"Let Me Tell You About Your Mental Health!" Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention

Manas Gaur
Knoesis Center
Dayton, Ohio
manas@knoesis.org

Ugur Kursuncu
Knoesis Center
Dayton, Ohio
ugur@knoesis.org

Amanuel Alambo
Knoesis Center
Dayton, Ohio
amanuel@knoesis.org

Amit Sheth
Knoesis Center
Dayton, Ohio
amit@knoesis.org

Raminta Daniulaityte
CITAR Center
Dayton, Ohio
raminta.daniulaityte@wright.edu

Krishnaprasad Thirunarayan
Knoesis Center
Dayton, Ohio
tkprasad@knoesis.org

Jyotishman Pathak
Cornell University
New York, NY
jyp2001@med.cornell.edu

ABSTRACT

Social media platforms are increasingly being used to share and seek advice on mental health issues. In particular, Reddit users freely discuss such issues on various subreddits, whose structure and content can be leveraged to formally interpret and relate subreddits and their posts in terms of mental health diagnostic categories. There is prior research on the extraction of mental health-related information, including symptoms, diagnosis, and treatments from social media; however, our approach can additionally provide actionable information to clinicians about the mental health of a patient in diagnostic terms for web-based intervention. Specifically, we provide a detailed analysis of the nature of subreddit content from domain expert's perspective and introduce a novel approach to map each subreddit to the best matching DSM-5 (Diagnostic and Statistical Manual of Mental Disorders - 5th Edition) category using multi-class classifier. Our classification algorithm analyzes all the posts of a subreddit by adapting topic modeling and word-embedding techniques, and utilizing curated medical knowledge bases to quantify relationship to DSM-5 categories. Our semantic encoding-decoding optimization approach reduces the false-alarm-rate from 30% to 2.5% over a comparable heuristic baseline, and our mapping results have been verified by domain experts achieving a kappa score of 0.84.[1]

[1]Resources created as a part of the study will be made available upon request to the corresponding author.

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271732

CCS CONCEPTS

• **Information systems** → *Information retrieval*; • **Computing methodologies** → *Natural language processing*; *Machine learning*; • **Applied computing** → *Health informatics*;

KEYWORDS

Reddit; Mental Health; DSM-5; Semantic Encoding and Decoding; Medical Knowledge bases; Drug Abuse Ontology; Semantic Social Computing

ACM Reference Format:
Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. *"Let Me Tell You About Your Mental Health!"*, Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention . In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271732

1 INTRODUCTION

Engagement of users in social media has grown from 22M in 2005 to 204M in 2015 and is expected to reach 220M (3 quarter of US population) by 2022. 80% of these social media users search for health-related information, and 50% of them look for medical specialists[2]. Reddit platform is extensively used to seek or give advice on a variety of health problems. Mental health-related conversations are particularly frequent[3]. Subreddits are forums dedicated to particular topics on Reddit, and they are created by informed users to consolidate posts on a domain of interest. For instance, "Ask a Doctor[4]" and "Mental Health[5]" are two of the most popular health-related subreddits that have nearly 50K members each. In

[2]https://goo.gl/gSFNgx
[3]https://health.good.is/features/internet-therapy-reddit
[4]https://www.reddit.com/r/AskDocs/
[5]https://www.reddit.com/r/mentalhealth/

this paper, we focus on 15 mental health-related subreddits: Suicide watch, Opiates, Opiates recovery, Schizophrenia, Crippling Alcoholism, BipolarSOs, BipolarReddit, Bipolar, Anxiety, Borderline Personality Disorder (BPD), Autism, Aspergers, Addiction, Self Harm, and Stop Self Harm. These subreddits have been identified as popular by domain experts [14].

Reddit platform enables free, unobtrusive, and honest sharing of mental health concerns because a patient is completely anonymous and so can open up without worrying about any social stigma or other consequences; thus, the content is less biased and of high quality compared to the content shared in survey questionnaires and interviews [17]. For example, consider the subreddit posts on mental health: (1) *Help needed for a guy who is ridden with anxiety, compulsivity, impulsive behavior, and bipolar.* (2) *How can I cope up with my unhealthy and chaotic thoughts?.*

When a patient visits a mental health clinic, practitioners interact with the patient and assess their condition utilizing responses to the questionnaires (e.g., PHQ-9) and interviews that are conducted in a clinical setting. Apart from these two traditional input sources, social media content of a patient can be *consensually* leveraged to gain additional insights. Extracting meaningful information from social media sources is important for learning about the epidemiology of mental health disorders, understanding the attitudes and informedness of patients about the conditions and treatment needs. It also has potential to create web-based interventions and resources for those who are identified as having serious mental health issues. However, the use of such information requires de-identification and the consent of patients following the patient confidentiality requirements. In this case, a patient can provide his/her handle on Reddit along with consent for the specified use, and his/her content can be collected and analyzed to predict presence and progression of mental health condition. Although we did not conduct a study on human subjects, such analysis can be performed after obtaining appropriate approvals (i.e., IRB). Thus, a mental health professional (MHP) can aggregate a variety of signals and personalized insights from diverse sources including patient Electronic Health Records (EHR), questionnaires, interviews, and social media.

Motivating Scenario: John Doe is a senior undergraduate student, who started experiencing suicidal thoughts and shared them on multiple subreddits in search of self-diagnosis. For instance, consider his Reddit post: *"The feeling of inadequateness and having wasted my 23 years of life in hopelessness, made me feel to kill myself"*. Informed mental health professionals (MHPs) on subreddits as well as users who had experienced similar feelings before, replied to this post sharing relevant information. These replies lead John to visit a mental health clinic that optionally provides a service to collect the social media handle of its clients following the patient's written informed consent, to enable web-based intervention. A custom tool can extract suicidal signals by comprehensively analyzing John's volunteered social media content and providing the findings to the mental health specialist. The clinic can, optionally request John to fill a corroboratory questionnaire and have an interview to obtain a reliable diagnosis of John's condition.

In particular, web-based intervention strategies that involve social media can immensely benefit the under-served population by addressing their mental health condition in a cost-effective manner [41]. Moreover, identifying symptoms early using social media content can augment proactive and preventative healthcare to the traditional reactive approach [43]. We are using our DSM-5 lexicon for robust classification of mental health disorders formalized in DSM-5 chapters. DSM-5 chapters (aka DSM-5 categories) are formal guidelines for categorization of mental health disorders that were created by domain experts. Further details about DSM-5 are provided in Section 3. Although the subreddits are created by informed MHPs from the medical community, they are not guaranteed to be based on the DSM-5 guidelines; hence, an appropriate mapping between the content of these subreddits to the corresponding DSM-5 categories, using reliable, globally-recognized and publicly available medical knowledge sources, is essential.

Our approach analyzes the subreddit posts to determine the DSM-5 categories and includes multiple stages of processing and analysis of posts. The processing is done by using two lexicons: one created from the subreddits and another related to medical knowledge bases. The lexicon from subreddits has been created by extracting N-grams and topics using Latent Dirichlet Allocation (LDA) [29]. On the other hand, the lexicon for DSM-5 has been constructed by utilizing publicly available knowledge bases, namely, ICD-10[6], SNOMED-CT[7], and DataMed[8], along with enriched Drug Abuse Ontology[9] (DAO)[4]. The analysis includes multi-class classification that maps subreddits into DSM-5 categories for labeling, utilizing features generated through word embeddings, lexicons, and the DAO ontology. Note that we use DSM-5 chapters and DSM-5 categories interchangeably (where DSM-5 chapters is the terminology used in DSM-5 manual for mental health categories). Accordingly, DSM-5 lexicon is the set of concepts related to DSM-5 categories.

We claim two primary contributions of this study: (1) We developed and evaluated a novel approach to map subreddits into DSM-5 categories (section 5.3), and (2) designed a semantic weighting mechanism that better relates Reddit and DSM-5 embedding spaces to improve multi-class DSM-5 classification (Section 5.6). We call the latter approach: Semantic Encoding and Decoding Optimization (SEDO). Apart from these technical contributions, we provide the following two resources to the medical research community for further use: (i) a domain-specific lexicon based on DSM-5 chapters utilizing ICD-10, SNOMED-CT and DataMed (section5.2), and (ii) an enriched Drug Abuse Ontology (DAO) with mental health-related terminology and slang terms from Reddit.

The paper combines the effectiveness of probabilistic language models, the richness of structured medical knowledge bases and a core optimization approach that was also utilized in Zero Shot Learning (ZSL) [36] to classify unstructured content to DSM-5 categories. ZSL[10] is a learning methodology that involves mapping between embedding spaces of data samples (image or text) and class labels assuming that data is unlabeled. At a high level, our approach can match a patient on social media platform to mental health resources and has potential for web-based intervention. In our study, we motivate the need and relevance of our methodology by first reviewing the necessary related work in Section 2. In Section

[6] https://bioportal.bioontology.org/ontologies/ICD10
[7] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[8] https://datamed.org
[9] http://wiki.knoesis.org/index.php/DAO
[10] http://isis-data.science.uva.nl/tmensink/docs/ZSL17.web.pdf

3, we elaborate on preliminary terminologies used in our approach. In Section 4, we detail the characteristic properties of the Reddit data and the need for entropy analysis. In Section 5, we describe our methodology addressing key contributions. In Section 6, we demonstrate the efficacy and domain-specificity of our approach and discuss its generality. In Section 7, we present our conclusions and identify directions for future work.

2 RELATED WORK

Predictive analysis on social communication platforms has attracted growing attention from research community of late. Previously, the issues in various domains that includes social, political and healthcare, have been studied to provide solutions for real world problems [21]. For performing such study, it was essential to transform the social media sphere to a data-science bubble where statistical and semantic learning can be employed. [40] details the feature engineering stages of the data science pipeline highlighting different feature types and features corresponding to different Twitter-based applications. Furthermore, [21] identifies the need for topical analysis for feature generation and enhancing classification in social media across multiple societal applications. Efficacy of topical analysis for classification of social media data has been further improved by utilizing the linked information in multiple domain-specific knowledge bases [43]. ConceptNet[11], and SentiWordNet[12] are some of the concept level semantic dictionaries for improving text categorization on Twitter/Reddit [3].

A significant body of work has been done in medical and clinical domains addressing mental health disorder classification. While traditional clinical records of patients suffering from mental health disorders provide an understanding of possible diagnosis of a mental health disorder to the doctor [37], additional insights can be obtained by analyzing the communication exchanged among users on social media. Subreddits[13] such as Meddit (r/medicine), r/physicianassistant, r/nursepractitioner, r/doctorsthatgame have become venues for patients to obtain immediate support, mostly before the onset of the disease. In this regard, the classification of the content in these subreddits based on a formal guideline for mental health diagnosis is useful for medical decision-making process. Accordingly, an association of the self-proclamations on social media with DSM-5 categories for legitimate specialist mediation is essential and an unexplored domain in computational medical information retrieval.

A recent study on Twitter employed a semi-supervised model to relate depression symptoms expressed on Twitter in terms of answers to PHQ-9 questionnaire[14] [42]. They analyzed 23M tweets from over 45K users to reveal nine depressive symptoms utilizing the background information along with a generative model. In [35], the Reddit platform was exploited for understanding the content related to anxiety. They employed N-grams, LDA, word embedding, and emotional analysis for binary classification of a post on Reddit. Besides, there has been work on understanding the linguistic features of the posts from patients suffering from mental health issues. In [35] Linguistic Inquiry Word Count (LIWC) has been utilized as a psycholinguistic knowledge base to improve the entropy of Anxiety related Reddit posts. In [30], LIWC, Brown clustering, and LDA, along with perplexity measures, are used for identifying people suffering from schizophrenia. One of the recent relevant topics on social media has been the "CopyCat Suicide," also known as "Werther effect." In a study by [20], a subreddit r/SuicideWatch was monitored for post-suicide progression in posting behavior on Reddit. A topical analysis approach was employed to quantify anxiety, anger, negative emotion, and demeaning tone that emphasizes suicidal and self-harm tendencies. Behavioral features of the content, like social engagement, ego network, linguistic style, vocabulary usage, emotions, sentiment, intention and mentions of medication was used to estimate the risk of mental health problems [11]. Social media acts as a rich source of users who declare their mental health conditions after being diagnosed. Utilizing their content, one can identify possible signals that could indicate depression. In [10], social activity and language features were extracted from self-reported depressed individuals on Twitter to train a probabilistic classifier to identify traits of clinical depression. However, mental health conditions are not only commorbid and etiologic but are also dynamic as one mental illness can cause another, making identification of conditions subordinate to diagnosis. [12] utilizes Reddit data, specifically interactional and linguistic features along with propensity score, to identify distinctive markers that estimate the likelihood of suicide ideation from current mental illness. Moreover, shared-task based exercises are promoting development of models for extracting mental health condition markers for classification. A CLPsych 2015 challenge developed the task comprising of three experiments: binary classification of content as (a) depression or control, (b) depression or Post Traumatic Stress Disorder (PTSD), and (c) control or PTSD [33]. While we rely on DSM-5 chapters for mental-health classification of Reddit main content, [14] utilizes subreddit labels for classification of mental-health conditions. Moreover, [14] uses word embeddings in their first phase of binary classification involving 5000 most frequent terms, 99.9% of which are stopwords. We anticipate that they were not able to extract a sufficient number of mental health-related terms from the text; so they included stop words to be able to train the word embedding model.

However, leveraging existing medical knowledge bases and universally accepted DSM-5 categories can improve classification accuracy substantially by contextualization. A DSM-5-based classification has been utilized to analyze posts shared on Twitter such as by gleaning depression traits and their variation over time in different tweets[15]. In [32], DSM-5 categorization was utilized to validate the thematic variation and equivalence between PTSD (Post-Traumatic Stress Disorder), Anxiety, and Depression on Reddit. We believe that DSM-5 classification of the social media content can ultimately support the health and well-being by timely and intelligent matching of a patient (help/advice seeker) with appropriate mental health specialist (caregiver).

3 PRELIMINARIES

Reddit posts can be used with supervised learning approach, due to the presence of labels for respective subreddits [14]. Bigrams (heroin addict, quit smoking) and trigrams (Narcissistic Personality Disorder, Cannabis Use Disorder) have been more informative

[11] http://conceptnet.io/
[12] http://sentiwordnet.isti.cnr.it/
[13] http://cliniciantoday.com/6-communities-for-healthcare-professionals-on-reddit/
[14] http://www.phqscreeners.com/sites/g/files/g10016261/f/201412/PHQ-9_English.pdf
[15] http://blog.medicalgps.com/social-media-and-healthcare-10-insightful-statistics

than unigrams for classification in the context of medical texts [23]. Apart from n-grams, Latent Dirichlet Allocation (LDA) has been used for topic-based feature extraction from text [5]. Perplexity measure can identify the number of topics, but [16] shows that perplexity of LDA model is not reflective of human judgment. However, a metric based on the coherence of the topics provides acceptable number of topics. Since, the vocabulary related to mental-health (e.g., symptoms, medication) are shared across multiple subreddits, it is difficult to generate good discriminative features using word statistics. Further, traditional features such as emotions, sentiments, part of speech tags, and morphological structure of sentences are not sufficient to distinguish posts across different mental health conditions. Instead, we propose to analyze the posts in a subreddit using relevant available medical background knowledge and label the subreddit in terms comprehensible to domain experts. The inclusion of appropriate context through semantic features utilizing medical knowledge bases, namely, SNOMED-CT, ICD-10, and DataMed, provides a better categorization of subreddit on mental health in terms of DSM-5 categories discussed below.

3.1 Diagnostic and Statistical Manual of Mental Disorders (DSM-5)

DSM-5 is the taxonomic and diagnostic manual developed and published by the American Psychiatric Association. It is an authoritative guide for mental healthcare professionals for the diagnosis of mental disorders. It includes 20 chapters (see Table 3), consistent with ICD-10 and NIH's Research Domain Criteria (RDoc)[16] for mental health.

3.2 Drug Abuse Ontology (DAO)

The Drug Abuse Ontology (DAO) is a domain-specific conceptual framework for interconnecting sets (named "classes") of drug-focused and health-related concepts. DAO was initially designed for the PREDOSE project [4] that analyzed web-forum posts related to buprenorphine use [7]. It was expanded further for eDrugTrends[17] [8, 22] and eDarkTrends[18] projects that focuses on cannabis [8], synthetic cannabinoid [22] and opioid-related data [8]. The DAO includes representations of mental health disorders and related symptoms that were developed following DSM-5 classification. The advantage of DAO is that it is not limited to medical terminology, but also includes commonly used lay and slang terms for mental health conditions and associated symptoms. For example, references for "Opioid Use Disorder" include such lay terms as "addicted to opioids," "addicted to heroin," "pain pill addict." References to the feeling of "Anxiety" or "Anxious" include such terms as "troubled," "with my stomach in knots", "antsy", "worried", and "agitated."

4 EXPLORATORY DATA ANALYSIS

In this section, we provide an overview of the data by performing topical and statistical analysis of the data. Then, we perform entropy analysis to illustrate randomness, and select a part of the data for the subsequent investigation. Data used in this paper has been consensually obtained from [14]. The data describes the social communication on Reddit related to mental health comprising 2.5M posts (main posts, replies, and comments) by 268,104 users. The dataset spans nine year period between 2006 and 2015.

Reddit Category	R	Avg.1	Avg.2	T. #Users
Addiction(ADD)	0.53	2.44	6.50	3211
Crippling Alcoholism(CRP)	0.39	15.16	3.09	17491
Anxiety(ANX)	0.02	1.38	6.13	50718
Opiates(OPT)	0.32	18.98	3.66	23701
Aspergers(ASP)	0.43	9.69	5.51	12849
Opiates Recovery(OPR)	0.43	10.27	6.80	5552
Autism(AUT)	0.41	4.85	5.63	8043
Schizophrenia(SCZ)	0.43	7.5	5.58	3275
Bipolar(BPL)	0.46	9.75	5.79	13699
Self Harm(SLF)	0.45	7.24	5.02	6389
BipolarSOs(BPS)	0.41	4.64	9.25	712
Stop Self Harm(SSH)	0.47	4.15	5.71	4240
Bipolar Reddit(BPR)	0.44	9.16	6.15	9750
Suicide Watch(SCW)	0.32	3.78	7.71	93789
BPD	0.43	7.25	6.78	6775
Median	**0.43**	**7.25**	**5.79**	**8043**

Table 1: User and content based statistical characteristics of the Reddit data. R: Ratio of number of main posts to total number of posts, Avg.1: Average number of main posts per user, Avg.2: Average number of sentences per main post and T.#Users: Total number of users.

4.1 Statistical Characteristics of the Dataset

As indicated in Table 1, "main" posts constitute about 40% of the total content in each subreddit. A user generates, on an average, seven main posts that start a conversation in each subreddit, and each main post has around six sentences. In particular, Crippling Alcoholism, Opiates, and Opiates-Recovery have large numbers of main posts per user, where number of sentences per main post is shorter than the overall median (5.79). This may be because people tend to be concise and tacit, and may end up using multiple main posts to make their conditions explicit. On the other hand, anxiety, addiction, bipolar-related, opiates recovery and suicide watch subreddits have above the median number of sentences per main posts, suggesting that these users are explicit and detailed about their conditions, medications, and symptoms. Addiction subreddit seems to be the most engaging subreddit compared to other 14 subreddits as it constitutes above 50% of the main posts, indicating much higher engagement from people (Table 1).

4.2 Topical Analysis

The dataset comprises of posts across 15 subreddits, and we performed a topical analysis to determine the relevance to mental health condition. We used the LDA, LDA over bigrams and Skip-grams to identify topics in each subreddit (as shown in Figure 2). Table 2 depicts the association of topics with subreddits.

We observe an overlap between topics of conversations on Aspergers and Autism. Moreover, BPD and crippling alcoholism show topical similarity with bipolar and addiction respectively. We also see an excessive use of the phrase "child autism" in posts, where the parents report symptoms shown by their child, or verify the presence/absence of disorder, or seek advice for their child. Bipolar disorder is also known as the manic-depressive illness characterized by extreme shifts in mood. The bipolar subreddit contains a set of topics like weight gain, bipolar 2, mood swings, depressive episode, medication or diagnosis, and rapid cycling[19]. BipolarSOs is a community of users where either one or both of the individuals who are in relationships have been diagnosed as bipolar. The topics in such a community include support groups, divorce-related issues,

[16] https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml
[17] https://medicine.wright.edu/citar/edrugtrends
[18] http://wiki.knoesis.org/index.php/EDarkTrends
[19] https://www.webmd.com/bipolar-disorder/guide/rapid-cycling-bipolar-disorder#1

SubReddit	Topics of Interest
Addiction	video game addiction, hypersexuality, drug addiction, work pressure, withdrawal symptoms
Anxiety	depression, anxiety, cognitive distortions, panic attacks, hopelessness, final exam, physical sensation
Aspergers	fear uncomfort, fear social interactions, maldevelopment fine motor, motor skills, trouble sleeping.
Autism	child autism, early intervention, ABA therapy, autistic son, sensory issues, eye contact.
Bipolar Reddit	weight gain, bipolar 2, mood swings, depressive episode, medication or diagnosis, rapid cycling.
Bipolar-SOs	support groups, bipolar relationships, divorce related issues, mood swings, mutual understanding.
BPD	impulsivity, mood swings, antisocial conduct, personality disorder.
Crippling-Alcoholism	drink day, quit drinking, cold turkey, liquor stores, steel reserve, mild depression
Opiates	buying oxycodone, selling oxycodone, pain management, chronic pain, opiate addiction, alienation
Opiates-Recovery	suboxone buprenorphine recovery, live life, alcohol anonymous, support.
Schizophrenia	auditory hallucinations, paranoid behavior, psychotic episodes, depression, anxiety, side effects, schizophrenia medications
Self-harm	cutting, scars, burning, frequency, feeling bad, emotional pain, self-injury, coping mechanisms.
Stop-Self-harm	started cutting, started feel worse, feel urge cut, feel guilty, treating scars, burns, propeling thought love, friendship, social interaction
Suicide-Watch	suicide ideation, suicide thought, commit suicide, substance addict, hang, life worth, meet people, talk family, talk friends.

Table 2: Sample of Topics identified from different subreddits.

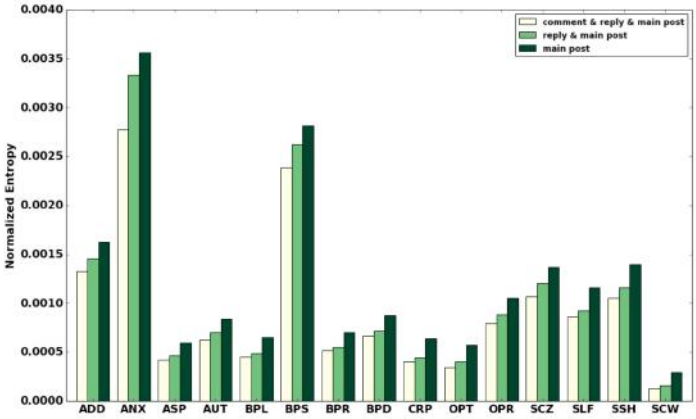

Figure 1: Entropy based analysis of different subreddits concerning change in information content of main posts after adding replies and comments.

mood swings (like bipolar), and how to develop mutual understanding. Opiates subreddit complements the opiate-recovery subreddit which comprises posts on recovery, live life, alcohol anonymous, and buprenorphine-related topics.

4.3 Entropy Analysis

Entropy measures the randomness in a dataset, which we can adapt to quantify the coherence of Reddit posts and their relevance to the Reddit topic. A distinction should be made between main posts that start a conversation, comments that follow the main posts and replies that follow comments since the text in these categories differ in their content as well as their length. Accordingly, we hypothesize that main posts contain more relevant information as compared to replies and comments. Based on the Reddit metadata[20], a reply is a type of post that owns a permalink whereas a comment owns permalink of the main post. To test our hypothesis, we initially performed an entropy analysis over the main posts, and gradually included the content from comments and replies by an iterative process [9]. We have appended comments and replies that have at least three sentences in the main posts. For computing Normalized Entropy (**NE**) of a subreddit **S**, we used the following Equation (1), which takes as argument a set of unique words for a subreddit.

$$NormalizedEntropyNE(S) = \frac{-\sum_{w \in UW_S} P_w \cdot logP_w}{|UW_S|} \quad (1)$$

where $\mathbf{P_w}$ is probability of occurrence of a word w in a Reddit main post file, $\mathbf{UW_S}$ is the set of unique words in **S**, and $|\mathbf{UW_S}|$ is total number of unique words in a subreddit **S**. As evident from Figure 1, the inclusion of comments and replies did not alter the entropy significantly. From Figure 1, we make two important observations: (1) Our data comprising of main posts is homogenous and predictable as entropy value is close to zero, and relevant to subreddit's focus. (2) Removing replies and comments will not affect the performance of the classification as they contribute very little to the predictability and homogeneity of the content. Upon the entropy analysis, we get 1.1M main posts that represents 56% reduction in content. In the subsequent section, we explain our methodology for analyzing the 1.1M Reddit main posts across 15 mental-health subreddits.

5 METHODOLOGY

In this paper, we have leveraged DSM-5 manual, along with domain-specific knowledge bases such as SNOMED-CT, ICD-10, and DataMed, to improve classification and mapping accuracy of subreddits to DSM-5 symptoms. In contrast, the inclusion of LIWC[21] features did not improve its performance. The overall architecture capturing our approach is shown in Figures 2 and 4. Sections 5.2 and 5.3 explain Figure 2 and Section 5.6 explains Figure 4[22].

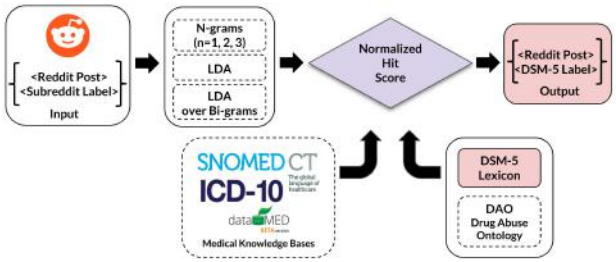

Figure 2: Procedure for generating DSM-5 Label/Category for each subreddit by calculating normalized hit score using N-gram, LDA, LDA over bigrams and DSM-5 Lexicon.

5.1 Characteristic Features

During our feature modeling phase, we define a set of 13 features that quantify the content within a subreddit syntactically rather than semantically. These features enable every post within a subreddit to be represented as a vector of dimension 1 X 13. We organized our feature set into three broad groups: Horizontal Linguistic Features (HLFs), Vertical Linguistic Features (VLFs), and Fine-Grained Features (FGFs). Contextual Features (or embedding of a subreddit post) with Modulations (CFwM) and without Modulations (CFw/oM) are two additional feature set created using Word2Vec on Reddit. We then carried out experiments to evaluate the performance of these 13 features for multiclass classification

[20] https://www.reddit.com/dev/api/

[21] LIWC Features:https://goo.gl/BT2d2Z, https://goo.gl/YQAZPN

[22] Images in Figure 4 are taken from Noun Project

(see Table 5). HLFs deal with the syntactic structure of a post revealing the expressiveness in a post. HLFs include (i) number of words used in a post, (ii) use of definite articles which aid in semantic association, (iii) presence of verb before a noun (e.g., drinking hell), (iv) presence of two nouns at the end of a sentence (e.g., bipolar and depression), (v) sequencing part of speech triples (e.g., Adjective-Noun-Verb (ANV) triple[2]), (vi) subordinate conjunctions that capture time, place, cause or effect in a sentence, (vii) average number of first-person pronouns, and number of pronouns in a sentence. VLF comprise those features which capture narrow though specific linguistic and non-linguistic attributes[23]. VLFs include (viii) number of part of speech tags, (ix) average similarity between different posts, (x) noun chunks[13], and (xi) document similarity. In FGF, we incorporate sentiment score of each post in 15 subreddits. (xii) Sentiment analysis is performed by comparing the words in a post to the word list with sentiment score in [31]. We also used (xiii) LabMT[24] to generate sentiment scores of a post. For generating the contextual features, we trained a word-embedding model[25] on a corpus containing all the posts from 15 subreddits, and the model generated a vocabulary of 12,808 words from the Reddit corpus having 300 dimension vector for each word. We used *sum* function to create word vector for each post from among the aggregation functions listed in [39], since it provided good performance. For CFw/oM features, we used Word2Vec model to generate 300 dimension embedding of a word in a post and summed the word-vectors to generate an embedding of the post of a subreddit. However, such embedding does not reflect the relative importance of a word in a post in a subreddit. In CFwM, we multiply each word's embedding with TF-IDF score and sum word vectors to generate a weighted embedding of the post in a subreddit.

5.2 Creation of DSM-5 Lexicon

We first created a DSM-5 lexicon by leveraging well-known resources such as SNOMED-CT[26], ICD-10[27], DataMed[28], and DAO[29]. DSM-5 lexicon contains n-grams associated with each of the DSM-5 categories. These medical knowledge bases (except DataMed) are stored and index in a graph structure and searching was performed with respect to each DSM-5 category. For each mental health disorder in the DSM-5 category, (1) we look for its two-hop parents and four-hop children and (2) extend the search using a Depth First Search graph traversal. It was set empirically so as to extract concepts that are contextually relevant to the DSM-5 category searched. Furthermore, we restricted our search space to those concepts that are labeled as *disorder* in the medical knowledge base. Table 3 shows the number of concepts that were identified following the search. These concepts are unigrams (U), bigrams (B), and/or tri-grams (T). Consider an intuitive example with concepts and associated SNOMEDCT IDs in brackets: *Mild Bipolar* [13313007] is the child of *Bipolar Disorder* [13746004] and have the following children: *Mild manic bipolar I* [71984005], *Mild depressed bipolar I* [74686005], *Mild mixed bipolar I* [43769008]. These children are siblings of *Rapid Cycling* [133091000119105]. In our Reddit corpus, the drug-abuse related categories form a substantial portion (48%) of the dataset in size, and the above knowledge bases have sparse information about the drug-abuse related concepts and relations. Hence, we incorporated the concepts and relationships from domain experts-curated Drug Abuse Ontology (DAO), to significantly improve the quality and precision of mapping. Additionally, the incorporation of slang terms from DAO to match and process the informal social media data improved both the coverage and recall. Table 3 documents the improvement on the number of concepts that relate to DSM-5 categories extracted from the aforementioned knowledge bases and the DAO, after the inclusion of slang terms from Reddit data.

DSM-5 Category	Before SL Terms	After SL Terms
Dissociative Disorders (DSD)	20	20
Anxiety Disorders (AXD)	40	87
Substance Use &Addictive Disorder (SAD)	39	123
Schizophrenia Spectrum (SCS)	77	77
Sleep Wake Disorder (SWD)	14	19
Paraphilic Disorders (PRD)	14	14
Trauma & Stressor Related Disorder (TSD)	25	28
Gender Dysphoria (GND)	15	15
Depressive Disorders (DPD)	71	107
Neurodevelopmental Disorders (NDD)	25	53
Sexual Dysfunctions (SXD)	23	23
Personality Disorders (PND)	76	98
Disruptive, Impulse, Control & Conduct Disorder (DICD)	34	34
Psychotic Disorders (PSD)	85	87
Bipolar & Related Disorders (BRD)	75	84
Elimination Disorders (ELD)	18	18
Obsessive-Compulsive & Related Disorder (OCD)	43	60
Feeding & Eating Disorders (FED)	32	39
Neurocognitive Disorders (NCD)	80	80
Suicidal Behavior/Ideation (SBI)	34	47
Number of concepts in DSM-5 lexicon & DAO	**840**	**1113**

Table 3: Improvement in number of concepts being captured after adding Slang Terms (SL) from Reddit and enrichment of DSM-5 Lexicon. SL: Slang Terms, Before SL: slang terms extracted from Medical Knowledge Bases and DAO and After SL: inclusion of DAO slang terms.

5.3 Mapping SubReddits to DSM-5 Categories

Our novel contribution is the automatic unsupervised mapping of subreddit labels to DSM-5 categories. This work is significant because manual mapping is very labor intensive given the large size of the dataset. Note that the mapping of a subreddit title cannot be done manually and exhaustively as the content is usually heterogeneous and the subreddit title does not have straightforward DSM-5 category analog. For example, in our study, we found BPD, Crippling Alcoholism, Schizophrenia, and SuicideWatch subreddits mapped to DSM-5 categories of DICD, SAD, SCS, and SBI (Table 3) respectively.

We have employed n-grams (n=1,2,3) language model to extract the most frequent and collocated terms that have matches in the DSM-5 lexicon. For the generation of n-grams, we utilized Skip-Gram model and subsampling of frequent words as explained in [28]. We performed preprocessing tasks such as removal of Stop Words[30], URLs and punctuations from the corpus before extraction of n-grams. We have found that there were no significant trigrams in main posts, so we focused on unigrams and bigrams, for the remaining analysis.

A simple approach to assess similarity between subreddit (label) and DSM-5 category. would be to count the number of exact and approximate matches between n-grams (U, B, T) in a subreddit main

[23] ftp://ftp.cs.indiana.edu/pub/gasser/Playpen/TR1/tr/node4.html
[24] http://neuro.imm.dtu.dk/wiki/LabMT
[25] https://radimrehurek.com/gensim/models/word2vec.html
[26] https://www.nlm.nih.gov/research/umls/licensedcontent/snomedpermanentpackages.html
[27] http://www.who.int/classifications/icd/en/
[28] API: https://datamed.org/
[29] http://wiki.knoesis.org/index.php/DAO
[30] http://www.lextek.com/manuals/onix/stopwords2.html,https://github.com/RxNLP/nlp-cloud-apis

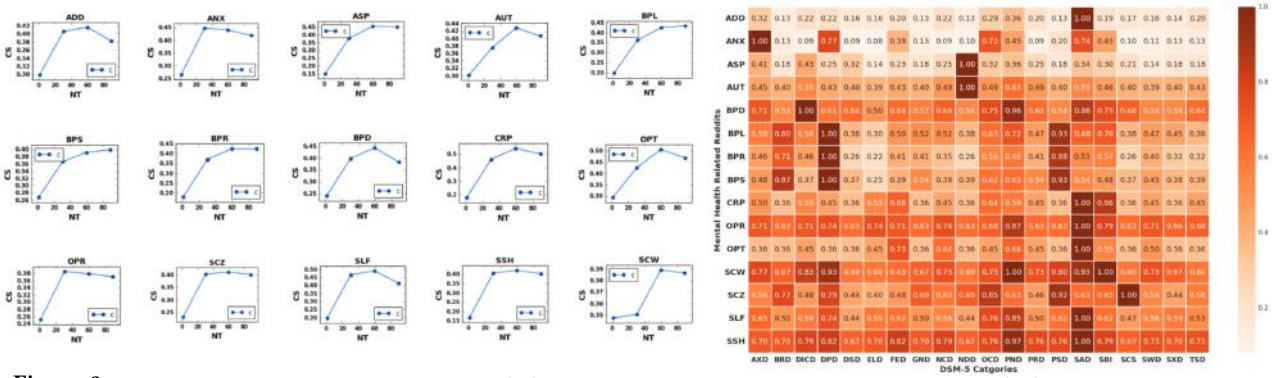

Figure 3: The graphs on the left shows coherence score (CS) based identification of optimal number of LDA topics (NT) for various subreddits. The heat map on the right shows the normalized hit score of each subreddit corresponding to each DSM-5 category.

posts and the corresponding DSM-5 lexicon. We call these matches as *Hits*(H), and they were calculated through string overlaps using set-intersection. In this procedure, approximate matches are identified when a unigram is a substring of a bigram or trigram(U-B, U-T), or a bigram is a substring of a trigram(B-T), while exact matches are simply exact string matches (U-U, B-B). The possible hits that we considered are: U-U, U-B, U-T, B-B, B-T.

For determining the best representative topics, we conducted coherence analysis using LDA[31]. Higher the topic coherence, the better it is for human interpretation.

A visualization of the "coherence score vs. num_topics" is depicted in Figure 3 (left) and the highest coherence scores were registered for number of topics in the range 50-60. We have chosen 55 as the average number of topics to pick for each subreddit for mapping them to the DSM-5 lexicon. We independently utilized LDA over subreddits and subreddit bigrams, and generated 110 sub-topics that comprises the top 2 sub-topics from the 55 topics [38]. We have computed the Hits (**H**) between LDA sub-topics of each subreddit and the DSM-5 lexicon to infer their corresponding DSM-5 category. As we have calculated the number of hits from n-grams, we combine the number of hits for LDA with the n-grams using the following equation, which we called "Normalized Hit Score" ($\mathbf{nhs}^{\mathbf{S}}_{\mathbf{D}}$):

$$H^S = \{H(ng^S, D) + H(LDA^S, D) + H(bLDA^S, D)\}_{D \in DSM-Lex}$$

$$nhs^S_D = \frac{H^S_D}{max(H^S)} \quad (2)$$

Where, **S** is the index of a particular subreddit, **D** is a set of concepts extracted from the aforementioned medical knowledge bases related to a particular DSM-5 category using the DSM-5 lexicon (DSM-Lex). $\mathbf{nhs}^{\mathbf{S}}_{\mathbf{D}}$ is defined as the ratio of $\mathbf{H}^{\mathbf{S}}_{\mathbf{D}}$ and $\mathbf{max}(\mathbf{H}^{\mathbf{S}})$, where $\mathbf{H}^{\mathbf{S}}_{\mathbf{D}}$ is the number of hits occurring between a particular subreddit **S**, and a DSM-5 category **D**, and $\mathbf{H}^{\mathbf{S}}$ is the collection of hit scores calculated for **S**, with all DSM-5 categories in the DSM-5 lexicon. $\mathbf{max}(\mathbf{H}^{\mathbf{S}})$ is the maximum hit score from the collection of a $\mathbf{H}^{\mathbf{S}}$. $\mathbf{H}(\mathbf{ng}^{\mathbf{S}}, \mathbf{D})$ is number of hits of n-grams in a subreddit **S** ($\mathbf{ng}^{\mathbf{S}}$) that matches to a **D** DSM-5 category in the lexicon. $\mathbf{H}(\mathbf{LDA}^{\mathbf{S}}, \mathbf{D})$ is the number of topics identified in a subreddit **S** ($\mathbf{LDA}^{\mathbf{S}}$), and **D**. $\mathbf{H}(\mathbf{bLDA}^{\mathbf{S}}, \mathbf{D})$ is the number of topics identified over bigram ($\mathbf{bLDA}^{\mathbf{S}}$) that overlapped with a set **D** of DSM-5 categories in the lexicon.

[31] https://radimrehurek.com/gensim/

The heat map in Figure 3 (right) depicts the results of our mapping. The normalized hit scores represented in the heat map ranges from 0 to 1 based on the likelihood of a subreddit to be mapped to a DSM-5 category. Ultimately, our approach maps a subreddit to a DSM-5 category only if the normalized hit score ($\mathbf{nhs}^{\mathbf{S}}_{\mathbf{D}}$) is 1.0. For instance, the subreddit *OPR* is mapped to the *SAD* as shown in Table 4. In general, for each subreddit, we obtain **nhs** for every DSM-5 category and assign the DSM-5 category with the highest score. However, it is noteworthy that **nhs** for the best and the second-best mappings may be close to each other, so disambiguation becomes necessary. For instance, while the **nhs** for the mapping SSH-SAD is 1.0, 0.97 for SSH-PSD, or 1.0 for both mappings of SCW with SBI and PSD. In such cases, we consult with the domain experts for disambiguation. In total, we have five mappings that have a second best **nhs** above 0.90, and we provided the n-grams, LDA and bLDA topics associated with them to the domain experts for resolution. Incidentally, in all the five cases, the best mapping selected by the domain expert was the one with the **nhs** of 1.0.

Reddit Main Posts	Subreddit Label	DSM-5 Label
I did not do **valium**, it scared me. I am doing **oxycontin**, and wnt to ask safe amount of **ghb**.	OPT	SAD
Aspergers run in my family. ..... My maternal grandFather was an **Aspire**,.... My mother ..... are aspire while another maternal uncle is not....	ASP	NDD
It's good to talk to her about her **anxiety attacks** when they're not happening. Kind of like coming up with a game plan..	ANX	AXD
It has been a few days **major depression**, a couple weeks 'fine', a couple weeks **mild depressed**, a few days fine, a week or two **manic**,Is it **worse bipolar?**.Seriously, if I've only barely survived **suicidal tendencies** with **short depressions**, I think would killed myself	BPL	DPD

Table 4: Paraphrased illustrative posts labeled with DSM-5 label after the mapping. For interpreting the acronyms, see Table 1 for subreddit labels and Table 3 for DSM-5 category.

5.4 Creation of a Coarser Dataset

Once the most prominent DSM-5 category has been identified for a subreddit, we replace the subreddit labels in the dataset with the corresponding DSM-5 category. We consider such a dataset as coarser dataset because fewer DSM-5 category labels than subreddit labels were identified based on our mapping procedure and will be used for training our model. The mapping procedure is employed at the subreddit level, whereas the classification is performed at the post level. Thus, posts in our dataset were labeled with their corresponding subreddit's label. The resulting dataset was imbalanced since some of the DSM-5 labeled categories have a larger size of

samples whereas some have smaller, particularly after the mapping procedure. Two domain experts for mental health have evaluated a random set of 118 samples with an average number of 5 sentences and 70 words per post, along with the subreddit and the DSM-5 labels, as shown in Table 4. We obtained 0.84 Kappa score with the distribution: the number of correct matches was 48, the number of incorrect matches was 9, and the remaining 61 posts were left blank by the domain experts because of their vague content or inadequate information for annotation.

5.5 Baseline Approach

The classification of mental health conditions over Reddit data was recently performed by [14] utilizing the feature sets HLF, VLF, and FGF. As mental health classification based on DSM-5 chapters has not been previously investigated, we have taken [14]'s methodology as our baseline, for comparison. We have also experimented with logistic regression, SVM (linear and radial basis kernels) and Adaboost, and found them to be ineffective for imbalanced and heterogeneous dataset. Hence, we eventually built our approach utilizing only Random Forest, as its effectiveness was also reported by prior works [15, 25, 26, 34]. As our dataset is highly imbalanced, we have utilized a resampling method, SMOTE[32], to synthetically oversample the minority portion of the dataset. Ensemble learning algorithms hybridized by sampling methods have proved to outperform in the cases of imbalanced data[19]. We have also employed the Term frequency and Inverse Document Frequency (TF-IDF) model that resulted in a feature space that comprises of 39,699 words before the application of singular value decomposition (SVD) to reduce the dimension down to 300. We augment this reduced feature space to the existing feature set of the baseline. The results of these experiments has been summarized in Table 5.

5.6 Semantic weighting through Encoding and Decoding Optimization (SEDO)

In this section, we explain our semantic weighting algorithm, called SEDO, and its role in the DSM-5 multi-class classification.

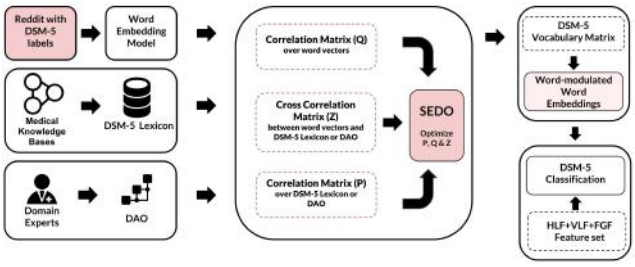

Figure 4: Proposed approach to DSM-5 classification using SEDO based word-vector modulation together with HLF, VLF and FGF features.

As the background knowledge is modeled through DSM-5 lexicon, we have incorporated this knowledge in the classification process utilizing SEDO. We introduce SEDO as an approach for obtaining a *discriminative weight* matrix between the DSM-5 lexicon and Reddit word embedding space after optimization utilizing the Sylvester equation [1]. Although the Sylvester equation has been used in computer vision within the context of ZSL [18], its utilization in creating a *discriminative weight* matrix between unstructured (e.g. Reddit) and structured data (DSM-5 Lexicon) has not been investigated. SEDO requires: (1) embedding space for each category in the DSM-5 lexicon, and (2) embedding space of each word in Word2Vec vocabulary created from Reddit data. SEDO formulate the function $\mathbf{E(R, D)}$ as minimizing the Frobenius norm[33] of difference between Reddit and DSM-5 embedding spaces (Equation (3)).

$$E(R, D) = min_W \{||R - W^T D||_F^2 + \delta||WR - D||_F^2\} \quad (3)$$

where R represents the Reddit word embedding space, D the DSM-5 embedding space, and W the weight matrix to be minimized.

As we are mapping the Reddit (unstructured) embedding space to the DSM-5 (structured) embedding space, we call this process as decoding, and from DSM-5 to Reddit data as encoding. In Equation (3), the part before the "+" represents the encoding of DSM-5 categories to Reddit data embedding space, while the part after "+" represents the decoding of Reddit data to DSM-5 categories. Furthermore, Equation (3) is a convex function; hence, we can expect a global optimal solution. Differentiating the Equation (3) with respect to "W" for minimization, involves following properties: $\mathbf{Tr(W^T D)} = \mathbf{Tr(D^T W)}$ (cyclic property of trace) [34] and $\mathbf{Tr(R)} = \mathbf{Tr(R^T)}$. A positive, symmetric and quasi-separable[35] matrix show such properties. Hence, Equation (3) is transformed to

$$E(R, D) = min_w \{||R^T - D^T W||_F^2 + \delta||WR - D||_F^2\} \quad (4)$$

$$\frac{d(E(R, D))}{d(W)} = -2(D)(R^T - D^T W) + 2\delta(WR - D)(R^T) \quad (5)$$

$\frac{\mathbf{d(E(R,D))}}{\mathbf{d(W)}} = \mathbf{0}$ can be solved using techniques for Sylvester equation. δ is a parameter for regularization during the optimization phase.

$$-DR^T + DD^T W + \delta WRR^T - \delta DR^T = 0 \quad (6)$$

$$(DD^T)W + W(\delta RR^T) = (1 + \delta)DR^T; 0 < \delta < 1 \quad (7)$$

Equation (7) represents the Sylvester equation form: $\mathbf{PX + XQ = Z}$ where $\mathbf{P}$ is $\mathbf{DD^T}$ and $\mathbf{Q}$ is $\mathbf{RR^T}$, which represents self-correlation between DSM-5 and Reddit embedding spaces respectively, and $\mathbf{Z}$ is $\mathbf{DR^T}$ represent cross-correlation between DSM-5 and Reddit embeddings.

DSM-5 Embedding Space: Each category in the DSM-5 Lexicon is represented by a set of concepts. These concepts can be U, B, or T. We created embedding of each category of DSM-5 using trained Word2Vec model on Reddit corpus. The embedding was created using summation operation over word vectors of all the concepts within a DSM-5 category. It resulted in a 300 dimension embedding for each category. Hence, DSM-5 embedding space is of dimensions 20 X 300. Self-correlation of DSM-5 embeddings ($\mathbf{DD^T}$) creates a matrix of dimensions 20 X 20. Similarly, self-correlation of Reddit word-embedding space ($\mathbf{RR^T}$) creates a matrix of dimension 12808 X 12808. Cross-correlation between $\mathbf{RR^T}$ and $\mathbf{DD^T}$ creates a matrix of dimensions 20 X 12808.

Since there are two self correlation matrices and a cross-correlation matrix, we utilize Sylvester optimization [27] function that generates a *discriminative weight* matrix (W) of dimension 20 X 12808.

[32] https://goo.gl/QWqXfq

[33] http://mathworld.wolfram.com/FrobeniusNorm.html

[34] http://www2.math.ou.edu/~dmccullough/teaching/slides/maa2010.pdf

[35] https://goo.gl/mcgvcZ

Approaches	Features	FL/V	Lexicon	Modulation	P	R/TPR	F	FAR
Baseline Model	HLF + VLF + FGF	13/-			0.42	0.47	0.44	0.30
Baseline Model with SMOTE	HLF + VLF + FGF	13/-			0.41	0.45	0.43	0.27
Balanced Random Forest	HLF + VLF + FGF	13/-			0.41	0.49	0.45	0.15
Balanced Random Forest	HLF + VLF + FGF + TF-IDF	313/39699			0.60	0.49	0.54	0.14
	HLF + VLF + FGF + CFw/oM	313/12808			0.60	0.54	0.57	0.13
	HLF + VLF + FGF + CFwM	313/12808		TF-IDF	0.55	0.50	0.52	0.13
	HLF + VLF + FGF + Tweet2Vec	313/3039345			0.54	0.48	0.51	0.15
Balanced Random Forest	HLF + FGF+ VLF	313/12808	DSM-5 lexicon w/o DAO	SEDO weights	0.87	0.77	0.82	**0.03**
	HLF + FGF+ VLF	313/12808	DSM-5 lexicon w/ DAO w/o SL	SEDO weights	0.87	0.80	0.83	**0.03**
	HLF + FGF+ VLF	313/12808	DSM-5 lexicon w/ SL w/o DAO	SEDO weights	0.85	0.82	0.83	**0.06**
	HLF + FGF+ VLF	313/12808	DSM-5 lexicon w/ DAO w/ SL	SEDO weights	0.88	0.83	0.85	**0.025**

Table 5: Classification Performance. FL: Feature Length, TPR: True Positive Rate, FAR : False Alarm Rate. w: with, w/o : without, FL: Feature Length, V: Vocabulary Size of the TF-IDF or word embedding model. The size of lexicon has been stated in Table 3.

Each value in the matrix gives the weight for a word in Reddit and DSM-5 category. A generic Twitter word2vec model[36] (Tweet2Vec) was also employed instead of the domain-specific model; however, Sylvester equation failed to converge to generate the desired weight matrix. It happened because the DSM-5 embedding space was sparse and not all concepts were able to generate vectors using Twitter Word2Vec model. The modulation through *SEDO weights* provides enrichment of the embedding space by a weight matrix created through our approach and the results of this procedure is reported in Table 5.

6 RESULTS AND DISCUSSION

Table 5 summarizes the results of our experiments and the improvement over the baseline discusssed in Section 5.5. We employed oversampling, TF-IDF for feature extraction and modulating word embeddings, and domain specific knowledge to generate discriminative weight matrix for modulating contextual features.

We evaluated our approach based on False Alarm Rate (FAR)[24], Precision (P), Recall(R)/TPR and F1-measure (F1). We emphasize Recall and FAR in our evaluation, since these metrics are essential for web-based intervention where reliability, robustness and efficacy are important. In particular, in clinical setting, a misclassification can cause serious risk to a patients' mental health because it can either lead to false diagnosis and wrong treatment, or to missed opportunity for early intervention. The baseline approach resulted in a FAR of 30% and sensitivity of 47% as the model has a bias towards some DSM-5 categories that cover a significant portion of the training and testing data. From Table 5, we observe that over-sampling procedure yielded a reduction of 10% in the FAR but sacrificed TPR by 4.2%. We also experimented with an existing implementation of balanced random forest [6] which bootstraps the process of oversampling by randomly drawing samples from majority class. Bootstrapping sampling procedure obtained 4% gain over the baseline and a significant reduction of FAR by 50%. In our baseline model, we have used VLF, HLF and FGF features which are generic and invariant to syntactic variations. We also employ TF-IDF to generate additional features for each post over the baseline achieving 16% reduction in FAR and 2% increase in Recall.

The addition of CFw/oM to the existing feature set that comprises of VLF, HLF, and FGF, improved the TPR by 13% and reduced the FAR by 57% compared to the baseline. It shows that involving domain knowledge in the form of contextual features resulted in a modest improvement in classifier performance. However, modulating the contextual features by TF-IDF scores (CFwM) decreased the TPR by 7% with no change in FAR. Such a behavior occurred as words identified as important by TF-IDF are not contextually relevant. Hence, it is pivotal to improve the weighting scheme of the word vectors by incorporating domain knowledge. Utilizing our novel semantic weighting scheme, we obtained a significant increase in TPR and F1-measure, by 39%, and 46% respectively. Further improvement of 4% was seen in TPR after enriching the DSM-5 lexicon with the DAO because a majority of the Reddit posts were related to opiates, opiates-recovery, and crippling alcoholism, and social media posts usually have *slang terms*[37] that can confuse the classifier. As the identification and addition of such lingo to the ontology is a challenging and tedious task, we have used the list of *slang terms* incorporated in the DAO (Table 3). As a result, the inclusion of slang terms reduced the FAR by 17%, compared to the experiment with absence of *slang terms*.

Based on a series of experiments, we observed two noteworthy points: (1) Modulating the word vectors in 300-dimensional space using information in Medical Knowledge Bases reduces false alarm, and (2) Contextualizing the word embedding using context-dependent slang terms and DAO, significantly reduces misclassification.

7 CONCLUSION AND FUTURE WORK

Our overall goal was to use main posts in mental health related subreddits, voluntarily shared by users, to be able to better assess mental health issues, uncover signals that may indicate mental health problems, and eventually determine appropriate mental health care providers. In order to operationalize this goal, we propose an approach to map the content to more rigorously defined DSM-5 categories to better characterize the nature of the mental health-related content. These DSM-5 categories can then be used to better reflect clinical aspects related to mental problems and accordingly point to appropriate mental health specialists for web-based intervention. Identification of mental health conditions using social media is not diagnostic, but can provide insights to the MHP based on social media content of the patient and potentially enable appropriate care. In this study we map a subreddit to a DSM-5 category, and label every post within the subreddit with the corresponding DSM-5 category. We sought to develop a semantic optimization technique (SEDO) that minimizes the distance between DSM-5 categories and Reddit content spaces, utilizing existing domain specific medical knowledge bases and Reddit main posts. Our approach generates discriminative weight matrix to perform multi-class classification by modulating the word embeddings of the Reddit content.

In the future, we plan to better interpret the presence of negation in Reddit posts to improve the classification accuracy. We will

[36] https://www.fredericgodin.com/software/

[37] https://mashable.com/2014/03/10/reddit-lingo-guide/#fgMTft2LNmqU

also explore adapting this approach for web-based intervention to the Twitter platform. Furthermore, we aim to enrich the DSM-5 lexicon with updated ICD-11, Unified Medical Language Systems (UMLS) concepts, relations and definitions, MedDRA, DrugBank, and Clinical Trials.

ACKNOWLEDGEMENT

We acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression", and National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02 "Trending: Social media analysis to monitor cannabis and synthetic cannabinoid use". Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, or NIDA.

REFERENCES

[1] Amrudin Agovic and Arindam Banerjee. 2012. Gaussian process topic models. *arXiv preprint arXiv:1203.3462* (2012).

[2] Melanie Andresen and Heike Zinsmeister. 2017. Approximating Style by N-gram-based Annotation. In *Proceedings of the Workshop on Stylistic Variation*.

[3] Erik Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. 2013. Knowledge-based approaches to concept-level sentiment analysis. *IEEE intelligent systems* (2013).

[4] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. 2013. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics* (2013).

[5] William B Cavnar, John M Trenkle, and others. 1994. N-gram-based text categorization. *Ann arbor mi* (1994).

[6] Chao Chen, Andy Liaw, and Leo Breiman. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley* (2004).

[7] Raminta Daniulaityte, Robert Carlson, Gregory Brigham, Delroy Cameron, and Amit Sheth. 2015. "Sub is a weird drug:" A web-based study of lay attitudes about use of buprenorphine to self-treat opioid withdrawal symptoms. *The American journal on addictions* (2015).

[8] Raminta Daniulaityte, Francois R Lamy, G Alan Smith, Ramzi W Nahhas, Robert G Carlson, Krishnaprasad Thirunarayan, Silvia S Martins, Edward W Boyer, and Amit Sheth. 2017. "Retweet to Pass the Blunt": Analyzing Geographic and Content Features of Cannabis-Related Tweeting Across the United States. *Journal of studies on alcohol and drugs* (2017).

[9] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Identifying relevant social media content: leveraging information diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*.

[10] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*.

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* (2013).

[12] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*.

[13] George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*.

[14] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports* (2017).

[15] Li Guan, Bibo Hao, Qijin Cheng, Paul SF Yip, and Tingshao Zhu. 2015. Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. *JMIR mental health* (2015).

[16] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning* (2014).

[17] Matthew R Jamnik and David J Lane. 2017. The Use of Reddit as an Inexpensive Source for High-Quality Data. *Practical Assessment, Research & Evaluation* (2017).

[18] Elyor Kodirov, Tao Xiang, and Shaogang Gong. 2017. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345* (2017).

[19] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* (2016).

[20] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*.

[21] Ugur Kursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2018. Predictive Analysis on Twitter: Techniques and Applications. *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, Springer-Nature* (2018).

[22] Francois R Lamy, Raminta Daniulaityte, Ramzi W Nahhas, Monica J Barratt, Alan G Smith, Amit Sheth, Silvia S Martins, Edward W Boyer, and Robert G Carlson. 2017. Increases in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis. *International Journal of Drug Policy* (2017).

[23] Raymond Lau, Ronald Rosenfeld, and Salim Roukos. 1997. Building scalable n-gram language models using maximum likelihood maximum entropy n-gram models. (1997).

[24] Neil A Macmillan and Howard L Kaplan. 1985. Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological bulletin* (1985).

[25] Matthew J Maenner, Marshalyn Yeargin-Allsopp, Kim Van Naarden Braun, Deborah L Christensen, and Laura A Schieve. 2016. Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PloS one* (2016).

[26] Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of the third workshop on computational lingusitics and clinical psychology*.

[27] Stefano Massei, Davide Palitta, and Leonardo Robol. 2017. Solving rank structured Sylvester and Lyapunov equations. *arXiv preprint arXiv:1711.05493* (2017).

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

[29] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*.

[30] M Mitchell, K Hollingshead, and G Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*.

[31] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* (2011).

[32] Albert Park, Mike Conway, and Annie T Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. *Computers in Human Behavior* (2018).

[33] D Preoţiuc-Pietro, M Sap, H A Schwartz, and L Ungar. 2015. Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

[34] Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. 2016. MIDAS: Mental illness detection and analysis via social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*.

[35] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*.

[36] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*.

[37] Joseph Thomas. 2009. Medical records and issues in negligence. *Indian journal of urology: IJU: journal of the Urological Society of India* (2009).

[38] Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*.

[39] Sanjaya Wijeratne, Lakshika Balasuriya, Derek Doran, and Amit Sheth. 2016. Word embeddings to enhance twitter gang member profile identification. (2016).

[40] Sanjaya Wijeratne, Amit Sheth, Shreyansh Bhatt, Lakshika Balasuriya, Hussein S Al-Olimat, Manas Gaur, AH Yazdavar, and Krishnaprasad Thirunarayan. 2017. Feature Engineering for Twitter-based Applications. *Feature Engineering for Machine Learning and Data Analytics* (2017).

[41] Marie Bee Hui Yap, Shireen Mahtani, Ronald M Rapee, Claire Nicolas, Katherine A Lawrence, Andrew Mackinnon, and Anthony F Jorm. 2018. A tailored web-based intervention to improve parenting risk and protective factors for adolescent depression and anxiety problems: postintervention findings from a randomized controlled trial. *Journal of medical Internet research* (2018).

[42] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.

[43] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.