanding the effects of COI and how to effectively manage COI across research contexts is critical for ensuring the integrity of the biomedical research enterprise. This dataset can facilitate the development of knowledge to reach these aims.

Methods

Study Identification

As previously mentioned, research on COI generally focuses on datasets of clinical trials or systematic reviews grounded in specific subspecialties or disciplinary journals [2]. For example, evaluations of COI's effects on research have focused on psychiatry [24], oncology [25], and plastic surgery [26]. In contrast, the aim of our work was to capture information about COI related to the most frequently prescribed drug products across all subspecialties and publication types. In adopting this approach, we sought to mirror common practices for conducting a review of the relevant literature on a specific product. Specifically, for each identified product, we conducted a targeted search of PubMed for the most relevant papers. Our search focused on the most commonly used drug products as indexed by ClinCalc [27]. ClinCalc curates an annually updated list of the Top 300 most commonly used drugs. The list lags 2-3 years behind the current year. Drug use data are derived from ClinCalc's analysis of the Agency for Health Care Research and Quality annual Medical Expenditure Panel Survey, which surveys the US residents on medical drug use (prescription and over-the-counter). To develop the COIPonD dataset, we used ClinCalc's Top 300 drugs lists, each for 2017 and 2018 (the latest available at the time of query), to develop a list of target products. As expected, there was significant overlap in the most commonly used drugs between these 2 years, and so the final number of target products was 319.

For each of these 319 products, we used the National Library of Medicine's (NLM's) MeSH (Medical Subject Headings) browser to identify the preferred or supplementary concept, as most relevant. The MeSH-controlled vocabulary was selected for this project as it represents the primary ontology used by the NLM to support topic indexing and search retrieval. While alternative vocabularies, such as RxNorm, can be useful for improving interoperability between heterogeneous datasets, those alternative vocabularies are not engineered into the

MEDLINE information architecture. MeSH terms group synonymous records in categories called “concepts” [28]. For most products, the generic name is the preferred concept, and associated trade names map automatically to that generic name in the MeSH ontology. In some cases (eg, combination neomycin, polymyxin B, and dexamethasone), the generic names map to a different preferred concept, such as the trade name. Occasionally, the preferred concept does not match the actual product in use, so we identified appropriate supplementary concepts (eg, the preferred concept for levothyroxine is thyroxine). Finally, in a few cases, no preferred or supplementary concepts were available. In these cases, we developed queries that would search for corresponding trade and generic names.

Our overall approach was to target, as closely as possible, highly relevant papers for each drug product. We identified the 150 most relevant papers per product, according to the PubMed relevance algorithm [29]. The intentional use of this algorithm to order search results was to mirror what researchers would discover in a search. In April 2021, we deployed an iterative query development strategy through which we identified the relevant preferred concept. If a preferred concept was available, we searched narrowly among MeSH terms and assessed if the available results were sufficiently large to secure a sample of 150 papers. If no preferred concepts were available, we attempted to identify a relevant supplementary concept and conducted the search again. In cases where the number of returned results for MeSH terms or supplementary concepts was insufficient (fewer than 150 papers returned), we then expanded our search to the title and abstract fields. If the search was still insufficient, we expanded our search again to all PubMed fields. Finally, in cases where the generic and trade names were not mapped to each other in MeSH, we conducted independent searches for each and aggregated the results. A complete list of primary and secondary search terms is available in Multimedia Appendix 1. For all searches, we collected up to 200 articles from PubMed, sorted by relevance. In cases where there were multiple searches per product (eg, where MeSH did not map the trade name to the generic name), searches were aggregated, and the number of times each article appeared was used to resort to relevance. All searches were executed using Rentrez, an R package designed to access the National Center for Biotechnology Information’s (NCBI’s) eUtilities application programming interface (API) [30]. The final dataset includes the most relevant papers per product, up to 150 per product. Some papers in the COIPonD database were identified for inclusion multiple times as part of different product searches. In such cases, the paper’s association with multiple products is recorded in the dataset. A total of 9/319 drug products returned fewer than 50 results even after this iterative search process, and these products were excluded from further data collection.

Data Extraction

Once target papers were identified, we used the eUtilities API to extract available PubMed metadata for each paper. Our disclosure statement data collection protocol proceeded in three steps that are: (1) in cases where a PubMed-indexed disclosure statement was available, it is included in the final dataset; (2) if no PubMed disclosure statement was available in the metadata, we developed an automated tool that attempted to locate and extract the relevant data from PubMed Central; and (3) if target disclosure statements were not available on either PubMed or PubMed Central, the paper was referred for manual collection.

While PubMed indexes all available disclosure statements under the XML tag <CoiStatement>, PubMed Central uses a variety of different possible tags. Therefore, for each paper indexed in PubMed Central, we searched all available XML tags for the following terms: conflict of interest, funding, conflicts of interest, disclosure statement, disclosure, author disclosure statement, and declaration of interest. All PubMed and PubMed Central data extraction was performed also using the Rentrez R package accessing the NCBI eUtilities API. In some cases, the multi-tag query approach led to multiple results being returned for a single paper. Any paper with multiple results was manually reviewed. In cases where the results were duplicate, they were reduced to a single entry. In cases where the results were unique, the results were combined into a single statement. This aggregation was most often required when a given paper’s XML schema separated each author’s disclosure into separate statements rather than combining them.

For each paper that did not have a disclosure statement available on PubMed or PubMed Central, we sought to locate the full text and extract the relevant data. Journal conventions for displaying disclosure statements vary widely, so we adopted the following heading prioritization schema to help ensure more uniform results. Specifically, we extracted all text under any heading of “Conflict of Interests,” “Conflicts of Interest,” “Competing Interests,” or “Duality of Interests.” If no such heading was available, the data extraction team then looked to assess the appropriateness of available information under “Disclosures,” “Funding,” or “Acknowledgements” headings. We extracted all data that specifically mentioned the terms “conflict of interests,” “conflicts of interest,” “competing interests,” or “duality of interests,” and any text that identified author-level financial relationships. We did not collect data from secondary headings if those data only identified study funding sources (as opposed to author-level relationships) or provided more general acknowledgments and statements of appreciation. In some cases, disclosure statements were hyperlinks to authors’ completed ICMJE (International Committee of Medical Journal Editors) forms. In these cases, the link was collected, but we did not seek to evaluate the data in the PDFs. Finally, if no disclosure information was available, the absence of available disclosures was recorded in the dataset.

Quality Assurance

Once data collection was complete, we conducted a quality assurance protocol to evaluate data accuracy (Table 1). This protocol involved first drawing a random sample of papers (N=381). The sample size was determined using the Cochran equation to determine representativeness at 95% confidence level and a 5% margin of error [31]. Where possible, the full text of each paper in the sample was located, and relevant data were re-extracted per the same data collection protocol described above. Re-extracted data were compared against previously collected data. The results of this protocol found that data collected for 349/381 (91.6%) of the papers were correct. Among the incorrect data 16/381 were classified as incorrect due to innocuous incomplete capture that would not adversely affect subsequent analyses (4.2%). Typically, this involved disclosure statements where additional assurances were provided and data entry only captured the reported COI. Another subset of 15/381 of the errors involved cases where data would have been partially or entirely missed (3.9%), and one was a case where a null result was incorrectly classified (no COI statement instead of non-English).

Table 1. Data collection error rates by error type.

Category	n (%)
Correct capture	349 (91.6)
Innocuous incomplete capture	16 (4.2)
Relevant data not captured	15 (3.9)
Incorrectly classified null result	1 (0.3)

Ethical Considerations

This project does not involve human subjects or animal research, and no ethics review was required.

Results

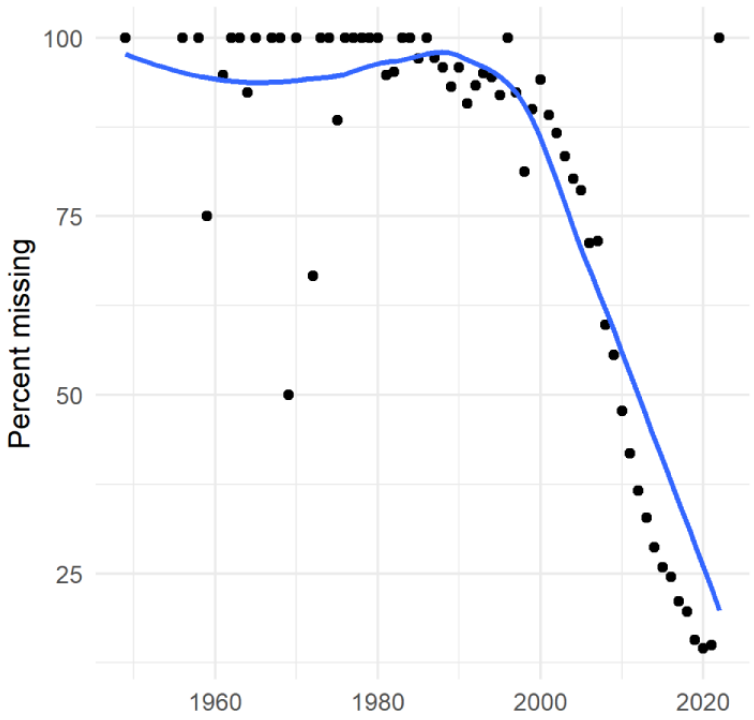
The total number of papers included in the dataset is 38,705. The final dataset includes papers indexed to 319 unique drug products, and the number of papers located per drug product ranges from 32 to 150 with an average of 139 (SD 18.99). The paper sample size is greater than 100 for 296 of the target drug products. Recall that 9/319 drug products returned fewer than 50 results and were excluded from the study and further data collection. The collected papers were published in 3769 unique journals. Each journal is represented between 1 and 488 times, with an average of 10.27 (SD 23.26) papers per journal. The most commonly represented journals are *PLoS One* (n=488), *Scientific Reports* (n=454), *BMJ Case Reports* (n=351), *Medicine* (n=319), and the *International Journal of Pharmaceutics* (n=309). Paper publication years range from 1949 to 2022, with an average publication date of 2016. Notably, the dataset includes 10,134 papers with no locatable disclosure statement. Missing rates vary widely by publication year (14%-100% with an average of 78%). As would be expected, we see a sharp decrease in missing rates as journal COI policies started to proliferate in the 1990s. Figure 1 lists the details (due to the timing of the query, there is only 1 article from 2022 in the dataset).

Notably, the presence of a disclosure statement does not definitively indicate that a COI was disclosed. In many cases, authors disclosed no specific relationships, typically asserting that there were no disclosures or indicating affirmatively that there was nothing to disclose. Accordingly, COIPonD contains approximately 15,500 entries that indicate no conflict existed, including many variations of “none” or “nothing to disclose.” These values are approximate because they are based on an analysis of disclosure statements that repeat exactly 2 or more times in the dataset. It is probable that a number of unique disclosure statements documenting either the absence of disclosure or the absence of COI are available in the data. These likely include author-specific phraseology such as “RV has nothing to disclose.” In terms of the repeated negative disclosures, we document over 600 variations, including “none” (n=1024), “the authors declare no competing interests” (n=430), and “no competing financial interests exist” (n=112). In addition, approximately 200 records contain URLs, and a substantial proportion of these are links to PDF disclosure forms maintained by the publishing journal. This leaves approximately 12,500 entities reporting the presence of COI of some variety.

Most statements were not available on PubMed or PubMed Central. Of the 38,705 papers for which we sought to retrieve disclosure statements, only 3184 were available on PubMed (8%). Of those not available on PubMed, approximately 10% (3871/38,705) were retrieved from PubMed Central. The vast majority of disclosure statements had to be retrieved manually.

The current release of the COIPonD dataset can be found on its Texas Data Repository page [32]. The data are provided in two data tables that include (1) paper metadata, (2) disclosures, and (3) drug look-up (Textbox 1).

Figure 1. Missing disclosure statement rates (%) by year. Fit with a Loess regression model for ease of visualization.



Textbox 1. The article, disclosure, and drug look-up data structures.

Paper metadata

- PMID: the unique PubMed identifier for each paper.
- Title: the title of the paper.
- Journal: the Internal Organization for Standardization journal abbreviation.
- Pub date: the paper's publication date.

Disclosure data structure

- PMID: the unique PubMed identifier for each paper.
- COI: the author-reported disclosure statement or indication of its absence, in its original form.

Drug look-up data structure

- PMID: the unique PubMed identifier for each paper.
- Drug ID: the unique identifier for each drug product.
- Drug name: the canonical name of each product (uses the name provided on ClinCalc, as derived from the Medical Expenditure Panel Survey).

Discussion

COIPonD indexes COI disclosure statements for the 38,705 most relevant papers related to 319 unique drug products that were the most commonly used in the United States in 2017 and 2018. The data come from journals across medical subspecialties. The dataset is a unique, extensive collection of disclosed COI that has otherwise been very difficult, if not impossible, to obtain previously. When combined with the associated article metadata and relevant product table, these data can serve as a foundation for new research on COI and the subsequent development of evidence-based COI management policies. COIPonD affords a range of benefits to investigators and can support a variety of study designs. Primarily, it offers researchers and policy makers an assortment of data for assessing potential bias and differential effects based on the type of COI. Results from these studies might be used to develop or implement evidence-based COI policy reform in biomedical publishing. Because the dataset focuses on author-level disclosures across a diverse range of drug products, it is possible to evaluate drug safety and risk in relation to disclosed COI and by type of COI [5]. COIPonD could be used to establish an evidence-based understanding of variability in disclosure requirements and authors' understandings of those requirements.

In addition, the data can assist government and institutional policy makers to make better-informed decisions about implementing new disclosure policies based on evidence. Much of the current research on COI policy is limited to comparisons of (1) the presence or absence of COI and (2) associated study features, such as positive results or methodological rigor. In contrast, the organization of this dataset provides opportunities for more granular and rigorous examinations of outcomes associated with different COI types. Subsequent analysis with COIPonD data could help address persistent and misleading calls for yet more evidence that COI has adverse effects on the biomedical research enterprise. As such, it can facilitate more robust COI research and reform. Associated metadata about disclosure statement provenance may also be helpful in evaluating the transparency of certain publications, publishers, and government transparency initiatives. These features of the dataset are relevant to researchers interested in developing an effective public management system for COI disclosure. For example, researchers might use it to conduct more comprehensive audits of studies comparing disclosed COI to data in the Open Payments Database. In addition, research using these data may contribute to the development of more robust transparency and reporting frameworks and data collection and aggregation systems. For example, while research has suggested the US NLM and the ORCID (Open Researcher and Contributor ID) researcher registry are particularly suited to a public registry, this dataset provides a starting point for those interested in building the necessary infrastructure and systems to support new institutional practices by making disclosures more accessible and streamlined in PubMed and PubMed Central [12,17].

These data were collected to pursue specific research questions and thus are not suitable for use in all study designs. For example, research on industry influence on the biomedical enterprise may wish to consider the effects of study sponsorship in addition to the disclosed COI. This dataset does not include information specifically related to study funding or sponsorship, but may contribute in combination with other data sources. The data are also limited to the articles related to the most commonly used drug products in the United States at a specific time (2017-2018). Data on new, less commonly used products and medical devices are not available. In particular, less frequently prescribed but more expensive biological products are not included in this dataset, even though their economic effects may be significant and are an important direction for future work. Although these are important components of broader research efforts related to industry influence on the biomedical research enterprise, they fall beyond the scope of the research effort that produced this dataset. The papers included in this dataset are limited to those available through a PubMed search at the time the query was initiated (April 2021). As millions of new papers are published each year and considering ongoing efforts of drug repurposing, we expect the COI picture to change with time, even for the same drug products.

Finally, the results of this research point to the need for more robust data management approaches for collecting information about COI disclosures. Although PubMed has included COI disclosure statements in its system since 2016, participation is optional and therefore low. PubMed Central offers some additional access to disclosure statements but is reliant on individual publisher data structures, leading to inconsistent field names for available disclosures. The fee-based Web of Science offers a further resource but has been noted for similar information gaps and low data quality with respect to COI [13-15]. These serious gaps in data availability led to substantial human labor hour requirements in order to complete the COIPonD dataset. It would be a significant benefit to the medical, bioethics, and scientometric research communities if repositories could explore ways to incentivize or even compel more complete participation from major publishers. For example, participation in citation metrics might be made contingent on the full reporting of COI data. Researchers and policy makers alike would benefit from ready access to COI disclosures as well as the ability to filter results based on the presence or absence of these relationships. However, even with greater participation in COI reporting systems, additional computational or labor investments would be required for researchers to make the most productive use of available data. We would, therefore, also endorse previous recommendations [18,33] to require COI disclosure reporting in structured formats that cleanly identify and link funders, recipients, and mechanisms of disbursement.

Acknowledgments

Dataset development was supported by the National Institute of General Medical Sciences of the National Institutes of Health (award R01GM141476).

Code Availability

The PubMed and PubMed Central query and data extraction code base is available at GitHub [34].

Authors' Contributions

SSG, ZPM, JBB, and JFR contributed to conceptualization and project administration. SSG, NS, and JSE performed data curation and validation and contributed to writing-original draft. SSG and NS conducted formal analysis and assisted with the software. SSG, ZPM, JBB, and JFR assisted with funding acquisition. SSG contributed to methodology, supervision, and visualization. SSG, ZPM, JBB, JFR, NS, and JSE assisted with writing-review and editing.

Conflicts of Interest

SSG reports grant funding from the National Institute of General Medical Sciences (NIGMS) and the Texas Health and Human Services Commission. ZPM reports grant funding from NIGMS. JBB reports grant funding from NIGMS, NSF, and Blue Cross Blue Shield/Health Care Service Corporation. JFR reports grant funding from NIGMS, NIMH, NIAID, NLM, Health Care Cost Institute, Austin Public Health, Texas Child Mental Health Care Consortium, Texas Alzheimer's Research and Care Consortium, and the Michael and Susan Dell Foundation. JFR also reports receiving an award from the NIH Division of Loan Repayment.

Multimedia Appendix 1
Primary and secondary search queries for included products.
[DOCX File , 48 KB - data_v5i1e57779_app1.docx]

References

1. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev 2017;2(2):MR000033 [FREE Full text] [doi: 10.1002/14651858.MR000033.pub3] [Medline: 28207928]
2. Graham SS, Karnes MS, Jensen JT, Sharma N, Barbour JB, Majdik ZP, et al. Evidence for stratified conflicts of interest policies in research contexts: a methodological review. BMJ Open 2022;12(9):e063501 [FREE Full text] [doi: 10.1136/bmjopen-2022-063501] [Medline: 36123074]
3. Bero LA, Rennie D. Influences on the quality of published drug studies. Int J Technol Assess Health Care 1996;12(2):209-237. [doi: 10.1017/s0266462300009582] [Medline: 8707496]
4. Waqas A, Baig AA, Khalid MA, Aedma KK, Naveed S. Conflicts of interest and outcomes of clinical trials of antidepressants: an 18-year retrospective study. J Psychiatr Res 2019;116:83-87. [doi: 10.1016/j.jpsychires.2019.05.029] [Medline: 31212249]
5. Graham SS, Majdik ZP, Barbour JB, Rousseau JF. Associations between aggregate NLP-extracted conflicts of interest and adverse events by drug product. Stud Health Technol Inform 2022;290:405-409 [FREE Full text] [doi: 10.3233/SHTI220106] [Medline: 35673045]
6. Fraguas D, Díaz-Caneja CM, Pina-Camacho L, Umbricht D, Arango C. Predictors of placebo response in pharmacological clinical trials of negative symptoms in schizophrenia: a meta-regression analysis. Schizophr Bull 2019;45(1):57-68 [FREE Full text] [doi: 10.1093/schbul/sbx192] [Medline: 29370436]
7. Gao Y, Ge L, Ma X, Shen X, Liu M, Tian J. Improvement needed in the network geometry and inconsistency of Cochrane network meta-analyses: a cross-sectional survey. J Clin Epidemiol 2019;113:214-227. [doi: 10.1016/j.jclinepi.2019.05.022] [Medline: 31150834]
8. Kapelios CJ, Naci H, Vardas PE, Mossialos E. Study design, result posting, and publication of late-stage cardiovascular trials. Eur Heart J Qual Care Clin Outcomes 2022;8(3):277-288. [doi: 10.1093/ehjqcco/qcaa080] [Medline: 33098422]
9. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003;326(7400):1167-1170 [FREE Full text] [doi: 10.1136/bmj.326.7400.1167] [Medline: 12775614]
10. Gyawali B, Tessema FA, Jung EH, Kesselheim AS. Assessing the justification, funding, success, and survival outcomes of randomized noninferiority trials of cancer drugs: a systematic review and pooled analysis. JAMA Netw Open 2019;2(8):e199570 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9570] [Medline: 31469391]
11. Mialon M, Vandevijvere S, Carriedo-Lutzenkirchen A, Bero L, Gomes F, Petticrew M, et al. Mechanisms for addressing and managing the influence of corporations on public health policy, research and practice: a scoping review. BMJ Open 2020;10(7):e034082 [FREE Full text] [doi: 10.1136/bmjopen-2019-034082] [Medline: 32690498]
12. Dunn AG. Set up a public registry of competing interests. Nature 2016;533(7601):9. [doi: 10.1038/533009a] [Medline: 27146998]
13. Lewison G, Sullivan R. Conflicts of interest statements on biomedical papers. Scientometrics 2014;102(3):2151-2159. [doi: 10.1007/s11192-014-1507-0]
14. Yegros YA, Lamers W, Díaz-Faes AA. Research integrity at stake: conflicts of interest and industry ties in scientific publications. 2022. URL: https://digital.csic.es/handle/10261/304326 [accessed 2024-08-23]
15. Álvarez-Bornstein B, Bordons M. Industry involvement in biomedical research: authorship, research funding and conflicts of interest. In: 17th International Conference ON Scientometrics & Infometrics, Proceedings Volume II. 2019 Presented at: 17th International Conference ON Scientometrics & Infometrics; 2-5 September 2019; Rome, Italy p. 1746-1751 URL: https://digital.csic.es/bitstream/10261/240184/1/ISSI-2019-Industry_involvement_in_biomedical_research_authorship%2C_research_funding_and_conflicts_of_interest.pdf
16. Collins M. NLM Technical Bulletin U.S. National Library of Medicine. 2017. URL: https://www.nlm.nih.gov/pubs/techbull/tb.html [accessed 2022-08-20]
17. Graham SS, Majdik ZP, Clark D, Kessler MM, Hooker TB. Relationships among commercial practices and author conflicts of interest in biomedical publishing. PLoS One 2020;15(7):e0236166 [FREE Full text] [doi: 10.1371/journal.pone.0236166] [Medline: 32706798]

18. Grundy Q, Imahori D, Mahajan S, Garner G, Timothy R, Sud A, et al. Cannabis companies and the sponsorship of scientific research: a cross-sectional Canadian case study. PLoS One 2023;18(1):e0280110 [FREE Full text] [doi: 10.1371/journal.pone.0280110] [Medline: 36626363]
19. Falciola L, Barbieri M. Disclosure of patenting activities within scientific publications as potential conflicts-of-interest: Evidences from biomedical literature. World Patent Information 2024;76:102251. [doi: 10.1016/j.wpi.2023.102251]
20. Tisherman RT, Wawrose RA, Chen J, Donaldson WF, Lee JY, Shaw JD. Undisclosed conflict of interest is prevalent in Spine literature. Spine (Phila Pa 1976) 2020;45(21):1524-1529. [doi: 10.1097/BRS.0000000000003589] [Medline: 32628433]
21. Probst P, Hüttner FJ, Klaiber U, Diener MK, Büchler MW, Knebel P. Thirty years of disclosure of conflict of interest in surgery journals. Surgery 2015;157(4):627-633. [doi: 10.1016/j.surg.2014.11.012] [Medline: 25704418]
22. Nejstgaard CH, Bero L, Hróbjartsson A, Jørgensen AW, Jørgensen KJ, Le M, et al. Association between conflicts of interest and favourable recommendations in clinical guidelines, advisory committee reports, opinion pieces, and narrative reviews: systematic review. BMJ 2020;371:m4234 [FREE Full text] [doi: 10.1136/bmj.m4234] [Medline: 33298430]
23. Oermann MH, Nicoll LH, Ashton KS, Edie AH, Amarasekara S, Chinn PL, et al. Analysis of citation patterns and impact of predatory sources in the nursing literature. J Nurs Scholarsh 2020;52(3):311-319. [doi: 10.1111/jnu.12557] [Medline: 32346979]
24. Ahmer S, Arya P, Anderson D, Faruqui R. Conflict of interest in psychiatry. Psychiatr Bull 2018;29(8):302-304. [doi: 10.1192/pb.29.8.302]
25. Hampson LA, Joffe S, Fowler R, Verter J, Emanuel EJ. Frequency, type, and monetary value of financial conflicts of interest in cancer clinical research. J Clin Oncol 2007;25(24):3609-3614. [doi: 10.1200/JCO.2006.09.3633] [Medline: 17704409]
26. Lopez J, Musavi L, Quan A, Calotta N, Juan I, Park A, et al. Trends, frequency, and nature of surgeon-reported conflicts of interest in plastic surgery. Plast Reconstr Surg 2017;140(4):852-861. [doi: 10.1097/PRS.0000000000003683] [Medline: 28953741]
27. The top 300 of 2020. URL: https://clincalc.com/DrugStats/Top300Drugs.aspx [accessed 2023-09-21]
28. Concept Structure in MeSH. U.S. National Library of Medicine. URL: https://www.nlm.nih.gov/mesh/concept_structure.html [accessed 2023-10-04]
29. U.S. National Library of Medicine. Updated algorithm for the PubMed best match sort order. URL: https://www.nlm.nih.gov/pubs/techbull/tb.html [accessed 2023-10-04]
30. Winter DJ. Rentrez: an R package for the NCBI eUtils API.: The R Journal; 2017. URL: https://journal.r-project.org/archive/2017/RJ-2017-058/RJ-2017-058.pdf [accessed 2024-09-27]
31. Cochran WG. Sampling Techniques. New York, NY: John Wiley & Sons; 1977.
32. Graham SS, Sharma N, Barbour JB, Edward JS, Majdik ZP, Rousseau JF. Texas Data Repository. Conflicts of interest publication disclosures (COIPonD) data set. 2023. URL: https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/GBSTTH [accessed 2024-09-27]
33. Dunn AG, Coiera E, Mandl KD, Bourgeois FT. Conflict of interest disclosure in biomedical research: a review of current practices, biases, and the role of public registries in improving transparency. Res Integr Peer Rev 2016(1):1 [FREE Full text] [doi: 10.1186/s41073-016-0006-7] [Medline: 27158530]
34. GitHub. COIPonD. URL: https://github.com/sscottgraham/COIPonD [accessed 2024-09-26]

Abbreviations

API: application programming interface
COI: conflicts of interest
COIPonD: Conflicts of Interest Publication Disclosures
ICMJE: International Committee of Medical Journal Editors
MeSH: Medical Subject Headings
NCBI: National Center for Biotechnology Information
NLM: National Library of Medicine
ORCID: Open Researcher and Contributor ID

Edited by A Coristine; submitted 26.02.24; peer-reviewed by T Burke, Y Jiang; comments to author 22.08.24; revised version received 28.08.24; accepted 30.08.24; published 31.10.24.

Please cite as:
Graham SS, Shiva J, Sharma N, Barbour JB, Majdik ZP, Rousseau JF
Conflicts of Interest Publication Disclosures: Descriptive Study
JMIR Data 2024;5:e57779
URL: https://data.jmir.org/2024/1/e57779
doi:10.2196/57779
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org