Theme-driven Keyphrase Extraction to Analyze Social Media Discourse

William Romano*[1], Omar Sharif*[1], Madhusudan Basak*[1,2], Joseph Gatto[1], Sarah Masud Preum[1,3,4]
[1]Department of Computer Science, Dartmouth College
[2]Department of Computer Science and Engineering, BUET, Bangladesh
[3]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College
[4]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College
{william.j.romano.gr, omar.sharif.gr, madhusudan.basak.gr, joseph.m.gatto.gr, sarah.masud.preum}@dartmouth.edu

Abstract

Social media platforms are vital resources for sharing self-reported health experiences, offering rich data on various health topics. Despite advancements in Natural Language Processing (NLP) enabling large-scale social media data analysis, a gap remains in applying *keyphrase extraction* to health-related content. Keyphrase extraction is used to identify salient concepts in social media discourse without being constrained by predefined entity classes. This paper introduces a theme-driven keyphrase extraction framework tailored for social media, a pioneering approach designed to capture clinically relevant keyphrases from user-generated health texts. Themes are defined as broad categories determined by the objectives of the extraction task. We formulate this novel task of theme-driven keyphrase extraction and demonstrate its potential for efficiently mining social media text for the use case of treatment for opioid use disorder. This paper leverages qualitative and quantitative analysis to demonstrate the feasibility of extracting actionable insights from social media data and efficiently extracting keyphrases using minimally supervised NLP models. Our contributions include the development of a novel data collection and curation framework for theme-driven keyphrase extraction and the creation of *MOUD-Keyphrase*, the first dataset of its kind comprising human-annotated keyphrases from a Reddit community. We also identify the scope of minimally supervised NLP models to extract keyphrases from social media data efficiently. Lastly, we found that a large language model (ChatGPT) outperforms unsupervised keyphrase extraction models, and we evaluate its efficacy in this task.

Introduction

Social media platforms like Twitter, Facebook, and Reddit gather spontaneous, self-reported lived experiences from thousands of individuals with a diverse array of medical and socio-demographic conditions. People seeking or undergoing treatment often resort to social media for informational and emotional support (Chen, Wang et al. 2021). Findings in public health research reports increased reliance on social media and other online health platforms for addressing health information needs (Neely, Eldredge, and Sanders 2021; Bhandari, Shi, and Jung 2014). Thus social media has become an exceptional source of clinically relevant data to understand population-level concerns, knowledge gaps, treatment perceptions, and barriers (Chen, Wang et al. 2021). In addition, social media platforms like Reddit support anonymity and thus encourage rich engagement among peers on stigmatized topics like mental health and substance use recovery (Naslund et al. 2016). Qualitative analysis of social media data has been used to understand various health issues, including COVID-19 (Sleigh et al. 2021), cancer (Levonian et al. 2020), depression, and other mental health conditions (Lachmar et al. 2017).

In parallel, recent advancements in natural language processing have enabled large-scale analysis of social media data, contributing significantly to areas like suicide risk detection, adverse drug reaction detection, and misinformation classification (Mathur, Sawhney, and Shah 2020; Aroyehun and Gelbukh 2019; Dharawat et al. 2022). However, a significant gap remains in applying keyphrase extraction to self-reported health-related content on social media, including online health communities.

Unlike topic modeling and Medical Named Entity Recognition (MedNER), keyphrase extraction concentrates on identifying salient concepts. This unique property of keyphrases makes it a potentially powerful tool for studying social media discourse, as pre-defined entity classes do not constrain it. Consider a social media post:

> Am I the only one taking **subs** that feels **nervus** all of the time? I know **anxiety** is a symptom of other opioid recovery drugs like methadone but I don't take those.

The keyphrases in this post are *subs*, *nervus*, and *anxiety*, where the term *subs* is a shorthand for Suboxone, a medication used to treat opioid use disorder (OUD). Standard MedNER or topic models may struggle to extract these keyphrases if they are not part of other texts in the corpus or are misspelled or presented in shorthand. However, applying keyphrase extraction from social media texts introduces its unique challenges, such as defining what constitutes a *keyphrase*, mitigating annotator biases, and serving end-goal applications.

This paper fills this gap by introducing a novel theme-driven keyphrase extraction framework specifically designed for social media. We present the concept of *themes* as

*Authors contributed equally to the paper.

broad categories determined by the objectives of the extraction task. We aim to identify clinically relevant keyphrases that will help uncover knowledge gaps, treatment perceptions, and experiences. This is particularly useful for researchers exploring social media discourse without being confined by pre-defined entity classes. To our knowledge, no existing models currently facilitate theme-driven, efficient keyphrase analysis of social media texts. We exemplify the value of theme-driven keyphrase extraction by analyzing Reddit posts on Medications for Opioid Use Disorder (MOUD), a topic of great relevance in the US due to the escalating opioid crisis.

Our study focuses on the subreddit r/Suboxone, a popular forum for discussing buprenorphine-based prescription medications. Reddit provides an ideal platform for our study due to its anonymous, publicly available relevant data from thousands of affected individuals considering or undergoing buprenorphine-based treatment for opioid use disorder. We employ a mixed-method approach consisting of (i) qualitative exploration of the annotated data to extract clinically relevant insights and (ii) quantitative analysis to demonstrate the feasibility of extracting keyphrases efficiently using minimally supervised NLP models. The quantitative analysis leverages Unsupervised Keyphrase Extraction and Large Language models. We demonstrate the application of Theme-driven Keyphrase Extraction on an online health community and make the following novel contributions:

1. We propose a unique framework for data collection and curation for theme-driven keyphrase extraction, encompassing a systematic approach for collecting data from social media, designing a theme-specific schema, annotating the data following the schema, and validating the annotation process.
2. Leveraging this framework, we develop a dataset, *MOUD-Keyphrase*, comprising human-annotated keyphrases from a Reddit community. This dataset, the first of its kind, brings a new resource to the field and available for public use[1].
3. We demonstrate the effectiveness of theme-driven keyphrases for mining social media data to extract domain-specific insights. Using a qualitative method, this application showcases theme-driven keyphrase extraction and its utility in a real-world context.
4. We explore the potential of minimal supervision for this task, utilizing ten off-the-shelf Unsupervised Keyphrase Extraction models and ChatGPT. This exploration is a first step towards understanding the role and limitations of minimal supervision in keyphrase extraction tasks.

Related Work

Keyword and Keyphrase Extraction

Keyphrase extraction has been a widely explored research area (Nomoto 2022; Hasan and Ng 2014), with the goal of extracting salient phrases that best summarize a document. Depending on the specific task and how the 'keyness' property is defined, the set of extracted keyphrases can vary significantly (Firoozeh et al. 2020).

Existing keyphrase extraction approaches generally work in two steps. First, they select candidate keyphrases through heuristic rules or manual annotation (Wang, Zhao, and Huang 2016). Second, they apply supervised (Lopez and Romary 2010) or unsupervised (Gu et al. 2021) approaches to rank keyphrases based on their relevance to the document and return the top-k keyphrases. Studies have also been conducted to expand keyword sets to increase the comprehensiveness of keywords relevant to a specific context (Bozarth and Budak 2022). Keyphrase extraction techniques can be characterized as statistical (Campos et al. 2020), graph-based (Mihalcea and Tarau 2004), or embedding-based (Bennani-Smires et al. 2018) methods. Statistical methods often leverage term frequency and document frequency measures, whereas graph-based methods model words or phrases and their co-occurrence relationships in a graph structure. Embedding-based techniques utilize neural embeddings to capture words' semantic and syntactic properties. Each of these methods has its strengths and weaknesses, and their performance can vary depending on the complexity and context of the text.

Recently, contextual embedding-based approaches have been used for keyphrase extraction (Zhang et al. 2022). Large language models like ChatGPT have been explored for keyphrase extraction and generation tasks (Martínez-Cruz, López-López, and Portela 2023). Early results demonstrate that ChatGPT can achieve state-of-the-art performance on keyphrase generation tasks through simple prompting without additional training or fine-tuning (Song et al. 2023).

Discourse Analysis for Opioid Use Disorder (OUD)

The growth of online health communities has provided a space for individuals to share their experiences, provide support, and engage in discussions on topics such as substance use, addiction, and recovery (Chancellor, Mitra, and De Choudhury 2016). Online health discourse on substance use refers to posts or discussions on social media platforms about drugs and other related topics such as addiction, harm reduction, treatment options, and recovery in social media platforms (Lavertu, Hamamsy, and Altman 2021). For example, Chen, Johnny, and Conway (2022) analyzed Reddit discussions about cannabis, alcohol, and opioids to examine the nature of stigma related to these substances. MacLean et al. (2015) found a positive correlation between forum use and recovery by analyzing online discourse. Chancellor et al. (2019) investigated posts from opioid recovery subreddits to uncover potential alternative treatment options for OUD. However, our study differs in the NLP task and scope, exploring a broader range of categories beyond treatment options, including psychophysical effects, medical history, and substance dependency & recovery.

Overall, our study is unique in several significant ways. Unlike previous studies focusing on author-assigned keyphrases in scientific literature (Augenstein et al. 2017), we extract keyphrases from rich, online data, which poses

[1] https://tinyurl.com/ymb4pn6s

unique challenges due to domain-specific vocabulary and colloquial linguistic style, including shorthand, and slang. Additionally, we use thorough manual annotation to identify theme-specific keyphrases instead of Twitter hashtags (Zhang et al. 2016). Hashtags are not always ideal candidates for theme-driven keyphrase extraction due to overreliance on popular hashtags, inherent ambiguity, and limited coverage of relevant information. Finally, we propose a framework for curating keyphrases from social media to enhance the interpretability and applicability of keyphrase extraction. Our analysis of the performance and errors of ChatGPT and other unsupervised models aims to contribute to the ongoing characterization of these models' capabilities. This is important given the increasing prevalence of online communities and the need for sophisticated tools to interpret such data.

Data Collection and Curation Framework for Theme-driven Keyphrase Extraction

We develop a new framework for curating theme-driven keyphrase datasets from social media. As shown in Figure 1, the framework consists of selecting data sources, data collection, theme-specific schema design, data annotation, and validation of the curated dataset. To determine theme-driven keyphrases, we follow the general criteria for 'keyness' provided by (Firoozeh et al. 2020) that include three components 'conformity', 'homogeneity', and 'univocity'. Conformity is reflected by capturing domain-specific terminology, homogeneity includes normalizing the diverse vocabulary, and univocity refers to specific and non-ambiguous keyphrases.

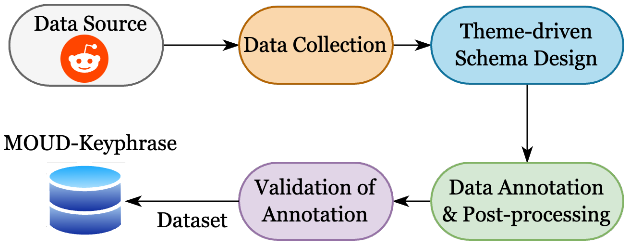

Figure 1: Major steps of theme-driven keyphrase dataset curation framework.

Data Source

Selecting appropriate data sources is paramount for addressing specific problems. For instance, Twitter is optimal for short text analysis, while Reddit offers a rich vein of detailed discussions. Integrating various sources can enhance analysis breadth and depth. Moreover, ethical guidelines and awareness of potential biases in user-generated content collection must be rigorously adhered to.

To develop our dataset, we choose Reddit since it is anonymous and generates a high volume of quality content regarding OUD treatment. As an anonymous platform, the only public information available for a given user is their username, join date, and activity (i.e., post and comment history). The subreddit from which we collected the data for this study is r/Suboxone, a community with over 30 thousand users as of May 2023. Created in 2011, r/Suboxone is a community forum for dialogue between users who share their relationship with the opioid recovery medication *Suboxone*. This subreddit is strictly moderated, and any posts about buying/selling illicit drugs, exposing other users' personal information, or bullying/abuse are removed. Analyzing activity on r/Suboxone makes it possible to achieve a unique and authentic perspective on a user's recovery. While this subreddit explicitly focuses on the treatment of OUD using Suboxone, users frequently discuss their experiences and concerns related to other treatment options too. This is because individuals with OUD may try other OUD treatment options. Discussions extend to various related topics and co-occurring substance and medication usage.

Data Collection

The subsequent step after selecting the data source is collecting the relevant information. It involves determining the timeline from which data should be curated, identifying features that can facilitate data collection (e.g., flairs, hashtags, and user-assigned tags), and deciding on the specific information to be collected, such as posts, comments, and upvotes. We scraped all posts between the 2nd of January, 2018, and the 6th of August, 2022, on the r/Suboxone subreddit using the PRAW and PushShift APIs (Boe 2022; Baumgartner 2022). We observed minimal interaction in this subreddit before 2018 and selected this timeline as it allows us to focus on more recent user data. We accumulated posts, comments, likes, upvotes, and unique post ids. Information that violates ethical concerns is not collected or stored. After filtering the corpus for irrelevant posts (e.g., those containing polls, links with no texts, or posts which had been deleted), we developed a corpus of 15,253 posts. Sample posts are presented in Table 2.

Theme-Driven Schema Design

Determining a phrase to be key to a given text considering various themes is subjective to annotator biases and target downstream applications. It is crucial to design a theme-specific schema to mitigate these biases. According to our exploration, the key steps of the design are: (i) defining the themes with iterative discussion with domain experts, (ii) creating annotation guidelines, and (iii) ensuring the *keyness* of the annotated keyphrases.

In this study, we aim to extract keyphrases relevant to Suboxone-assisted OUD recovery. We collaborate with a team of five domain experts to define the themes for medication-based OUD recovery. All of our collaborators are well-versed in substance use disorder. Together, they cover a wide variety of expertise, including psychiatry, biomedical data science, digital technology for substance use and mental health, epidemiology, public health policy, addiction medicine, and addiction psychiatry. Two of our collaborators are clinicians and administer MOUD treatment. All of them reviewed our samples and helped us to contextualize different aspects of opioid recovery and the shared lived experiences of the affected individuals. Based on guidance from our collaborators, we identify four main themes: *Treatment Options*, *Substance Dependency & Recovery*, *Medical History*, and *Psychophysical Effects*.

The definitions of these themes and the motivation behind choosing them are presented below. Additionally, we have defined another category, *Others*, to include keyphrases not fitting the four target categories but still salient to the original post. We believe this schema and guidelines can extend to other communities with health-related discourse.

- **Treatment Options:** This category covers keyphrases related to different treatment options used for recovery. Such as medications used to treat OUD (e.g., *Buprenorphine*, *Methadone*, or their formulations), psychotherapy, behavioral counseling, or other medications used to cope with withdrawal or other psychophysical effects (e.g., using *melatonin* to help with insomnia while in recovery). We consider prescribed medications, over-the-counter medications, herbal supplements, and other therapeutic options as potential candidates for keyphrases. Keyphrases related to this category can facilitate the analysis of people's perceptions of various treatment options for OUD recovery, including their effectiveness and or lack thereof.
- **Substance Dependency & Recovery:** This category covers keyphrases related to the history of substance use (e.g., *fentanyl*), co-occurring substance use (e.g., *tobacco*, *alcohol*), and critical factors in recovery (e.g., *relapse*). We consider both prescribed and self-administered substances. Analysis of these keyphrases can aid in identifying salient themes relevant to substance usage history, and trajectory of recovery, e.g. tapering, and relapse.
- **Medical History:** Keyphrases concerning medical history, including diagnosis and self-diagnosis of any physical and mental health conditions, relevant medical procedures (e.g., *major surgery*), or critical family medical history. This category covers medical history beyond substance dependency/recovery. Examining medical history-related keyphrases can assist in studying the impact of other health conditions on OUD recovery.
- **Psychophysical Effects:** Keyphrases regarding any physical or psychological effects and symptoms associated with OUD recovery, e.g., psychological effects relevant to withdrawal, precipitated withdrawal, and side effects of medications. We aim to better understand the common or rare (if any) effects of OUD treatment options by examining these keyphrases.
- **Others:** All keyphrases related to OUD but do not belong to any of the above four categories are assigned to this category.

The insights drawn from analyzing theme-specific keyphrases can aid in devising effective intervention strategies for treatment induction, adherence, and retention. In the second step, we thoroughly iterated over our definition of keyphrases via multiple rounds of trial annotations followed by annotator discussions to create the annotation guidelines. Finally, we ensured that keyphrases were annotated to satisfy the general criteria for '*keyness*' inspired by Firoozeh et al. (2020). Conformity, for our purposes here, is reflected by annotating theme-specific terminology. We create homogeneous keyphrases by normalizing slang and misspellings associated with informal discussions online. Finally, to ensure univocity, annotators are instructed to choose non-ambiguous keyphrases.

Data Annotation and Post-Processing

Choosing the optimal strategy to annotate theme-specific keyphrases manually can be difficult (Firoozeh et al. 2020). The most commonly adopted annotation strategy includes annotation by domain experts, hybrid (combination of domain experts and hired annotators), and crowd workers (Chau et al. 2020). We chose the second option as it yields more high-quality annotation while not putting too much burden on domain experts.

We randomly selected 1,000 posts from the collected samples for keyphrase annotation, a set of comparable size to other notable works in keyphrase extraction (Augenstein et al. 2017). Larger social media datasets for keyphrase extraction also exist (Zhang et al. 2016), but those datasets leverage automated approaches like using *Twitter hashtags* as keyphrases. However, our theme-driven definition of keyphrases requires rigorous manual annotation. Thus we limit our sample size for manual annotation to 1,000 posts. The complete data annotation was manually carried out by four graduate students who are both regular social media users and active in NLP research. They studied/used opioid recovery subreddits, so they possessed better domain knowledge than mTurk annotators. Moreover, they were trained on the annotation task and provided background on MOUD and suboxone through multiple sessions led by experts. We used LightTag, an online platform, to label the keyphrases (LightTag 2023). Two annotators annotated each sample to generate high-quality keyphrases. Experts resolved any confusion during annotation through discussion. Example keyphrases from each category of MOUD-Keyphrase are exhibited in Table 1.

Theme	**Frequency**	**Example Keyphrases**
Treatment Options	182	adderall, antidepressant, naloxone
Substance Dependency & Recovery	77	cocaine, fentanyl, heroin
Medical History	35	covid, osteoarthritis, ptsd
Psychophysical Effects	331	aches, constipation, panic attack
Others	256	scam, travel, boyfriend

Table 1: Example keyphrases from each theme. Frequency denotes the number of unique keyphrases after normalization for each theme on the dataset.

Depending on the task, post-processing might also be required to ensure the 'homogeneity' property of the annotated keyphrases. For our task, we have to normalize the extracted

Title	Post	Annotation-1	Annotation-2	JI
Getting on **suboxone**	I just came from Florida and have been **clean** from **dope** for 9 months but the **cravings** are setting in. Can anyone suggest me the best way to go about getting on **suboxone**??	craving, heroin, clean, suboxone	suboxone, craving, heroin	0.75
Tapering from 24mg	Will I go through any **withdrawals** tapering off 24mg of **suboxone** if I taper down to 22mg? Been on 24mg for 4 months.	suboxone, taper, withdrawal	suboxone, taper, withdrawal	1.0

Table 2: Sample excerpts with titles, posts, and annotations. Annotated keyphrases are normalized (e.g., *dope* is normalized to *heroin*). The Jaccard Index (JI) is calculated over the normalized keyphrase list and indicates the similarity between the annotations.

keyphrases manually. Reddit users commonly use various unique phrasings of the same word, e.g., shorthand, slang, and misspellings. For example, the opiate *heroin* may also be referred to as *h*, *dope*, *smack*, and *speedball*, and also it is often misspelled as *herion*, *heroine* etc. We manually map all keyphrase variations to their most meaningful representative parent phrases to solve this problem. Due to space constraints, the steps and guidelines we followed to normalize keyphrases are provided in the supplementary materials[2].

Validation of Annotation

The validation process can include quantitative and qualitative measures to ensure the quality of annotations. Frequently used quantitative measures are inter-annotator agreements, recall, diversity, etc. The qualitative measure depends on the overarching goal and how the keyphrases can be utilized to achieve this. As such, we aim to discover data-driven knowledge from extracted keyphrases, which we demonstrate further through exploratory analysis (RQ1).

Since our annotation involved multiple annotators, we calculated the inter-annotator agreement score to ensure the dataset's quality. We used two lists of normalized keyphrases for each sample from the annotators. We use the Jaccard index to measure the agreement/similarity between annotations (Sarwar, Noor, and Miah 2022). Jaccard index is defined as:

$$JI = \frac{len(A \cap B)}{len(A) + len(B) - len(A \cup B)} \tag{1}$$

Let A and B be the respective set of keyphrases from annotators 1 and 2 for a given sample i in the dataset. JI computes the Jaccard index for sample i and represents the annotator agreement for that sample. While calculating the intersection and union of the two sets, we considered the exact string match between the elements of the sets as used in Schopf, Klimek, and Matthes (2022). We used $Avg.(JI)$ to capture the average Jaccard index for the whole dataset of n samples. The average Jaccard similarity index obtained using the exact string match approach was 61.36%. Given the complexity and subjective judgment of the task, this score indicates a moderate lexical similarity between keyphrases extracted by multiple annotators (Sarwar, Noor, and Miah 2022). Sample posts with extracted keyphrases and JI-score are presented in Table 2.

Methods

In this section, we outline our research questions and describe the methodology we employed to address them, utilizing our meticulously compiled *MOUD-keyphrase* dataset. Employing a mixed-method approach, we initially adopt a qualitative strategy to uncover the clinical insights with our first research question. Subsequently, we proceed with a quantitative analysis to address the second research question, examining the potential of minimally supervised NLP models in efficiently extracting keyphrases.

RQ1: What Clinical Insights Can Theme-Driven Keyphrases From Social Media Provide?

To answer this question, we take a qualitative approach. We look into the list of normalized keyphrases and their frequencies across the dataset for each theme. We investigate frequently-recurring keyphrases as well as those with limited occurrences to gain better insights into the recovery process of OUD. Frequent keyphrases (e.g., *heroin*) offer a ground to study the associated concerns and problems. In contrast, infrequent keyphrases (e.g., *sex drive*) can aid in finding rare, new, and clinically undocumented evidences. Furthermore, we consider the keyphrases that reflect individuals' opinions and emotions toward specific treatment options. We map the keyphrases to the motivating usecases for each theme (referring to the section defining theme-driven schema design) and manually review the associated posts to uncover insightful information about the themes.

We also examine the co-occurrence of keyphrases both within and across different themes. This analysis could lead to discovering previously unknown side effects and new treatment options that can be further explored for hypothesis-driven research.

RQ2: Can We Effectively Extract Keyphrases Using Minimal Supervision?

Qualitative analysis can potentially uncover valuable insights from data, and a comprehensive dataset can aid in grounding these insights. Nonetheless, manual annotation requires a lot of time and effort. Exploring quantitative methods to facilitate data annotation with minimal supervision is crucial. Therefore, we study the scope of unsupervised keyphrase extraction techniques and large language models (e.g., ChatGPT). ChatGPT has demonstrated outstanding performance in NLP tasks across diverse domains (Qin et al. 2023). However, limited research has evaluated ChatGPT's keyphrase extraction ability on theme-specific

[2] https://tinyurl.com/ymb4pn6s

long documents from online health communities. We aim to assess the performance of ChatGPT and ten off-the-shelf unsupervised keyphrase extraction models on our dataset.

Keyphrase extraction using unsupervised off-the-shelf methods: We experimented with three types of methods: statistical, graph-based, and embedding. These techniques represent standard approaches in keyphrase extraction and provide a diverse range of technical implementations for thoroughness. We evaluate the performance of two statistical methods, TfIdf (Manning and Prabhakar 2010), and YAKE (Campos et al. 2020). We explore four graph-based approaches to keyphrase extraction, namely TextRank (Mihalcea and Tarau 2004), TopicRank (Bougouin, Boudin, and Daille 2013), PositionRank (Florescu and Caragea 2017), and MultipartiteRank (Boudin 2018). These classic models are commonly used as baselines in modern unsupervised keyphrase extraction studies (Zhang et al. 2022). We employ the following pipeline for all four embedding methods: 1) Extract candidate keyphrases, 2) Embed each candidate and the full Reddit post, and 3) Return the top-k candidate keyphrases with the highest cosine similarity to the embedding of the entire Reddit post. We evaluate the performance of DistilBERT (Sanh et al. 2019), BERT (Devlin et al. 2018), DeBERTa-v3 (He, Gao, and Chen 2021), and SBERT (Reimers and Gurevych 2019), as they are all widely-used transformer models with proven ability to produce meaningful contextual word embeddings.

Keyphrase extraction using ChatGPT: We use ChatGPT (Ouyang et al. 2022) (OpenAI API model version *gpt-3.5-turbo-0301*) to extract theme-specific keyphrases from the posts. Our experimentation involved both zero-shot and few-shot approaches, using various prompts (Brown et al. 2020). The optimal prompt was determined by a process of trial and error, guided by empirical observations of the model's outputs. We tested several styles of prompts, discarding those that produced outcomes significantly deviating from our expectations and refining those that showed promise. Two different prompt templates, namely *'basic'* and *'guided,'* were utilized to conduct the experiments. In the *'basic'* template, the model prompted to generate OUD-related keyphrases, while in the *'guided'* version, the detailed definitions of various themes were provided in the prompt. The full details and prompt templates are provided with the supplementary materials. The temperature value was set to 0.0 during the experiments, ensuring the deterministic extraction of keyphrases by the model.

Experimental Setup: For the statistical and graph-based methods, we use the PKE library (Boudin 2016) to perform keyphrase extraction. For the embedding methods, we use KeyBERT (Grootendorst 2020) to extract and rank keyphrases. All models use the same candidate keyphrases. Specifically, we use PKE's grammar selection tool to extract only noun phrases as candidates for each post. Finally, we investigate the performance of ChatGPT in zero-shot and few-shot settings and compare it with other models.

Evaluation: We compare the models using various metrics (i.e., precision, recall, F1 score), but the primary focus will be on the F1 score. This metric considers both precision and recall, providing a more comprehensive view of model performance. The ground truth for these metrics is established based on the human-annotated data. We present the F1@k metric for unsupervised keyphrase extraction models, where the value of k represents the top k keyphrases extracted by the model. This approach allows us to effectively compare the quality and relevance of the keyphrases produced by each model. In the evaluation of ChatGPT, we assess precision (P), recall (R), and F1 score, as the model does not produce a ranked list of keyphrases.

By focusing on the effectiveness of different keyphrase extraction models and investigating the conditions under which they excel, we aim to contribute a nuanced understanding of keyphrase extraction from online health communities.

Results

RQ1: What Clinical Insights Can Theme-Driven Keyphrases From Social Media Provide?

After normalization, we obtained a total of 881 unique keyphrases. The distribution of the keyphrases among different groups can be found in Table 1. The findings of our thematic analysis of frequent keyphrases are described in the following.

1. *Treatment options:* Many users posted about treatment options (182 times), indicating that they were concerned about their treatments and thus tried different treatment options. Commonly observed keyphrases from this theme were *suboxone*, *kratom*, *buprenorphine*, etc. An interesting observation is that users discussed *subutex* (84) more than *kratom* (69). But using theme-based analysis, we found *suboxone* co-occurred more with *kratom* (61) than *subutex* (57), potentially indicating *kratom* is a more popular alternative of *suboxone*.
2. *Substance dependency & recovery:* We also found reasonable evidence (77) of keyphrases related to *substance dependency & recovery*. Frequently occurring keyphrases in this category are *taper*, *heroin*, *oxycodone*, and *fentanyl*. From the data, we observed that the majority of posts discussed relapsing to *heroin*.
3. *Medical history:* We found the lowest number (35) of keyphrases in the *medical history* category. The possible reason is that users focused on the ongoing problems or effects and did not mention medical history, or the reported medical histories were not often the keyphrase in a post. Examples of keyphrases from *medical history* theme include but are not limited to *pregnancy*, *adhd*, and *bipolar disorder*.
4. *Psychophysical effects:* We observed that the highest frequency keyphrase category is *psychophysical effects* with 331 in total. The prevalence of *psychophysical effects* can be attributed to the fact that users discuss different types of effects they had faced due to the concurrent treatment or substance use, thus seeking suggestions. *Withdrawal*, *sleep*, and *precipitated withdrawals* were the top mentioned keyphrases from this theme. Furthermore, the

top keyphrases describing the psychophysical effects of withdrawal are *sleep*, *depression*, and *anxiety*.

Our systematic theme-specific analysis of frequent keyphrases is valuable for clinicians and researchers focusing on MOUD treatment. For example, it is advantageous for researchers to understand which other treatment options patients consider while undergoing suboxone-based treatment for opioid recovery. This information can uncover potentially harmful trends in self-prescribed or new treatment options critical to a patient's recovery experience. Similarly, quantifying the prevalence of various psychophysical effects can guide researchers toward emerging, potentially significant side effects which require attention from public health officials. Identifying medical/substance use histories can inform about clinically relevant subpopulations seeking MOUD treatment who frequently engage on Reddit, e.g., individuals with fentanyl or heroin dependency, pregnant women seeking MOUD treatment, or individuals interested in tapering their medications. These findings can inform tailored intervention design for MOUD treatment, i.e., who can be reached through Reddit and when.

What insights can we draw from keyphrase co-occurrences? Our analysis found that at least one instance of the keyphrase *suboxone* co-occurs with each of the known adult side-effects listed on the U.S. Substance Abuse and Mental Health Service Administration webpage[3]. This relationship to standard substance abuse guidelines validates the quality of our annotated data. Additionally, such annotations allow one to gather many self-reports based on suboxone and a corresponding psychophysical effect. This facilitates analysis of the context and severity of various opioid-related issues. For example, one user posted:

> Is anyone getting **heart or chest pain**, **back pain** while **breathing** in sometimes and pain in right upper abdomen randomly while you're **detoxing** yourself with **subs**?...day 5 [and] still finding quite a few of these scary symptoms ... 1st time I was kicking **tranq dope** so idk if it's just that? or the **subs** somehow?

Here we find a user experiencing a physical side-effect of starting suboxone (i.e., lingering withdrawal symptoms) being perceived as a general suboxone side-effect. MOUD researchers are interested in such cases as misperceptions in MOUD treatment can negatively impact treatment induction, adherence, and retention.

Additionally, we find many side effects co-occurring with suboxone which are *not* officially listed as known psychophysical responses to suboxone. For example, almost 2% of user posts in our dataset expressed issues with their sex drive. Other emerging co-occurring psychophysical effects of suboxone include depression, depersonalization, anger, skin crawling, mood swings, hunger, and memory issues. Researchers interested in monitoring rare adverse drug reactions or unknown drug effects may benefit from keyphrase extraction co-occurrence analysis. Accurate extraction of keyphrases in recovery-related social media discourse can help identify similar discussions and inform the design of qualitative, prospective studies to identify population-level perceptions and misperceptions related to MOUD treatment.

Model	F1@5	F1@10	F1@15
Statistical Methods			
TfIdf	0.362	0.352	0.333
YAKE	0.293	0.309	0.303
Graph-Based Methods			
TextRank	0.148	0.224	0.259
TopicRank	0.293	0.3099	0.303
PositionRank	0.265	0.288	0.298
MultipartiteRank	0.301	0.317	0.311
Embedding Methods			
DistilBERT	0.329	0.332	0.319
BERT	0.226	0.284	0.307
DeBERTa-v3	0.207	0.249	0.281
SBERT	0.250	0.319	0.332

Table 3: F1@k for each unsupervised keyphrase extraction model. We varied the value of *k* during testing to evaluate the model's performance using different sets of ranked candidate keyphrases.

In addition to co-occurrence patterns with the keyphrase *suboxone*, there are intriguing patterns with *tapering*. Co-occurrence patterns—such as between *tapering* and *withdrawal* are unsurprising given that dose reduction usually induces withdrawal symptoms. More interesting is the co-occurrence pattern that *tapering* has with both *sleep* and *anxiety*. Such significant associations present in the dataset but not explored by the literature are primary candidates for exploration in future work.

RQ2: Can We Efficiently Extract Keyphrases Using Minimal Supervision?

Our experimental results exhibit that the TfIdf method produces the highest F1-score across all unsupervised baseline experiments presented in Table 3. The method that achieved the second-best performance is SBERT. This is expected as 1) embedding approaches are unique in producing contextual representations of the whole post, which explicitly encode textual semantics—a valuable feature for unsupervised keyphrase extraction; 2) SBERT is the only embedding method employed that was trained to output embeddings for longer sequences of texts (other methods are only explicitly trained to produce only high-quality token embeddings). Unfortunately, none of the standard unsupervised models achieved F1 above 0.36. This limits the use of off-the-shelf models to extract theme-driven keyphrases.

This motivates us to explore the capabilities of large language models (e.g., ChatGPT) for the keyphrase extraction task. The performance of ChatGPT with both zero-short and few-shot approaches is presented in table 4. From the results, it is evident that the few-shot examples improve the model's performance. In both few-shot experiments, we ran-

[3] https://tinyurl.com/ms6bcrcj

Model	**P**	**R**	**F1**
TfIdf	0.238	0.555	0.333
MultipartiteRank	0.222	0.519	0.311
SBERT	0.237	0.557	0.332
Zero-shot (Basic)	0.180	0.313	0.229
Few-shot (Basic)	**0.478**	0.554	**0.510**
Zero-shot (Guided)	0.162	0.597	0.256
Few-shot (Guided)	0.173	**0.622**	0.271

Table 4: ChatGPT performance comparison on Zero-shot and Few-shot settings with best unsupervised keyphrase extraction models. Since ChatGPT does not provide a ranked set of keyphrases, we focus on precision (P), recall (R), and F1 scores for evaluation purposes. Experiments labeled *basic* were only prompted for basic keyphrase extraction without additional context. Those labeled *guided* utilized theme-specific guidelines in their prompts.

domly sampled three examples from the dataset. The average score of 5 runs with the same temperature is reported to mitigate potential bias to specific samples. The basic prompt with the few-shot examples acquired the highest F1 score of 0.51 and outdid all the other models. The guided few-shot model score is surprisingly low, only 0.271. This could be due to the ChatGPT model being too sensitive to their input prompts. Furthermore, the success of prompting heavily relies on the familiarity of the label space (Min et al. 2022). Regarding keyphrase extraction, the label space is vast and challenging to cover with few-shot samples. As a consequence, this might lead to suboptimal performance.

Although the basic prompt with the few-shot performs better than the unsupervised models, the overall performance of this model with an F1-score of 0.51 is poor. To get more insights, we further analyze the outputs generated by this setting of ChatGPT. We found that it suffers from issues like missing important keyphrases, focusing on irrelevant keyphrases, showing weaknesses in filtering out specific keyphrases, and performing poorly in determining the occurrences of the keyphrases for all the themes.

- **ChatGPT often misses important keyphrases:** Being a generalized extractor, ChatGPT often misses identifying keyphrases that carry useful information about the specific context. For example, in the following example, *clean* and *pregnancy* carry valuable information about the context of the post, but those have been missed by ChatGPT.

  > ... Like probably many people on this sub and on suboxone in general I have had a long struggle with my addiction. I have been on Suboxone for 3 years and I think that the last three years have been the most stable years that I've had with my addiction since it started (other then during my **pregnancy**, which was before the Suboxone but I am proud to say that I was **clean** the whole time)...

- **ChatGPT often overpredicts keyphrases**: ChatGPT often over-predicts keyphrases. For each sample, this model extracts on average 6.25 keyphrases (with a standard deviation of 1.11) while the average number of keyphrases in our ground truth is 4.87. Following is an example where along with some relevant keyphrases (*upset stomach*, *withdrawals*, *achey*), ChatGPT extracts a lot of irrelevant keyphrases (*oatmeal*, *yougurt*, *protein bars*, *vitamins*) which can confuse the future models trained on the data and lead to an incorrect training.

  > . . . So day 8, the hardest is still adjusting and dealing with insomnia, as well as off and on **upset stomach**. Can't eat meat yet, lots of **oatmeal**, bananas, **yougurt**, **protein bars** and **vitamins**. I went 4 nights of 2-3hrs of sleep. It's a response of fight or flight in early **withdrawals**, but today I felt a bit better. I do get pretty **achey** and exhausted doing too much. But I can keep up with some chores. Force myself outside for 15min. . .

- **ChatGPT performs poorly for some keyphrases**: Along with the qualitative observations, we tried to quantitatively identify the blindspots of the ChatGPT for theme-driven keyphrase identification. We aimed to determine the keyphrases ChatGPT has a minimal success ratio. We considered the keyphrases which occurred in more than 3 sample texts. For all the runs, we found ChatGPT failed to identify a high number of keyphrases more than half of the time, e.g., *fear*, *crap*, *hospital*, *cold turkey*, and *doctor*.

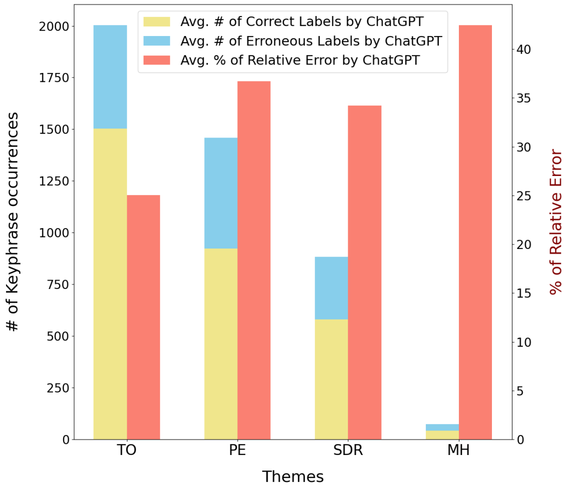

Figure 2: The performance of ChatGPT Few-shot (Basic) model in extracting keyphrase occurrence for different themes (TO: Treatment Options, PE: Psychophysical Effects, SDR: Substance Dependency & Recovery, and MH: Medical History). For each theme, the first bar presents the number of keyphrase occurrences the model failed to identify stacked on the number of keyphrase occurrences correctly identified by it, and the second bar shows the relative error (in percentage) made by the model.

- **ChatGPT struggles to determine keyphrase occurrences for all themes:** We were curious to see whether ChatGPT over-predicts or under-predicts any particular themes in our used dataset. Dividing the ground-truth keyphrases of the whole dataset into different themes, we determined the number and percentage of keyphrase occurrences missed by ChatGPT, as shown in Figure 2. ChatGPT faces more or less difficulties in extracting keyphrase occurrences across all themes. In particular, the model fails to identify a significant percentage of occurrences under the themes of 'medical history' (53.15%) and 'psychophysical effects' (41.91%), highlighting areas for improvement.

Ethical Considerations

This research was conducted under Institutional Review Board (IRB) approval at the authors' institution. Since Reddit users can create an account using only an email address, which Reddit does not disclose, the users are largely anonymous. These users can only be identified by the data they voluntarily disclose in their posts. Even if a user shares self-identifying information (e.g., location, age, gender identity), it remains nearly impossible to ascertain the true identity of a Reddit user without further details. This inherent anonymity provides an additional layer of privacy protection.

Regarding the use of Reddit data, the user agreement requires users to consent to share their public posts and comments via the Reddit API, allowing for their utilization in research like ours under specific conditions that respect user privacy and ethical research guidelines. Additionally, as a standard precaution, posts included as examples in the paper were paraphrased to prevent the reader from directly identifying the Reddit user account that posted each example.

Bias and Fairness Bias and fairness, especially concerning large models, are increasingly gaining traction in machine learning research (Gehman et al. 2020). This study bears the potential for bias from three primary sources: research question bias, data bias, and model bias. Research question bias might inadvertently favor conventional clinical relevance as a group of clinical researchers suggested the themes. Data bias can emerge if the training data poorly represents the intended population, for example, data collection focuses only on specific demographics, location, or time that was not intended or part of the study design. With Reddit data typically skewing toward young white males, it could limit its relevance to other demographics[4]. Since Reddit data is anonymous, identifying potential bias towards a specific subpopulation is challenging. Model bias could arise if the comparison model is already biased. The absence of control or transparency over the large text corpus used for training some of the large language models could introduce undetected biases, an ongoing issue with large models (Dodge et al. 2021).

Due to the unavailability of demographic data from Reddit, a complete fairness analysis is currently unfeasible and will be addressed in future research. In response to potential biases, we applied several strategies, including selecting multiple themes, randomly sampling from the more extensive data collection, and diversifying data temporally to minimize time-related bias. Despite these efforts, achieving absolute bias-free research is elusive. We are dedicated to ongoing bias monitoring and mitigation in our studies, hoping that our work aids in detecting and rectifying potential biases in computing research.

Broader Impact

Implications for keyphrase extraction on Social Media: Existing work on keyphrase extraction in social media has two limitations: (1) Posts are often derived from Twitter, where low character limits preclude capturing rich information from long, complex self-narrated text from people with lived experiences. Analysis of our data requires understanding texts of variable length, some of which reach even 10,000 characters. Thus it will promote the creation of better HealthNLP models capable of modeling keyphrases in more extended social media discourse on health. (2) Large Twitter datasets utilize hashtags as surrogate keywords—a strategy based on the error-prone assumption that hashtags are always indicators of keyness or saliency. This is the first dataset with keyphrases extracted by human annotators. This work thus provides reliable annotation for clinically relevant keyphrase extraction from Reddit on MOUD-based treatment for opioid recovery.

Implications for MOUD Research: Our *MOUD-Keyphrase* dataset can inform the development of clinician-facing tools that facilitate the discovery of tangible insights that can inform MOUD research and practice. For example, a keyphrase extraction tool based on our dataset could facilitate the discovery of the perceived effectiveness of different MOUD treatment options, strategies to cope with side effects, rare/new adverse drug reactions, and uncover patterns in the patient-reported experience with different MOUDs. Such findings may guide future research in opioid recovery by clinicians and public health researchers. Also, results from such theme-driven keyphrase extraction may guide the development of tailored patient communication tools and programs, e.g., when and how to taper or potentially severe side effects of suboxone.

Limitations and Future Work

The limitations of this study primarily stem from the inadequate supply of labeled data, which restricts the application of supervised models and might affect the generalizability of our findings. Another limitation is the focus on Reddit posts, potentially missing nuances of health-related discussions on other social media platforms or online health communities. Furthermore, the unsupervised models explored for keyphrase extraction might struggle with Reddit's informal language and prevalent health-related terminology.

Future work could explore the scope of the study to include various social media platforms or online communities to capture a broader spectrum of online health discussions. Annotating more data to enable the deployment of supervised models could also enhance the model's performance.

[4] https://tinyurl.com/3n2vwe9x, https://tinyurl.com/5n8vxzak

Additionally, developing or adapting keyphrase extraction models to better handle the unique characteristics of social media discourse is a necessary avenue for future research.

Our study focused on four themes, suggesting that other themes and subthemes in online discourse may have been overlooked. The inconsistent inflections or word choices of extracted keyphrases present another area of improvement. Future studies could leverage large language models for keyphrase generation, and conduct user studies to evaluate the perceived usefulness of extracted keyphrases, thereby enhancing both the quality and utility of the analysis.

References

Aroyehun, S. T.; and Gelbukh, A. 2019. Detection of Adverse Drug Reaction in Tweets Using a Combination of Heterogeneous Word Embeddings. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, 133–135. Florence, Italy: Association for Computational Linguistics.

Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; and McCallum, A. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555. Association for Computational Linguistics.

Baumgartner, J. M. 2022. Pushshift API.

Bennani-Smires, K.; Musat, C.; Hossmann, A.; Baeriswyl, M.; and Jaggi, M. 2018. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 221–229. Brussels, Belgium: Association for Computational Linguistics.

Bhandari, N.; Shi, Y.; and Jung, K. 2014. Seeking health information online: does limited healthcare access matter? *J. Am. Med. Inform. Assoc.*, 21(6): 1113–1117.

Boe, B. 2022. PRAW: The Python Reddit API Wrapper.

Boudin, F. 2016. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 69–73. Osaka, Japan.

Boudin, F. 2018. Unsupervised Keyphrase Extraction with Multipartite Graphs. *CoRR*, abs/1803.08721.

Bougouin, A.; Boudin, F.; and Daille, B. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 543–551. Nagoya, Japan: Asian Federation of Natural Language Processing.

Bozarth, L.; and Budak, C. 2022. Keyword expansion techniques for mining social movement data on social media. *EPJ Data Science*, 11(1).

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

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.

Chancellor, S.; Mitra, T.; and De Choudhury, M. 2016. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, 2111–2123. New York, NY, USA: Association for Computing Machinery. ISBN 9781450333627.

Chancellor, S.; Nitzburg, G.; Hu, A.; Zampieri, F.; and De Choudhury, M. 2019. Discovering Alternative Treatments for Opioid Use Recovery Using Social Media. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, 1–15. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Chau, H.; Balaneshin, S.; Liu, K.; and Linda, O. 2020. Understanding the Tradeoff between Cost and Quality of Expert Annotations for Keyphrase Extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, 74–86. Barcelona, Spain: Association for Computational Linguistics.

Chen, A. T.; Johnny, S.; and Conway, M. 2022. Examining stigma relating to substance use and contextual factors in social media discussions. *Drug and Alcohol Dependence Reports*, 3: 100061.

Chen, J.; Wang, Y.; et al. 2021. Social media use for health purposes: systematic review. *Journal of medical Internet research*, 23(5): e17917.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Dharawat, A.; Lourentzou, I.; Morales, A.; and Zhai, C. 2022. Drink Bleach or Do What Now? Covid-HeRA: A Study of Risk-Informed Health Decision Making in the Presence of COVID-19 Misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1218–1227.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Firoozeh, N.; Nazarenko, A.; Alizon, F.; and Daille, B. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3): 259–291.

Florescu, C.; and Caragea, C. 2017. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), 1105–1115. Vancouver, Canada: Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; au2, H. D. I.; and Crawford, K. 2021. Datasheets for Datasets. arXiv:1803.09010.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Online: Association for Computational Linguistics.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Gu, X.; Wang, Z.; Bi, Z.; Meng, Y.; Liu, L.; Han, J.; and Shang, J. 2021. UCPhrase: Unsupervised Context-Aware Quality Phrase Tagging. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 478–486. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.

Hasan, K. S.; and Ng, V. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1262–1273. Baltimore, Maryland: Association for Computational Linguistics.

He, P.; Gao, J.; and Chen, W. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *CoRR*, abs/2111.09543.

Lachmar, E. M.; Wittenborn, A. K.; Bogen, K. W.; and McCauley, H. L. 2017. #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR Ment Health*, 4(4): e43.

Lavertu, A.; Hamamsy, T.; and Altman, R. B. 2021. Monitoring the opioid epidemic via social media discussions. *medRxiv*.

Levonian, Z.; Erikson, D. R.; Luo, W.; Narayanan, S.; Rubya, S.; Vachher, P.; Terveen, L.; and Yarosh, S. 2020. Bridging Qualitative and Quantitative Methods for User Modeling: Tracing Cancer Patient Behavior in an Online Health Community. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 405–416.

LightTag. 2023. The text annotation tool for teams. https://www.lighttag.io/. Accessed: 2023-01-15.

Lopez, P.; and Romary, L. 2010. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 248–251. Uppsala, Sweden: Association for Computational Linguistics.

MacLean, D.; Gupta, S.; Lembke, A.; Manning, C.; and Heer, J. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, 1511–1526. Association for Computing Machinery. ISBN 9781450329224.

Manning, C.; and Prabhakar, R. 2010. Introduction to information retrieval.

Martínez-Cruz, R.; López-López, A. J.; and Portela, J. 2023. ChatGPT vs State-of-the-Art Models: A Benchmarking Study in Keyphrase Generation Task. arXiv:2304.14177.

Mathur, P.; Sawhney, R.; and Shah, R. R. 2020. Suicide Risk Assessment via Temporal Psycholinguistic Modeling (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10): 13873–13874.

Mihalcea, R.; and Tarau, P. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Barcelona, Spain.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Naslund, J. A.; Aschbrenner, K. A.; Marsch, L. A.; and Bartels, S. J. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences*, 25(2): 113–122.

Neely, S.; Eldredge, C.; and Sanders, R. 2021. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: Survey study. *J. Med. Internet Res.*, 23(6): e29802.

Nomoto, T. 2022. Keyword Extraction: A Modern Perspective. *Sn Computer Science*, 4.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv:2302.06476.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sarwar, T. B.; Noor, N. K. M.; and Miah, M. S. U. 2022. Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. *PeerJ Computer Science*, 8.

Schopf, T.; Klimek, S.; and Matthes, F. 2022. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Sleigh, J.; Amann, J.; Schneider, M.; and Vayena, E. 2021. Qualitative analysis of visual risk communication on twitter during the Covid-19 pandemic. *BMC Public Health*, 21(1): 810.

Song, M.; Jiang, H.; Shi, S.; Yao, S.; Lu, S.; Feng, Y.; Liu, H.; and Jing, L. 2023. Is ChatGPT A Good Keyphrase Generator? A Preliminary Study. arXiv:2303.13001.

Wang, M.; Zhao, B.; and Huang, Y. 2016. PTR: Phrase-Based Topical Ranking for Automatic Keyphrase Extraction in Scientific Publications. In Hirose, A.; Ozawa, S.; Doya, K.; Ikeda, K.; Lee, M.; and Liu, D., eds., *Neural Information Processing*, 120–128. Cham: Springer International Publishing. ISBN 978-3-319-46681-1.

Zhang, L.; Chen, Q.; Wang, W.; Deng, C.; Zhang, S.; Li, B.; Wang, W.; and Cao, X. 2022. MDERank: A Masked Document Embedding Rank Approach for Unsupervised Keyphrase Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, 396–409. Dublin, Ireland: Association for Computational Linguistics.

Zhang, Q.; Wang, Y.; Gong, Y.; and Huang, X.-J. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 836–845.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost“ of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? N/A
   (b) Did you mention the license of the assets? N/A
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes

6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? N/A
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A
   (d) Did you discuss how data is stored, shared, and unidentified? Yes