Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6387–6396
Marseille, 20-25 June 2022

CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts

Muskan Garg[1,7], **Chandni Saxena**[2], **Veena Krishnan**[3], **Ruchi Joshi**[4],
Sriparna Saha[5], **Vijay Mago**[6], **Bonnie J Dorr**[7]
[1]Thapar Institute of Engineering & Technology, India,
[2]The Chinese University of Hong Kong, Hong Kong SAR,
[3]University of Petroleum And Energy Studies, India [4]Amity University Rajasthan, India,
[5]Indian Institute of Technology, Patna, [6]Lakehead University, Canada, [7]University of Florida, USA.
{muskangarg, bonniejdorr}@ufl.edu, chandnisaxena@cuhk.edu.hk, sriparna@iitp.ac.in,
vkrishnan@ddn.upes.in, rjoshi@jpr.amity.edu, vmago@lakeheadu.ca

Abstract

Research community has witnessed substantial growth in the detection of mental health issues and their associated reasons from analysis of social media. We introduce a new dataset for Causal Analysis of Mental health issues in Social media posts (CAMS). Our contributions for causal analysis are two-fold: *causal interpretation* and *causal categorization*. We introduce an annotation schema for this task of causal analysis. We demonstrate the efficacy of our schema on two different datasets: (i) crawling and annotating 3155 Reddit posts and (ii) re-annotating the *publicly available SDCNL dataset* of 1896 instances for interpretable causal analysis. We further combine these into the CAMS dataset and make this resource publicly available along with associated source code: `https://github.com/drmuskangarg/CAMS`. We present experimental results of models learned from CAMS dataset and demonstrate that a classic Logistic Regression model outperforms the next best (CNN-LSTM) model by 4.9% accuracy.

Keywords: clinical depression, clinical psychology, intent classification, suicidal tendency

1. Introduction

With substantial growth in digitization of psychological phenomena, automated Natural Language Processing (NLP) has been applied by academic researchers and mental health practitioners to detect, classify or predict mental illness on social media. However, there is a critical need for identifying underlying *causes* of mental illness in the face of dire outcomes. For example, a person commits suicide every 11.1 minutes in the US[1] and 23% of deaths in the world are associated with mental disorders according to the World Health Organization. The pandemic lockdown has heightened the mental health crisis in UK (Pierce et al., 2020) and US (McGinty et al., 2020).

In this context, people with mental disorders who decide to visit mental health practitioners for social well-being may face difficulty due to *social stigma* or *unavailability of mental health practitioners*, leaving those most in need to be neglected by the community. As a result, sufferers of mental health conditions are unable to take necessary steps for their treatment. Unfortunately, 80% of those with mental health conditions do not undergo clinical treatment and about 60% of those who take their own lives previously denied having any suicidal thoughts (Sawhney et al., 2021). Accordingly, social media platforms (e.g., *Reddit, Twitter*) are important resources for investigating the mental health of users based on their writings.

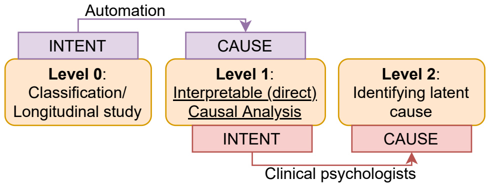

Automation
INTENT
CAUSE
Level 0: Classification/ Longitudinal study
Level 1: Interpretable (direct) Causal Analysis
Level 2: Identifying latent cause
INTENT
CAUSE
Clinical psychologists

Figure 1: The intent-cause analysis of mental health on social media

1.1. Motivation

The research community has witnessed tremendous growth in the study of mental health classification on social media since 2013 (Garg, 2021). However, there is minimal automation for identifying potential causes that underlie mental illness. Online users suffering from depression may express their thoughts and grievances on social media unintentionally, for instance, the post (P).

> P: *I cannot deal with this breakup anymore and want to finish my life*

The reason behind depression in P is clearly interpreted from the word *breakup*, which serves as an indicator of a cause related to the notion of *relationship*. Through the application of automatic *causal analysis*, underlying reasons of this type may be extracted and potentially leveraged to address mental health problems.

Social, financial and emotional disturbances have a

[1] https://suicidology.org/wp-content/uploads/2021/01/2019 datapgsv2b.pdf

Dataset	Task	Avail.
CLPsych (Coppersmith et al., 2015)	Depression detection for suicide risk	S
MDDL (Shen et al., 2017)	Depression candidate detection (D1, D2, D3)	A
RSDD (Yates et al., 2017)	Depression detection from Reddit data	ASA
SMHD (Cohan et al., 2018)	Multi-task mental illness from Reddit data	ASA
eRISK (Losada et al., 2018)	Early risk detection: CLEF	A
Pirina18 (Pirina and Çöltekin, 2018)	Depression detection from Reddit data	A
Ji18 (Ji et al., 2018)	Suicide risk detection from Reddit data	AR
Aladag18 (Aladağ et al., 2018)	Suicide risk detection	AR
Sina Weibo (Cao et al., 2019)	Identifying candidates with suicide risk	AR
SRAR (Gaur et al., 2019)	Suicide risk from Reddit posts	ASA
Dreaddit (Turcan and McKeown, 2019)	Stress detection from Reddit posts	A
UMD-RD (Shing et al., 2020)	Suicide risk detection from Reddit data	ASA
SDCNL (Haque et al., 2021)	Suicide v/s depression from Reddit	A
CAMS (Ours)	Interpretable Causal analysis from Reddit	A

Table 1: Different mental health datasets and their availability. A: Available, ASA: Available via Signed Agreement, AR: Available on Request for research work

huge impact on the mental health of online users. Here, we consider three levels of mental disorder analysis from *automation* to *latent cause* as shown in Figure 1. We identify the *intent (level 0 task)* of a user by mental illness prediction and classification of social media posts. We further automate the process of identifying and categorizing the *direct cause (Level 1)* that a user may mention in the post. In the future, causal analysis may discern crucial protective factors for mental health and address some important societal needs. Domain experts refer to *level 1* as a *direct cause* mentioned by a user, often accompanied by a *latent cause (Level 2)* when they are posted on social media. In this work, we focus on automation by introducing a dataset for *Level 1: interpretable causal analysis*.

1.2. Challenges and Contributions

Mental health illness detection and analysis on social media presents many linguistic, technical and psychological challenges. Among many under-explored dimensions, some substantial research gaps are addressed as follows:

1. Dataset availability may be limited due to the sensitive nature of personal information.

2. There are many existing Level 0 studies for mental health detection but no substantial study for Level 1, e.g., in-depth causal analyses of disorders.

To address these challenges, we introduce the task of *causal analysis*. We first introduce an *annotation scheme* for causal analysis. The dataset annotations are carried out in two ways: (i) crawling and annotation of the Reddit dataset (ii) re-annotation of the existing SDCNL dataset (Haque et al., 2021) for the proposed task of *Causal Analysis of Mental health on Social media* (CAMS). There are no existing studies for this task as observed from Table 1. To the best of our knowledge, our work is the first to address *causal analysis* and to provide a publicly available dataset for this purpose. Our major contributions are:

1. Definition of *Interpretable Causal Analysis* and construction of an annotation schema for this new task.

2. Annotated web-crawled Reddit dataset of 3362 instances using our annotation schema.

3. Re-annotation of the existing SDCNL dataset as a robustness test for our annotation schema.

4. Combination of the datasets above and introduction of our new, publicly available CAMS dataset.

5. Demonstration of the performance of machine learning and deep learning models using CAMS.

Below we discuss relevant background (Section 2) and introduce the annotation scheme for causal analysis (Section 3). Section 4 presents our new CAMS resource, annotation, and validation. Annotations are verified by experts (clinical psychologist and rehabilitation counselor) and validated using statistical testing of Fleiss' Kappa agreement (Falotico and Quatto, 2015). We further use existing multi-class classifiers for interpretable causal analysis in Section 5. Section 6 provides concluding remarks and future research directions.

2. Background

Reddit has become one of the most widely used social media platforms. Haque et al. (2021) use two subreddits $r/depression$ and $r/suicidewatch$ to scrape the SDCNL data and to validate a label correction methodology through manual annotation of this dataset for *depression* versus *suicide*. They also address ethical issues impacting *dataset availability* and make their dataset publicly available. In this section, we discuss the *evolution of mental health studies* and the *historical perspective of causal analysis*.

2.1. Evolution of Mental Health Studies

New NLP questions have emerged from investigations into predicting depression (De Choudhury et al., 2013) and suicidal tendencies (Masuda et al., 2013). Researchers consider users' profiles (Conway and O'Connor, 2016) to introduce the CLPsych shared task dataset (Coppersmith et al., 2015) for solving the problem of Mental Illness Detection and Analysis on Social media (MIDAS). MIDAS has further benefited from exploiting social network features (Lin et al., 2017), attention mechanisms (Nam et al., 2017), handling imbalanced dataset (Cong et al., 2018), and explainability (Cao et al., 2019).

Additional research directions have emerged from the use of knowledge graphs (Cao et al., 2020), feature optimization techniques (Shah et al., 2020), longitudinal studies (Sawhney et al., 2021), and handling noisy labels (Haque et al., 2021). The 3-step theory (3ST) (Klonsky et al., 2021) of suicide supports the argument of gradual development of suicidal tendencies (over time) associated with a range of potential causes.

2.2. Historical Perspective: Causal Analysis on Social Media

Our work is relevant to causal analysis of human behavior on social media. Recent approaches are developed to study *'why online users post fake news'* (Cheng et al., 2021), *beliefs and stances behind online influence* (Mather et al., 2022), and *causal explanation analysis on social media* (Son et al., 2018). The work of Son et al. (2018) is the closest to ours in that it detects a connection between two discourse arguments to extract a causal relation based on annotated Facebook data. However, the dataset is limited and is not publicly available; thus, no recent developments are observed. To address this issue, we annotate the Reddit dataset (the nature of Reddit data is different from that of Facebook) and further categorize causal explanations.

3. Annotation Scheme

3.1. Inferences from Literature

Potential reasons behind mental illness may be detected in posts that refer to insomnia, weight gain, or other indicators of worthlessness or excessive or inappropriate guilt. Underlying reasons may include: *bias or abuse* (Radell et al., 2021), loss of *jobs or career* (Mandal et al., 2011), physical/emotional illness leading to, or induced by, *medication* use (Smith, 2015; Tran et al., 2019), *relationship* dysfunction, e.g., marital issues (Beach and Jones, 2002), and *alienation* (Edition and others, 2013). This list is not exhaustive, but it is a starting point for our study, giving rise to five categories of reasons (plus 'no reason') for our automatic causal analysis: *no reason*, *bias or abuse*, *jobs and careers*, *medication*,[2] *relationship*, and *alienation*.

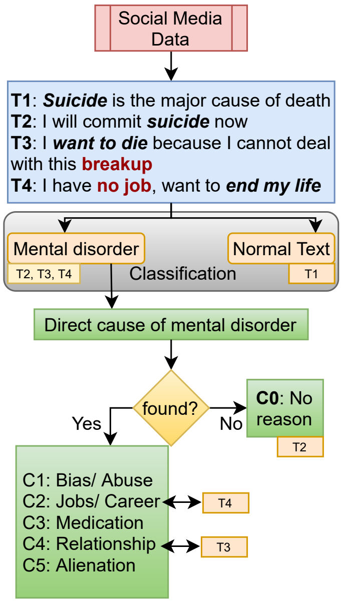

Social Media Data
T1: ***Suicide*** is the major cause of death
T2: I will commit ***suicide*** now
T3: I ***want to die*** because I cannot deal with this **breakup**
T4: I have **no job**, want to ***end my life***
Mental disorder
T2, T3, T4
Classification
Normal Text
T1
Direct cause of mental disorder
found?
Yes
No
C0: No reason
T2
C1: Bias/ Abuse
C2: Jobs/ Career
C3: Medication
C4: Relationship
C5: Alienation
T4
T3

Figure 2: Architecture of the causal analysis for mental health in social media posts

3.2. Annotation Task

Table 2 presents examples of data annotation involving the labeling of *direct causes* of mental health disorders in social media posts. There are two types of annotations: *cause category* and *Inference*. The *Inference* column contains textual data which represents the actual *reason* behind mental disorders. This *inferred reason* is further classified as one of the six different causal categories.

3.3. Problem Formulation

The architecture for our automatic causal analysis is shown in Figure 2. Social media text is provided to prediction/classification algorithms that filter out *non-mental disorder* from posts. The remaining *mental disorder* posts are then analyzed to detect reasons behind users' depression or suicidal tendencies. Finally, the reasons are classified into 5 causal categories and one 'no reason' category.

More formally, we introduce the problem of Causal Analysis of Mental health on Social media (CAMS) as a multi-class classification problem. We extract a set of social media posts as $p = p_1, p_2, p_3, ..., p_n$ for n number of posts. We *interpret the cause* for every i^{th} post p_i as C_{p_i} and *classify* it into one of the predetermined categories $y = y_0, y_1, y_2, y_3, y_4, y_5$ where y_0: 'no reason', y_1: 'bias or abuse', y_2: 'Jobs and careers', y_3: 'Medication', y_4: 'Relationship', and y_5: 'Alienation' as y_{p_i}.

3.4. Guidelines for Annotation

Our professional guidelines support annotation of the post p_i with *causal inference* C_{p_i} and class y_{p_i}. The

[2] We recognize 'medication' as both an *indicator* of physical/emotional illness (e.g., an intent to alleviate illness) and a potential *cause* of illness (e.g., medication-induced depression).

Text	Cause	Inference
That’s all I can really say. Nothing is worth the effort... I don’t think I am capable of taking steps to improve my life, because I just don’t even fucking care. Why try... I just... ugh...	Alienation	Nothing is worth the effort
Does anyone feel like the only person that could understand your depression would be someone else that was depressed? It might suck you into a place that you don’t want to be in again.	No reason	-
God help me.... I know I should go to the hospital. I know I have to keep fighting....if only to prove to my children, cursed with these genetic tendencies of mine, that life is worth living. My scars and cynicism are just a little too hard for anyone who tries to stay around too long.	Medication	go to the hospital, scars and cynicism, genetic tendencies
I hate my job .. I cant stand living with my dad.. Im afraid to apply to any developer jobs or show my skills off to employers..I dont even have a car rn... I just feel like a failure.	Jobs and Careers	hate my job, feel like failure
5 of my closest friends from high school have stopped responding to my calls or texts. i thought it was just a phone issue at first, but it is too unlikely of just coincidence.	Relationships	5 of my closest friends, stopped responding
...Then, on the way to the pub, a group of girls basically called me unattractive. Funny how girls are never shy about calling me ugly, but they’re apparently too shy to “approach me”.	Bias or Abuse	girls call me unattractive

Table 2: Sample of the CAMS dataset for causal analysis in the format of $< text, cause, inference >$

Range	Interpretation
< 0:	Less than chance agreement
0.01–0.20:	Slight agreement
0.21– 0.40:	Fair agreement
0.41–0.60:	Moderate agreement
0.61–0.80:	Substantial agreement
0.81–0.99:	Almost perfect agreement

Table 3: Interpretation of resulting values of Fliess’ Kappa agreement study (McHugh, 2012).

guidelines are developed through collaborative efforts with a clinical psychologist and a rehabilitation counselor. Student annotators label posts with their *causal inference* and *causal class*. The latter are annotated as follows:

1. **No reason**: When there is no reason that identifies the cause of mental disorder in the post.

   > C_{y_0}: [“I just want to die”, “Want to end my life”.]

2. **Bias or abuse**: A strong inclination of the mind or a preconceived opinion about something or someone. To avoid someone intentionally, or to prevent someone from taking part in the social activities of a group because they dislike the person or disapprove of their activities. It includes body shaming, physical, sexual, or emotional abuse.

   > C_{y_1}: [“No one speak to me because I am fat and ugly.”, “It has been 5 years now when that horrible incident of ragging shattered down all my confidence”.]

3. **Jobs and careers**: Financial loss can have catastrophic effects on mental illness, relationships and even physical health. Poor, meaningless and unmanageable education, unemployment, unaffordable home loans, poor financial advice, and losing a job are some of the major concerns. It includes gossiping and/or social cliques, aggressive bullying behavior, poor communication and unclear expectations, dictatorial management techniques that don’t embrace employee feedback. The educational problems like picking up courses under some external pressure and poor grades are also part of this category.

   > C_{y_2}: [“cant afford food or home anymore”, “I do not want to read literature but my parents forced me to do so. Not happy with my grades”]

4. **Medication**: The general drugs and other antiviral drugs can increase the risk of depression. The habit of using substances and alcohols can aggravate the problem of mental disorders. Moreover, medical problems like tumors, cancer, and other prolonged diseases can boost the presence of mental depression.

   > C_{y_3}: [“I am chain smoker, want to quit, but I cant. My life is mess”, “tried hard to leave drugs but this dire craving is killing me..”]

5. **Relationships**: When two people or a group of people fight, it may lead to a relationship or friendship drifting apart, for example, regular fights, breakups, divorce, mistrust, jealousy, betrayal, difference in opinion, inconsistency, conflicts, bad company, non-commitment, priority, envy. Problems like bad parenting and childhood trauma are also part of this category.

   > C_{y_4}: [“Cannot deal with this breakup anymore”, “He dumped me and its killing me”]

6. **Alienation**: Alienation is the feeling of life being worthless even after doing everything. There may be indicators of meaninglessness, loneliness, tired of daily routines, powerlessness, normlessness, isolation, and cultural estrangement.

   > C_{y_5}: [“I don’t know why am I living, everything seems to be meaningless”]

The student annotators were trained by experts (clinical psychologist and rehabilitation counselor) to pick those words and/ or phrases through which they have identified the y_{p_i} of the post p_i, and rank them. The student annotators followed these guidelines thoroughly.

3.5. Annotation Perplexity

The judgement of reasons behind online users' intentions is a complex task for human annotators, generally due to mentions of *multiple reasons* or the presence of *ambiguity in human interpretations*. Causal analysis can be viewed as a multivariate problem, resulting in multiple labels. The annotation scheme is not sophisticated enough to capture all the aspects of this phenomenon. We thus propose perplexity guidelines to simplify the task and facilitate future annotations. Our mental health therapists and social NLP practitioners have constructed perplexity guidelines to handle the trade-off between task complexity and simplicity of the annotation scheme. The *perplexity* guidelines are:

1. **Multiple reasons in the post**: There are some posts with multiple reasons for conveyed feelings. To resolve this, annotators must find a *root cause* among the *direct causes* mentioned by the user.

   > Example 1: I was of 11 years since when i realized and facing constant ignorance of my parents. 8 yrs later i lost my first girlfriend and alcoholic since then. My beer belly and obesity has made people biased towards me. I have lost everything and want to end up now.

   In Example 1, the root cause of the mental disorder is *negligence of parents*. Thus, this post is assigned the *Relationship* category. That is, we handle multiple causes by prioritizing the root cause or the most emphasized reason by the user. This *perplexity guideline* reduces annotation ambiguity and helps develop better models for automation of this task in the near future.

2. **Ambiguity in human interpretations**: The subjective nature of causal analysis makes this task even more complicated for human annotators. The six different causes are not atomic in nature. The human interpretation of the same post and the same inference may vary, even among experts.

   > Example 2: I wish I could stay alone somewhere and cry my self to sleep. I wish i won't wake up.

   Example 2 contains some important words like *alone* and *cry* which convey the category as *Alienation*. However, two out of three annotators considered this to be the *No reason* category. As this is the subjective decision of every human annotator, we leave it at their discretion. However, the final category assigned by the human expert (following human annotation) is based on "majority rules," in this case, 'no reason.'

3. **Subject of intent in the post**: Many posts refer to the depression of loved ones and other acquaintances. Given that the goal of this task is to identify the cause behind mental depression of online users, experts agree that such posts are candidates for causal analysis.

   > Example 3: I love to do prepare meals for my cousin because I think he is suffering from depression due to his car accident last month.

   In Example 3, the user is talking about their cousin who is purportedly suffering from depression. This text is passed through classification to detect *depression*; however, the third person usage precludes detection of a reason behind this condition. Although the user presents their own perspective on the reason for their cousin's condition (car accident), the input must include self-reported evidence for the reason. Thus, this example is annotated as *No reason*.

The professional training and guidelines are supported by perplexity guidelines. We have further deployed student annotators to label the dataset after they annotate the first 25 posts under the supervision of experts.

4. CAMS Dataset

We introduce a new language resource for CAMS and elucidate the process of data collection (4.1) and data annotations (4.2). We further discuss the challenges and discuss future research directions (4.3). We make our dataset and the source code publicly available for future use.

4.1. Overview of Data Collection

We demonstrate the efficacy of our annotated scheme as follows:

1. We collect 3362 Reddit posts which are available with subreddit r/depression using Python Reddit API Wrapper (PRAW). Experts remove empty and irrelevant instances from the crawled dataset resulting in 3155 samples.

2. We leverage the existing SDCNL dataset comprised of $1,896$ posts: 1517 training samples and 378 testing samples, assumed as the cleaned dataset.

3. We combine these two corpora, introducing them as the CAMS dataset, which is further annotated by our trained student annotators.

4. We consult mental health practitioners, a clinical psychologist and a rehabilitation counselor, to verify the combined dataset.

4.2. Dataset Annotations

After verification of the dataset by experts, three duly trained student annotators manually annotate the data in the format: <text, cause, inference> as shown in Table 2. In this section, we discuss the annotation process,

Class	Crawled corpus			SDCNL Training data			SDCNL Test data		
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
No reason	1	508	59.78	1	1785	68.58	1	1562	84.85
Bias or Abuse	6	2109	347.48	5	4378	227.24	6	578	149.80
Jobs and career	13	2258	228.28	17	2771	255.70	20	1481	206.95
Medication	5	1552	213.83	3	3127	205.86	11	1124	165.60
Relationship	2	3877	229.35	14	2739	240.08	9	756	202.56
Alienation	3	1592	153.86	1	899	147.01	12	683	145.67

Table 4: Word length variation in posts across causal classes for each dataset.

verification by experts, and validation using statistical tests.

Annotation is carried out manually by annotators who are proficient in the language. They work independently for each post and follow the given guidelines. Each annotator takes one hour to annotate about $15 - 25$ Reddit posts and 180 and 90 hours to annotate the *crawled* and *existing* dataset, respectively. The annotations are obtained as three separate files.

Several challenges have emerged during the annotation process, e.g., a countable (< 10) set of non-English posts. For such cases, augmented guidelines instruct the annotator to mark the post as *No reason*. We recommend the removal of non-English posts as the CAMS dataset is proposed for English only. The annotated files are verified by a clinical psychologist and a rehabilitation counselor. This verification is performed over the annotations given by our trained annotators without bringing this to their knowledge and experts have given the final annotations.

We further validate three annotated files using Fliess' Kappa inter-observer agreement study. The agreement study for the crawled dataset is found to be 64.23%. We also study the inter-agreement annotations for the existing dataset as 73.42% and 60.23% for testing and training data of SDCNL, respectively. The trained annotators agree with 61.28% agreement among the annotators for the CAMS dataset. The resulting values are interpreted as per Table 3. Despite the increased subjectivity of the task, the student annotators *substantially agree* with their judgements.

4.3. Discussion

Existing work on causal analysis is associated with finding discourse relations among words and identifying the segments that represent the reason behind the intended information. We extend this work to find the category of the cause using these interpreted segments. Since we are extending the causal interpretations to automatic categorization, our work is not directly comparable to any of the existing works. However, we glean insights into the characteristics of the CAMS dataset through further analysis. This section examines the word length of posts in the dataset (4.3.1) and varying number of instances in each class (4.3.2). Additionally, we discuss the social nature of the dataset (4.3.3).

4.3.1. Length of the Posts

The length of posts varies from a few characters to thousands of words. One of the major challenges for automation is the construction of a multi-class classifier that is suitable for posts of varying word lengths. One of the possible solutions to this challenge is to extract the inference from the post and classify it using the inference text. We choose to explore this in the near future. Table 4, shows that, although there is consistency in the average number of words among all the classes, there is a huge variation in the word counts across posts.

4.3.2. Imbalanced Dataset

Table 5 shows that the number of posts for every cause varies widely, perhaps signifying that causes of mental health disorders are not well-distributed in society.

Cause	CC	Train_S	Test_S	CAMS
No reason	292	332	70	694
Bias or abuse	122	194	35	351
Jobs/careers	399	181	48	628
Medication	410	170	43	623
Relationship	956	297	91	1344
Alienation	976	340	92	1408
Total	**3155**	**1517**	**379**	**5051**

Table 5: Sample distribution of the CAMS dataset for different causes where *CC* is Crawled Corpus, *Train_S* is the Training data of SDCNL dataset, *Test_S* is the Test data of SDCNL dataset, and *CAMS* column contains the total number of samples in the dataset for each cause.

In the crawled corpus, the highest number of samples is observed for the *Relationship* and *Alienation* causal categories, which is perhaps an indicator that our society is less equipped to deal with issues pertaining to *'near & dear ones'* and *'loneliness / worthlessness'*, respectively. The number of posts with *'no reason'* is smaller in the crawled corpus due to the cleaning of the dataset. Interestingly, there are fewer posts assigned *'Bias or abuse'*—less than half of each of the two additional categories: *'Jobs and careers'* and *'Medication'*.

4.3.3. Social nature of the dataset

Our expert clinical psychologists have explored the social nature of the dataset, in light of our analysis above. During re-annotation of existing dataset, the prevalence of some causes, e.g., *Alienation* and *Relationship*, point to the importance of the ability to take a societal pulse on a regular basis, especially in these unprecedented times of pandemic-induced distancing and shut-downs. Other problems, e.g., *Jobs and careers* and *bias and abuse* depend upon good governance. The problem of *medication* depends upon technological/medical advances and accessible healthcare, or lack thereof. Additionally, online users often feel depressed but do not address any specific reason behind it, indicating that inferring relevant causes is a challenge if one uses NLP alone.

4.4. Ethical Considerations

We emphasize that the sensitive nature of our work necessitates that we use the publicly available Reddit dataset (Haque et al., 2021) in a purely observational manner(Broer, 2020). We claim that the given dataset does not disclose the user's personal information or identity. We further acknowledge the trade-off between privacy of data and effectiveness of our work (Eskisabel-Azpiazu et al., 2017). We ensure that our CAMS corpus is shared selectively and is subject to IRB approval to avoid any misuse. Our dataset is susceptible to the biases and prejudices of annotators who were trained by experts. There will be no ethical issues or legal impact with this causal analysis of mental illness.

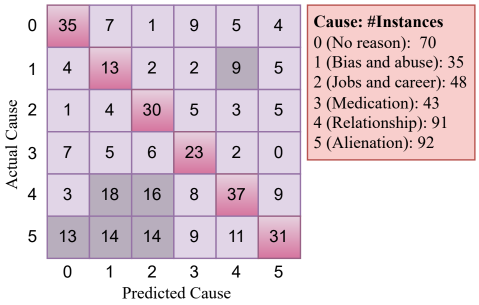
5. Corpus Utility for Machine Learning

We include traditional multi-class classifiers trained on CAMS training dataset and evaluate it on the CAMS test data. We choose the following Recurrent Neural Network (RNN) architectures: Long Short Term Memory (LSTM) model, Convolution Neural Network (CNN), Gated Recurrent Unit (GRU), Bidirectional GRU/LSTM (Bi-GRU/Bi-LSTM) and other hybrid models. In this section, we discuss the experimental setup (5.1) and analyze the results (5.2).

5.1. Experimental Setup

We use the existing re-annotated SDCNL dataset for experimental results and analysis. We clean the dataset, pre-process the posts and then use GloVe[3] word embedding with 100 dimensions trained on Wikipedia for each token. We further set up RNN architectures with default settings ($lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-08$) for the batch-size of 256. The categorical *Cross Entropy* loss function and the *ADAM* optimizer are used to perform the back-propagation learning on 20 epochs.

The number of samples for three classes of the existing SDCNL dataset (*'Bias or abuse', 'Jobs and careers', and 'Medication'*) are very few in comparison to the other three classes. We add 120+120+120 samples from the crawled corpus, to help balance the number of instances across the classes. As a result, the number of training samples increases from 1517 to 1877. We use these training data to build and validate the multi-class classifier. We test this classifier on the 379 sample test data and analyze the results. The evaluation metrics used for this task are *F1-measure* and *Accuracy*.



Actual Cause \ Predicted Cause	0	1	2	3	4	5
0	35	7	1	9	5	4
1	4	13	2	2	9	5
2	1	4	30	5	3	5
3	7	5	6	23	2	0
4	3	18	16	8	37	9
5	13	14	14	9	11	31

Cause: #Instances
0 (No reason): 70
1 (Bias and abuse): 35
2 (Jobs and career): 48
3 (Medication): 43
4 (Relationship): 91
5 (Alienation): 92

Figure 3: Confusion matrix of CNN+LSTM model on test data

5.2. Results and Discussion

We use multi-class classifiers to find causal categories and obtain the results reported in Table 6. We test our performance with both machine learning and deep learning approaches. The two top-performing machine learning algorithms are based on *Logistic Regression* and *Support Vector Machine*. Whereas the former outperforms all existing techniques, the latter shows comparative results with deep learning models. The hybrid model, *CNN-LSTM*, attains the best performance among all deep learning mechanisms with 47.78% accuracy. It is interesting to observe that the results are consistent for all the classifiers with few exceptions for classes *Bias or abuse* and *Medication*. We further analyze the best performing deep-learning classifier *(CNN+LSTM)* below.

5.2.1. Error Analysis

The accuracy of multi-class classification is found to be around 40% to 50%. We undertake a comprehensive error analysis to explore the intricacies of our task.

1. *Cause classification error*: We obtain the confusion matrix for *CNN+LSTM* as shown in Figure 3. We highlight the cells with more than 40% incorrect predictions. The predictions for *Alienation* and *Relationship* are incorrect and overlap with *Bias or Abuse* and *Medication*. This is due to complex interactions, as illustrated in the following perceivable overlap between *Bias or Abuse* and *Relationship*:

   > Example 4: My friends are ignoring me and I am feeling bad about it. I have lost all my friends.

   Example 4 is associated with *biasing* and *friendship*, in a case where someone feels ostracized by their friends. The emphasis on*friends* tips the balance in favor of the class *Relationship*. However,

[3] https://nlp.stanford.edu/projects/glove/

Classifier	F1: C0	F1: C1	F1: C2	F1: C3	F1: C4	F1: C5	Accuracy
LR	**0.63**	**0.28**	0.54	**0.46**	0.46	**0.53**	**0.5013**
SVM	0.54	0.23	**0.56**	0.44	**0.48**	0.45	0.4670
LSTM	0.54	**0.27**	0.52	0.46	0.42	**0.51**	0.4595
CNN	0.56	**0.27**	0.51	0.42	0.46	0.38	0.4378
GRU	0.51	**0.27**	0.54	**0.47**	0.48	0.42	0.4541
Bi-LSTM	0.55	0.12	0.41	0.23	0.44	0.50	0.4351
Bi-GRU	**0.57**	0.14	**0.55**	0.46	0.49	0.39	0.4568
CNN+GRU	0.51	0.14	0.49	0.36	0.27	0.45	0.4027
CNN+LSTM	0.54	0.22	0.54	**0.47**	**0.54**	0.47	**0.4778**

Table 6: Experimental results with CAMS dataset. F1 is computed for all six causal classes: 'No reason' (C0), 'Bias or abuse' (C1), 'Jobs and careers' (C2), 'Medication' (C3), 'Relationship' (C4), 'Alienation' (C5).

the major challenge is to train the model in such a way that it understands the inferences and then chooses the most emphasized *causal category* using optimization techniques. We view this challenge as an open research direction.

2. *Overlapping class*: The *overlapping problem* of classes is observed with ambiguous results for samples, e.g., for *Relationship* and *Alienation*. This class representation problems can be resolved with data augmentation (Ansari et al., 2021) and demarcation of boundaries among classes. In a real-time scenario, demarcation of fixed boundaries is not possible due to subjectivity of the task. We recommend the approximation of a newly built model over handcrafted / automated features accordingly. Further, we obtain low performance for *class 1 ('Bias or abuse')* due to annotators' perceived overlap with classes 4 and 5 (*'Relationship' and 'Alienation'*). Future work is needed to mitigate such uncertainty. For example, delineation of discourses within the text would support a more definitive interpretation and reliable annotation.

3. *Semantic Parsing*: In a multi-class classification task, as the length of posts varies over a wide range, one may choose to summarize every post before applying a classifier. Our experiments with YAKE (Campos et al., 2020) for keyword extraction yielded results that are further deteriorated. From this we determine that it is important to identify the *causal interpretation* from the full text in order to perform multi-class classification. A future avenue of research involves the exploration of *discourse relations* to identify segments that represent independent causes that underlie mental health disorders.

5.2.2. Implications and Limitations

The CAMS dataset provides a means for exploring the identification of reasons behind mental health disorders of online users. The notion of *causal categorization* is defined and used to proactively identify cases where users are at potential risk of mental depression and suicidal tendencies. The results of this work may be employed to explore the impact of *unemployment*, *low grades*, etc. Our analysis may also be useful for the study of online behavior. A major limitation of the CAMS dataset is that the users may intentionally post their intent of mental disorder on social media, e.g., for deliberately making new friends. In this work, we have assumed that the data has no such biasing.

6. Conclusion

This paper introduces the task of causal analysis to identify the reasons behind mental *depression and suicidal tendencies* (intent). We have introduced CAMS: a dataset of 5051 instances, to categorize the *direct causes* of mental disorders through mentions by users in their posts. We transcend the work of Level 0 studies, moving to the next level (Level 1) for causal analysis. Our work is the combined effort of experts in the field of Social Natural Language Processing (Social NLP), including a rehabilitation counselor and clinical psychologists (CPsych). We have further implemented machine learning and deep learning models for causal analysis and found that *Logistic Regression* and *CNN+LSTM* gives the best performance, respectively. In the future, we plan to extend the problem of causal analysis of mental health detection on social media as a multi-task problem. Another major future challenge for this work is the generation of explanations for multi-class classification, by leveraging causal analysis within the CAMS framework.

7. Acknowledgement

We acknowledge our three student annotators: Simran jeet Kaur, Astha Jain and Ritika Bhardwaj. We also acknowledge Amrith Krishna for his kind support and for proofreading this manuscript. Publication costs are funded by NSERC Discovery Grant (RGPIN-2017-05377), held by Vijay Mago, Department of Computer Science, Lakehead University, Canada.

8. References

Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O., and Bingol, H. O. (2018). Detecting suicidal ideation on forums: proof-of-concept study. *Journal of medical Internet research*, 20(6):e215.

Ansari, G., Garg, M., and Saxena, C. (2021). Data augmentation for mental health classification on social media. *arXiv preprint arXiv:2112.10064*.

Beach, S. R. and Jones, D. J. (2002). Marital and family therapy for depression in adults.

Broer, T. (2020). Technology for our future? exploring the duty to report and processes of subjectification relating to digitalized suicide prevention. *Information*, 11(3):170.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Cao, L., Zhang, H., Feng, L., Wei, Z., Wang, X., Li, N., and He, X. (2019). Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. *arXiv preprint arXiv:1910.12038*.

Cao, L., Zhang, H., and Feng, L. (2020). Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*.

Cheng, L., Guo, R., Shu, K., and Liu, H. (2021). Causal understanding of fake news dissemination on social media.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018). Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *27th International Conference on Computational Linguistics*, pages 1485–1497. ACL.

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., and Tao, C. (2018). Xa-bilstm: A deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1624–1627. IEEE.

Conway, M. and O'Connor, D. (2016). Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Edition, F. et al. (2013). Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21.

Eskisabel-Azpiazu, A., Cerezo-Menéndez, R., and Gayo-Avello, D. (2017). An ethical inquiry into youth suicide prevention using social media mining. *Internet Research Ethics for the Social Age*, 227.

Falotico, R. and Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Garg, M. (2021). Quantifying the suicidal tendency on social media: A survey. *arXiv preprint arXiv:2110.03663*.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R., and Pathak, J. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525.

Haque, A., Reddi, V., and Giallanza, T. (2021). Deep learning for suicide and depression identification with unsupervised label correction. *arXiv preprint arXiv:2102.09427*.

Ji, S., Yu, C. P., Fung, S.-f., Pan, S., and Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Klonsky, E. D., Pachkowski, M. C., Shahnaz, A., and May, A. M. (2021). The three-step theory of suicide: Description, evidence, and some useful points of clarification. *Preventive medicine*, 152:106549.

Lin, H., Jia, J., Qiu, J., Zhang, Y., Shen, G., Xie, L., Tang, J., Feng, L., and Chua, T.-S. (2017). Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833.

Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pages 343–361. Springer.

Mandal, B., Ayyagari, P., and Gallo, W. T. (2011). Job loss and depression: The role of subjective expectations. *Social Science & Medicine*, 72(4):576–583.

Masuda, N., Kurahashi, I., and Onari, H. (2013). Suicide ideation of individuals in online social networks. *PloS one*, 8(4):e62262.

Mather, B., Dorr, B. J., Dalton, A., de Beaumont, W., Rambow, O., and Schmer-Galunder, S. M. (2022). From stance to concern: Adaptation of propositional analysis to new tasks and domains. In *Findings of the Association for Computational Linguistics: Human Language Technologies*, Dublin, Ireland, May.

McGinty, E. E., Presskreischer, R., Han, H., and Barry, C. L. (2020). Psychological distress and loneliness reported by us adults in 2018 and april 2020. *Jama*, 324(1):93–94.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Nam, H., Ha, J.-W., and Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307.

Pierce, M., Hope, H., Ford, T., Hatch, S., Hotopf, M., John, A., Kontopantelis, E., Webb, R., Wessely, S., McManus, S., et al. (2020). Mental health before and during the covid-19 pandemic: a longitudinal probability sample survey of the uk population. *The Lancet Psychiatry*, 7(10):883–892.

Pirina, I. and Çöltekin, Ç. (2018). Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Radell, M. L., Abo Hamza, E. G., Daghustani, W. H., Perveen, A., and Moustafa, A. A. (2021). The impact of different types of abuse on depression. *Depression research and treatment*, 2021.

Sawhney, R., Joshi, H., Flek, L., and Shah, R. (2021). Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428.

Shah, F. M., Haque, F., Nur, R. U., Al Jahan, S., and Mamud, Z. (2020). A hybridized feature extraction approach to suicidal ideation detection from social media post. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 985–988. IEEE.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., and Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Shing, H.-C., Resnik, P., and Oard, D. W. (2020). A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137.

Smith, H. R. (2015). Depression in cancer patients: Pathogenesis, implications and treatment. *Oncology letters*, 9(4):1509–1514.

Son, Y., Bayas, N., and Schwartz, H. A. (2018). Causal explanation analysis on social media. *arXiv preprint arXiv:1809.01202*.

Tran, B. X., Ho, R., Ho, C. S., Latkin, C. A., Phan, H. T., Ha, G. H., Vu, G. T., Ying, J., and Zhang, M. W. (2019). Depression among patients with hiv/aids: research development and effective interventions (gapresearch). *International journal of environmental research and public health*, 16(10):1772.

Turcan, E. and McKeown, K. (2019). Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.