Age-Specific Linguistic Features of Depression via Social Media

Charlotte Rosario
The Nueva School
San Mateo, United States
charlottearosario@gmail.com

Abstract

Social media data has become a crucial resource for understanding and detecting mental health challenges. However, there is a significant gap in our understanding of age-specific linguistic markers associated with classifying depression. This study bridges the gap by analyzing 25,241 text samples from 15,156 Reddit users with self-reported depression across two age groups: adolescents (13-20 year olds) and adults (21+). Through a quantitative exploratory analysis using LIWC, topic modeling, and data visualization, distinct patterns and topical differences emerged in the language of depression for adolescents and adults, including social concerns, temporal focuses, emotions, and cognition. These findings enhance our understanding of how depression is expressed on social media, bearing implications for accurate classification and tailored interventions across different age groups.

1 Introduction

Depression, a prevalent mental health condition that impacts over 300 million individuals worldwide (World Health Organization, 2017), has historically been approached with a generalization that it impacts all individuals in the same way. However, in recent years, there has been a growing recognition of depression's impact on specific subgroups, including teens and young children (Ahrens, 2002). Recent research has found that depression manifests differently for individuals at different stages of life (Rice et al., 2019; Kaufman et al., 2001). This shift in perspective underscores the importance of exploring the causes and correlates of emotional distress specific to different age groups.

It is crucial to observe and understand the unique linguistic markers of depression in different age groups, as indicators of depression for adults may not hold true when applied to adolescent populations, and vice versa. This "one-size-fits-all" approach can lead to misdiagnosis, inappropriate interventions, and limited access to suitable support services, thus jeopardizing the well-being of individuals (Spector, 1991; Taylor and Brown, 1988). Given the absence of a comprehensive global surveillance system and the rapid evolution of global health trends, there is an imperative need to gather up-to-date insights on the impact of mental health disorders of different subpopulations (Ormel et al., 1994; Patel et al., 1999).

Recently, there has been a surge of interest in leveraging social media data to gain valuable psychological insights (Karim et al., 2020; Homan et al., 2014). Analyzing social media data has been proven to complement traditional mental health assessments, including the Patient Health Questionnaire-9 (Liu et al., 2022) and the Behavioral Risk Factor Surveillance System (BFRSS) survey (Eichstaedt et al., 2015; Culotta, 2014). Due to the growing popularity of social media platforms for receiving online peer support among those facing mental health challenges, social media data has demonstrated potential for reflecting greater sociological trends.

Most social media investigations of mental health have either overlooked age as a feature or focused solely on the platforms' predominant young adult populations (Low et al., 2020; Coppersmith et al., 2014). However, it is important to recognize that age-based expression of different subgroups may significantly differ; per the principles of postcolonial computing, analytical insight and intervention strategies cannot be simply transplanted from one subgroup to another without modification (Irani et al., 2010). Each age group has its unique context and characteristics that must be considered to ensure the relevance and effectiveness of intervention strategies.

This study quantitatively examines language use in individuals sharing their experiences with depres-

Proceedings of the Student Research Workshop associated with RANLP-2023,
pages 33–43, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_004

sion on social media, specifically on Reddit. Two age groups—adolescents (below 21) and adults (above 21)—are analyzed to explore two primary questions: (1) How do these groups differ in expressing their depression on social media? and (2) How do they differ compared to neurotypical individuals of the same age group? The contributions of this paper are twofold: first, prior research on the existence of language markers that distinguish depression on social media is affirmed through comparative analyses with control users; second, linguistic insights into age-specific markers of depression are revealed using closed- and open-vocabulary approaches. The findings provide a nuanced understanding of how depression manifests on online social media platforms, laying the foundation for future hypothesis-driven research on mental health diagnosis and treatment strategies tailored to different age groups.

2 Related Works

2.1 Language for Psychological Assessment

Language has been extensively used in previous research to identify linguistic patterns associated with psychological traits, such as emotional stability and personality (Eichstaedt et al., 2015). The profound influence of language, particularly one's native language, on thoughts, actions, and social relationships is widely acknowledged (Maynard and Peräkylä, 2006). Studies, including the work of Boroditsky et al. (2003), have demonstrated the intricate relationship between language perception and its impact on social processes. Linguistic styles encompass various indicators of lexical density, temporal references, social support and connectivity, and environmental awareness. Previous research has emphasized the significance of these linguistic cues in comprehending mental health, both in everyday and social media contexts (Ramírez-Esparza et al., 2008).

2.2 Social Media and Mental Health

Numerous studies have turned to online social media data, particularly in terms of language and conversational patterns, as a valuable source for understanding and detecting global health trends. This line of inquiry encompasses various areas, such as utilizing social media to gain insights into diseases (Paul and Dredze, 2011), substance abuse (MacLean et al., 2015; Murnane and Counts, 2014), postpartum depression (Choudhury et al., 2014), eating disorders (Chancellor et al., 2016), and other mental health conditions (Coppersmith et al., 2014; Liu et al., 2022; Tsugawa et al., 2015).

Social media language has revealed several distinctive ways individuals with depression express themselves. For instance, Coppersmith et al. (2014) discovered that depressed users tend to exhibit higher levels of self-focus, anxiety, and anger in their online writings. Social media users with depression often exhibit various symptoms and behaviors, including anhedonia, social difficulties, health and sleep problems, inactivity, thoughts of death, perceived hopelessness, tentativeness, overall negativity, sadness, interpersonal hostility, and disinterest in self-care and leisure (Schwartz et al., 2014; Preoţiuc-Pietro et al., 2015; Resnik et al., 2015; Chancellor et al., 2016; Choudhury et al., 2013). They also use first-person singular pronouns and swear more frequently than neurotypical users (Chung and Pennebaker, 2007).

These investigations have revealed compelling evidence that individuals with depression display unique linguistic patterns in their online communication that distinguish them from the broader population. However, it is important to recognize that the majority of existing research has generalized these findings across various social groups. This prompts the question of whether linguistic markers of depression remain consistent when examined within specific demographic cohorts.

2.3 Demographic Language Markers of Depression

Previous research has sought to investigate nuances in the language markers of depression specific to different demographic subgroups. Choudhury et al. (2016) identified differences in experiences of depression between users of different genders and countries of origin on Twitter. Loveys et al. (2018) explored cultural differences in online language data regarding depression of users from different racial and ethnic backgrounds. Ramírez-Esparza et al. (2008) found that Spanish-speaking depressed individuals were more likely to mention relational concerns than English-speaking individuals. Mittal et al. (2023) discovered differential mental health language markers around race, politics, violence, employment, and affordability for immigrant populations. While these studies highlight the variation in linguistic features of depression across different gender, cultural, racial, and political subgroups,

Regex Patterns
(i) I (was | am) born in <four digit year>
(ii) I (was | am) born in <two digit year>
(iii) I am <age>(years old | yrs old | yo)
(iiii) I am a <age>(year old | yo)
(v) I am <age>(<punctuation>| <conjunction>)

Table 1: Regular expression patterns for identifying age disclosures in Reddit posts.

there is a gap in research regarding the language features of depression that vary among different age groups.

The present study presents a data-driven exploratory analysis of the social media language of diverse age profiles who express their symptoms of depression online.

3 Data

Reddit, an online discussion-based social media platform, was selected as the data source for this study. The platform encourages self-disclosure by enforcing anonymity of users, which can contribute to obtaining less biased results (Gaur et al., 2018). An initial collection of 512,876 Reddit posts was collected from the r/depression, r/suicidewatch, and r/mentalhealth subreddits. This included the 252,459 posts extracted from the r/depression, r/mentalhealth, and r/SuicideWatch subreddits of the Reddit Mental Health Dataset proposed by Low et al. (2020) and 7,000 r/depression Reddit posts from the dataset of Pirina and Çagri Çöltekin (2018). The rest of the posts were obtained directly from Reddit using the PRAW API.

3.1 Annotating for age

Demographic information of users on Reddit is unavailable due to the platform's policy of anonymity, which introduces the need for inferring the age of users. Prior research in this field has sought to identify the best automated method of annotating age in social media data, with several studies using machine learning or lexicon-based approaches (Yatam and Reddy, 2014; Chen et al., 2015). These methods, however, were not used in the present study to avoid adding additional uncertainty or bias into the dataset. Since Reddit users often explicitly disclose their age in their post, stating "I'm 15 years old" or their age in brackets (e.g., "I [17f] just broke up with bf [18m]"), age was inferred based on natural language patterns. Using a similar approach as Tigunova et al. (2020), user's age was extracted using five variations of regular expressions patterns, as listed in Table 1, with pattern (v) specifically designed to avoid false positives (e.g., "I am 60 miles away" or "I feel like I am 60"). All users younger than 13 were excluded, as this violates Reddit's terms of service.

3.2 Validation of age inference

Following a similar approach as Chew et al. (2021), the age inference methods were validated based on annotations obtained from two independent raters on a sample of 100 users. Agreement was found between the raters' annotations and the one given by our method with a total accuracy of 97%.

3.3 Filtering for genuine mental health disclosure

Posts were filtered for explicit key phrases indicating depression based on the method of Choudhury et al. (2016), which was sourced from consultation with a psychiatrist. This was necessary given that the Reddit corpus is susceptible to significant noise from sarcastic, humorous, or flippant usage (e.g., "i have to do the laundry, kill me now" is not indicative of genuine depression). Additionally, regular expression patterns were used to ensure content and age disclosures were representative of actual individuals with depression (e.g., avoiding "my best friend is depressed right now").

After filtering for genuine age-disclosure and mental health expression, a total of 25,241 posts with 15,156 unique Reddit users were extracted. This dataset will be referred to in the present study as MID (i.e. mental illness disclosure).

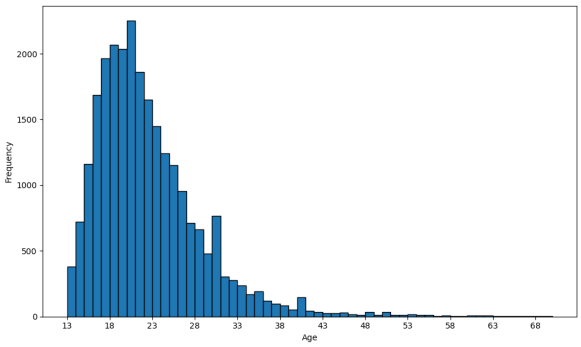

Figure 1: Distribution of ages of Reddit users. (Ages 70+ not shown due to negligible frequencies.)

3.4 Categorizing into age groups

Users were then categorized into age groups. Due to binary age groups (under/over 20 years old) resulting in more distinct results in previous research (Cesare et al., 2017) and the unique distribution of users' ages in the MID dataset, as seen in Figure 1, users were ultimately split into two groups, adolescent (13-20) and adult (21+).

3.5 Gathering a control dataset

To enable robust statistical comparisons between depressed and neurotypical users of different ages, we also obtained a control data sample from non-mental health-related subreddits. This data was taken from the Reddit Mental Health Dataset (Low et al., 2020) and included 11 non-mental health subreddits, resulting in a total of 302,172 posts initially. To ensure data integrity, posts with age-disclosure that were not present in the MID dataset and did not match any key phrases listed in Table 1 were extracted. In total, the control dataset (referred to as CTL) yielded 30,489 posts from 17,092 unique authors, with 10,327 adolescent and 20,162 adult users.

4 Methods

Reddit posts were characterized using two sets of language features: (a) dictionary-based psycholinguistic features, and (b) open-vocabulary topics.

4.1 Closed vocabulary: LIWC Analysis

An established language dictionary known as Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022) was employed as a closed-vocabulary approach to analyze the language patterns. This top-down approach has been extensively used in language research to detect emotions and personality (Eichstaedt et al., 2020). Three categories of linguistic measures derived from the LIWC2022 software were analyzed: (1) affective attributes, (2) cognitive attributes, and (3) linguistic style attributes. As seen in Table 2, the **affective measures** include negativity, anger, anxiety, sadness, and swearing; the two **cognitive measures** include Cognition (insight, tentativeness, and differentiation) and Perception (spacial, visual, auditory, time, past-focused, present-focused, and future-oriented); and the three **linguistic measures** include Lexical Density (articles, prepositions, verbs, auxiliary verbs, nouns, and adverbs), Social/Personal Concerns (words belonging to family, friends, death, work, money, substances, health, wellness, and sexual), and Interpersonal Focus (personal or impersonal pronouns). The relative frequency of each LIWC measure was calculated by taking the mean of each measure within each of the two age groups. Since multiple features are examined concurrently (31 LIWC categories), coefficients are deemed statistically significant if their values fall below a Benjamini-Hochberg-corrected two-tailed p-value of 0.05.

	μ (adolescent)	*μ (adult)*	*effect size*	*p*
Affective attributes				
Negativity	2.113	2.005	0.073	***
Anger	0.283	0.248	0.069	***
Anxiety	0.460	0.431	0.042	**
Sadness	0.691	0.646	0.053	***
Swearing	0.476	0.382	0.118	***
Cognitive attributes				
Cognition				
Insight	3.502	3.335	0.095	***
Tentative	3.129	2.957	0.093	***
Differentiation	4.226	4.134	0.054	***
Perception				***
Visual	0.531	0.573	-0.053	***
Auditory	0.180	0.148	0.078	***
Time	5.572	5.765	-0.087	***
Focus Past	5.423	4.802	-0.106	***
Focus Present	7.439	7.144	0.115	***
Focus Future	1.576	1.496	0.059	***
Linguistic style attributes				
Lexical density				
Articles	4.022	4.369	-0.219	***
Prepositions	12.600	12.920	-0.127	***
Auxiliary Verbs	11.705	11.466	0.090	***
Adverbs	7.838	7.488	0.148	***
Negations	3.168	3.054	0.073	***
Verbs	21.893	21.415	0.142	***
Social/Personal Concerns				
Family	0.708	0.659	0.050	***
Friend	0.541	0.478	0.083	***
Work	1.425	1.594	-0.113	***
Money	0.225	0.402	-0.297	***
Health	1.803	1.870	-0.038	**
Wellness	0.037	0.062	-0.117	***
Substances	0.072	0.085	-0.042	***
Sexual	0.117	0.131	-0.032	*
Death	0.632	0.535	0.117	***
Interpersonal Focus				
Personal Pronouns	15.630	14.791	0.271	***
Impersonal Pronouns	5.534	5.292	0.120	***

Table 2: Differences between posts from youth and adult MID users based on linguistic measures. All effect sizes (measured as Cohen's D) between are significant at p <.05, two-tailed t-test, Benjamini-Hochberg corrected.

4.2 Open vocabulary: LDA Topic Modeling

The second method employed an open-vocabulary approach utilizing Latent Dirichlet Allocation topic modeling (LDA) (Blei et al., 2003) to generate data-driven linguistic features known as topics. LDA identifies common topics present in text documents by capturing sets of words (e.g., 'september', 'october', 'november') that frequently co-occur, allowing for manual inspection and assessment of recurring themes. Although topic modeling may not capture the same level of detailed insights as human observation due to its unsupervised nature, it offers the potential to discover patterns in users' concerns that may otherwise go unnoticed (Low et al., 2020).

Data was pre-processed by removing high-frequency words (occurring in more than 75% of the documents), words that occur fewer than five times, URLs, @mentions, #hashtags, emoticons, emojis, numbers, punctuation marks, and special characters. Reddit posts were then tokenized using *happierfuntokenizing* from the DLATK Python library (Schwartz et al., 2017). We employed Gensim's multi-core LDA implementation with default hyper-parameter settings and 10 topics, determined through the coherence score and the average corpus likelihood value over ten runs. The words associated with each topic generated by the LDA model were inspected by two researchers familiar with mental health content to extract descriptive topical themes.

5 Results

Adolescent and adult MID users show considerable differences in the linguistic features and topics discussed in their Reddit posts.

5.1 Affective Attributes

As shown in Table 2, adolescent MID users received higher LIWC scores across all affect measures than the adult MID subgroup, including higher negativity (effect size = 0.073; 5.4% higher), anger (effect size = 0.069; 14.0% higher), anxiety (effect size = 0.042; 6.7% higher), and sadness (effect size = 0.053; 7.08% higher). While all four affect measures had small effect sizes, the frequency of swearing (effect size = 0.118) for adolescents was significantly greater than adults by 24.57%. This mean relative difference in the frequency of swearing was also seen within the adolescent age group between MID and CTL users, where adolescents with depression swore 260.2% more than those of the same age without depression.

The extent to which the adolescent and adult CTL users differ along the aggregate of these affective attributes is also reported. Per Figure 2, this mean difference separating the two age groups in the control cohort is only 1.13%, which is lower in comparison to 2.46% in the case of the MID cohort. This indicates that the adolescents and adults in the MID cohort show differences beyond that accounted for in the control sample.

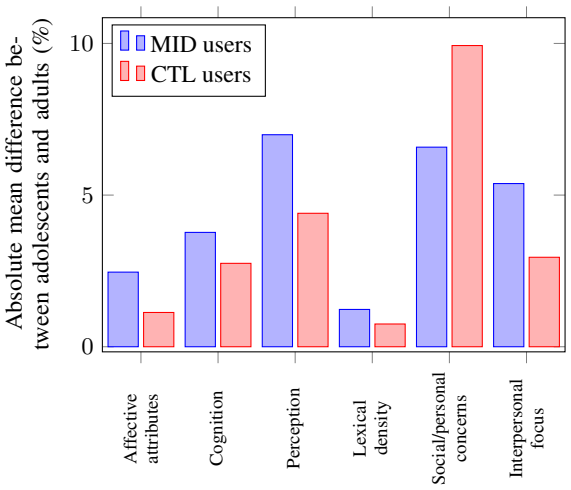

Figure 2: Mean absolute differences between adolescent and adult users with depression (MID) versus control (CTL) users per the various categories of linguistic measures. Difference for a specific measure is calculated as the ratio of the difference between the values of the measure for adolescents and adults, to the value of the measure among adults.

5.2 Cognitive Attributes

Adult MID users showed lower cognition (i.e. greater cognitive impairment) in their Reddit posts compared to adolescent MID users. For instance, insight (effect size = 0.095) is higher in adolescents relative to adults by 5.01%. Similarly, adolescents showed greater tentativeness (effect size = 0.093; 5.81% higher). Overall, the adolescent and adult users differ by 3.77% in the MID cohort, whereas this difference is significantly smaller (2.25%) in the CTL cohort across the various cognitive attributes, per Figure 2, indicating that these results are significant to depression.

5.3 Perception

Adolescent MID users used more words relating to auditory experiences (effect size = 0.078) in their Reddit posts when discussing depression compared to adults by 21.12%. Adult MID users, on the other hand, were more likely to reference language relating to visual attributes (effect size = -0.053; 7.8% higher). Adult MID users also focused more on time (effect size = -0.087; 3.5% higher) and used more past-tense (effect size = -0.106; 6.2% higher) to express their depression, whereas adolescent MID users used more future-oriented language (effect size = 0.059; 5.4% higher) in their posts.

> "I'm afraid of what I might do to myself when I leave for college" (↑ adolescent)

> "What really hurt me [...] was, after all these years, coping with an inner bully" (↑ adult)

The adolescent and adult subgroups in the CTL cohort did not exhibit noticeable differences in their use of temporal language (4.4%) compared to adolescents and adults in the MID cohort (6.99%).

5.4 Lexical Density

Lexical density in the Reddit posts of adult MID users is higher compared to the adolescent subgroup, as observed through the usage of prepositions (effect size = -0.127; 2.5% higher) and articles (effect size = -0.219; 8.6% higher). However, compared to the adolescent MID users, adults had a lower proportion of verbs (effect size = 0.142; 2.3% lower), auxiliary verbs (effect size = 0.090; 2.1% lower), and adverbs (effect size = 0.148; 4.7% lower). Per Figure 2, the aggregrate of the mean differences in lexical density between adolescents and adults for the MID cohort is greater than the difference between the two subgroups in the CTL sample.

5.5 Social and Personal Concerns

In contrast to the previous four categories of LIWC measures, the mean difference for the aggregate of social LIWC measures between adolescents and adults is higher in the control group (9.93%) compared to MID users (6.58%), as depicted in Figure 2. However, when examining specific LIWC measures within the social/personal category, adolescent and adult MID users exhibit distinct linguistic patterns associated with social and personal concerns in their Reddit posts.

To start, adolescent MID users used more words relating to 'friendship' (effect size = 0.083) and 'family' (effect size = 0.050) than adult MID users, by 13.2% and 7.5% respectively, and discussed social relationships 60.5% more than neurotypical adolescents.

> "I found out that my father had been cheating on my mum and now my life is going downhill" (↑ adolescent)

Adult MID users, on the other hand, associated depression with 'work' (effect size = -0.113; 11.8% higher) more than the adolescent subgroup. They also used more words relating to 'money' (effect size = -0.297) by 78.5%.

> "my mind is going haywire. Money, my career, school, should I quit my job, I hate my job" (↑ adult)

The Reddit posts of adult MID users concerning depression exhibited a higher emphasis on wellness (effect size = -0.117; 66.6% higher), substances (effect size = -0.042; 17.7% higher), and sexual matters (effect size = -0.032; 12.5% higher).

5.6 Interpersonal Focus

Adolescent MID users showed a higher use of personal pronouns (z = 0.271; 5.68% higher), an indicator of self-focus. Per Figure 2, for the CTL cohort, the interpersonal focus measures account for a difference of 2.95% between adolescent and adult users, while for MID users, the difference is more considerable at 5.38%, which indicates distinct variations beyond the control sample.

5.7 Topical Differences

Topic modeling revealed significant differences in the topics discussed in relation to depression between adolescents and adults in the MID and CTL cohorts.

Four topics from both the adolescent MID and adult MID subgroups stood out as having apparent clusterings and term overlap. For adolescent MID users, the terms associated with topic #1 relate to friendships, specifically the transience of social relationships (e.g., 'friend', 'year', 'day', 'still', 'long', trying'). The rest of the topics center largely around parents and family, with topic #7 relating to parental pressures (e.g., 'parent', 'dad', 'mom', 'always', 'talk'), topic #3 relating to family responsibilities (e.g., 'help', 'parent', 'work',

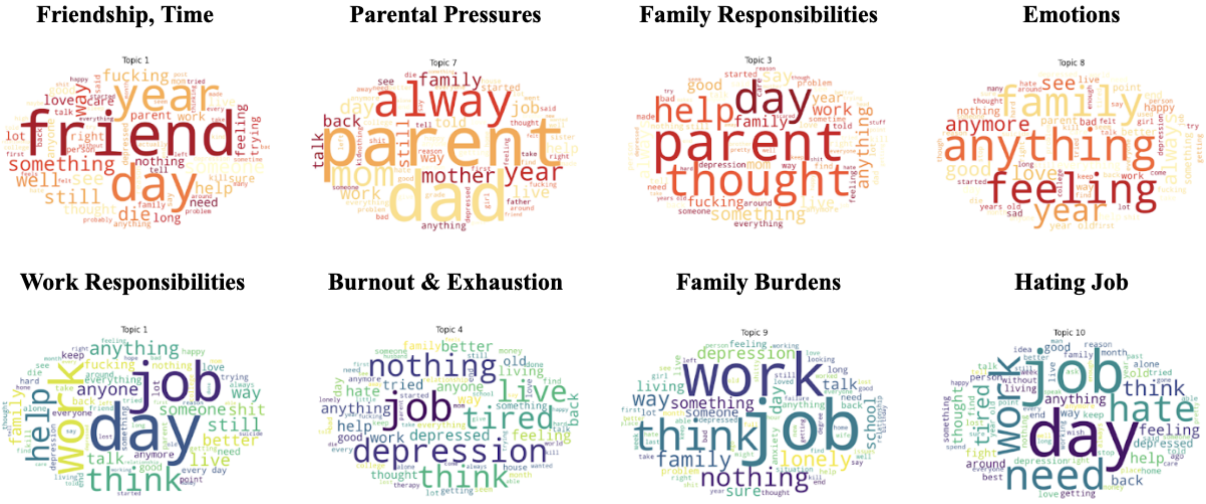

Figure 3: Topics significantly associated with depression posts of adolescents (top) and adults (bottom).

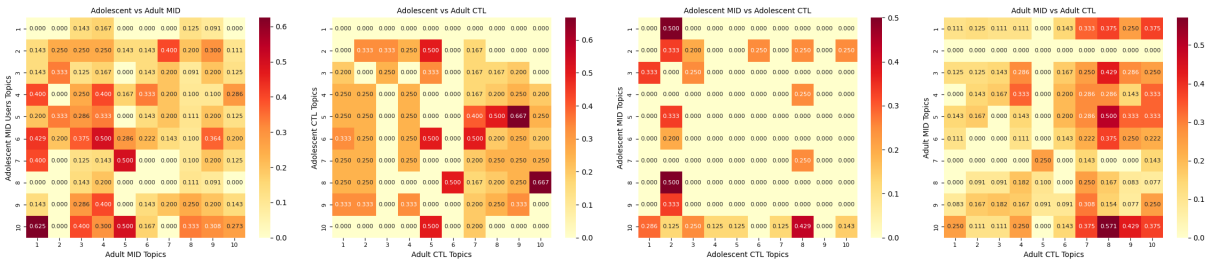

Figure 4: Jaccard similarity coefficients of topics comparing MID adolescent vs adult, CTL adolescent vs adult, adolescent MID vs CTL, and adult MID vs CTL. Darker shades indicate stronger similarities between topics.

problem', 'bad'), and topic #8 relating to emotions (e.g., 'family', 'anymore', 'sad', 'love', 'feeling'), per Figure 3. For adult MID users, the topics center largely around work (e.g., 'work', 'day', 'job', 'help', 'think' in topic #1) and exhaustion (e.g., 'nothing',' tired', 'live', 'depression' in topic #4), as seen in Figure 3. Topic #9 and topic #10 also focus on work, but in the context of family burdens (e.g., 'work', 'family', 'lonely', 'living') and negative emotions (e.g., 'job', 'hate', 'need', 'getting') respectively.

Figure 4 depicts the term overlap of topics discussed within the adolescent and adult groups in the MID and CTL cohorts. Approximately 21 out of 100 topic comparisons between adolescents and adults with depression exhibited a similarity coefficient above 0.3, indicating about a third of shared terms. This implies that 79% of topic comparisons had statistically insignificant term overlap, indicating distinct topics for adolescents relative to adults. In contrast, when comparing adolescents and adults in the control group, only 17 out of 100 topic comparisons had a similarity coefficient of 0.3 or higher. Comparisons between MID and control users within the specific age groups showed even lower term overlap. Specifically, among the topic comparisons of adolescents MID and CTL users, only 7 topics had a similarity coefficient of 0.3 or higher. For the comparison between adult MID and CTL, 15 out of 100 topic comparisons had a similarity coefficient of 0.3 or higher.

6 Discussion

This study has three main findings. First, adolescents and adults with depression share several topic similarities with each other (21% of topics in the MID sample exhibited significant term overlap for the two age groups, while this was only 17% in the control sample). The shared topics revolved predominantly around family matters and negative emotions, aligning with previous research that depressed users tend to exhibit higher levels of sadness, anxiety, and anger in their online writings (Schwartz et al., 2014; Preoţiuc-Pietro et al., 2015). This finding may suggest that depression has language markers that distinguish it from standard

social media content, while also highlighting the universality of negative emotion and family challenges as markers of depression that transcend age groups.

Second, adolescents and adults with depression possess several distinct linguistic features of depression specific to their age group. Adolescent social media users with depression were more likely to associate their condition with social relationships, using words relating to friendships more frequently than adults with depression. This finding remained significant when controlling for neurotypical users, indicating that the topic of social relationships may be uniquely correlative with depression for adolescents. In contrast, adults with depression were more likely to associate their condition with job responsibilities, financial status, and health, evident in the higher prevalence of terms related to money and physical health (i.e. exhaustion, sexual matters, and substances). This not only aligns with prior research that depressed adults display more somatic symptoms than younger individuals with depression (Fiske, 2009), but also highlights how adults discuss their depression differently than adolescent users on social media.

Adolescent social media users with depression exhibited a higher usage of future-tense language when discussing their mental health experiences, suggesting a preoccupation with future stressors potentially correlative with the onset of depression. In contrast, adults with depression employed more past-tense language, indicating a focus on past grievances or regrets. This differs from prior research with neurotypical social media users, where as users increased in age, they used more future-tense (Pennebaker and Stone, 2003), indicating that these temporal linguistic features may be unique to depression and can be age-specific language markers.

Third, the linguistic features that distinguish depression were more prominent and effective among adolescent social media users compared to adults. While adults with and without depression had several similarities in the topics they discussed (15% of topics had significant term overlap), the topics discussed by adolescents with and without depression had considerably less overlap (only 7%). This suggests that among older social media users, the correlative topics of depression may become less discernible and intertwined with normal discussion topics. Prior research suggests that the reason for this discrepancy is because adults are less open to expression on social media compared to younger individuals (Pennebaker and Stone, 2003; Barbieri, 2008); other results from the present study also supports this finding, as adults were shown to use more impersonal and less emotional expression when discussing their depression compared to adolescent users. Overall, this finding indicates that identifying depression expression on social media platforms may be more difficult within adult subpopulations compared to adolescent groups.

7 Limitations

The present study had limitations that should be acknowledged. First, the infrequency of age self-disclosure on Reddit made it challenging to obtain samples for a wider range of age categories, such as individuals over the age of 60. This resulted in a dataset predominantly composed of users aged 13-40, as seen in Figure 1, necessitating a focus on binary age groups rather than narrower ranges. The age group of ”adults” in the present study may be more advantageously interpreted as ”young adults.” Further research is thus needed with a larger amount of age-labeled data for analysis of more precise age ranges.

Another limitation was the topic modeling process, where the interpretation of results relied on observation rather than a psychologist's expertise. This introduces potential bias, and the topics should therefore be validated by further research. Future studies should also employ more rigorous linguistic models and methodologies for precise topic categorization.

Moreover, due to Reddit's interactive forum structure, the posts analyzed in the present study may have belonged to larger chains/dialogues that influenced the user's behavior and language use. Future studies should incorporate additional contextual information to avoid hidden confounding variables.

Finally, the sample of Reddit users may not be representative of the global population, limiting the generalizability of our findings. The present study focused solely on Reddit and the English language, while mental health discussions may occur on other platforms or in different languages. Future research should consider incorporating data from multiple platforms and languages for a more comprehensive understanding.

8 Conclusion

This exploratory study contributes to a growing body of literature on mental health expression in online language data by quantitatively examining age-based linguistic differences in expressing depression on Reddit. By investigating social media data through psycholinguistic feature extraction and open-vocabulary topic modeling, adolescents and adults with depression were found to engage in discussions about several common topics; however, they also exhibited distinct age-specific linguistic markers, with adolescents being more prone to associating their depression with friendships, while adults were more likely to focus on work and health concerns. Depression language markers were also found to be more discernible among adolescent social media users compared to adult users. These findings highlight the importance of recognizing language variations across age groups when detecting depression on social media.

Acknowledgements

I would like to express my gratitude to Audrey Acken and Matthew Hesby for their encouragement, help and valuable feedback for my research.

References

Cheryl Ahrens. 2002. Age differences and similarities in the correlates of depressive symptoms. *Psychology and aging*, 17:116–24.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in american english1. *Journal of Sociolinguistics*, 12:58 – 88.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Lera Boroditsky, Lauren A. Schmidt, and Webb Phillips. 2003. *Sex, syntax and semantics.*, chapter Sex, syntax and semantics. The MIT Press.

A. Boyd, Ashokkumar, Seraj, and Pennebaker. 2022. The development and psychometric properties of LIWC-22.

Nina L. Cesare, Christan Earl Grant, and Elaine Okanyene Nsoesie. 2017. Detection of user demographics on social media: A review of methods and recommendations for best practices. *ArXiv*, abs/1702.01807.

Stevie Chancellor, Zhiyuan Jerry Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. In *International Conference on Web and Social Media*.

Rob Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, and Mario Navarro. 2021. Correction: Predicting age groups of reddit users based on posting behavior and metadata: Classification model development and validation. *JMIR public health and surveillance*, 7:e30017.

Munmun Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. pages 626–638.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*.

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, René Clausen Nielsen, and Georgia Tech. 2016. Quantifying and understanding gender and cross-cultural differences in mental health expression via social media.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Social communication*.

Glen A. Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *CLPsych@ACL*.

Aron Culotta. 2014. Estimating county health statistics with twitter. *Conference on Human Factors in Computing Systems - Proceedings*.

Johannes Eichstaedt, Margaret Kern, David Yaden, H. Schwartz, Salvatore Giorgi, Gregory Park, Courtney Hagan, Victoria Tobolsky, Laura Smith, Anneke Buffone, Jonathan Iwry, Martin Seligman, and Lyle Ungar. 2020. Closed and open vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations.

Johannes Eichstaedt, H. Schwartz, Margaret Kern, Gregory Park, Darwin Labarthe, Raina Merchant, Sneha Jha, Megha Agrawal, Lukasz Dziurzynski, Maarten Sap, Christopher Weeg, Emily Larson, Lyle Ungar, and Martin Seligman. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26.

Gatz Fiske, Wetherell. 2009. Depression in older adults. *Annual Review of Clinical Psychology*, pages 363–389.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, A. 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. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Christopher Homan, Naiji Lu, Xin Tu, Megan C. Lytle-Flint, and Vincent Michael Bernard Silenzio. 2014. Social structure and depression in trevorspace. *CSCW : proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, 2014:615 – 625.

Lilly C. Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial computing: a lens on design and development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Fazida Karim, Azeezat Abimbola Oyewande, Lamis F Abdalla, Reem Chaudhry Ehsanullah, and Safeera Khan. 2020. Social media use and its connection to mental health: A systematic review. *Cureus*, 12.

J. Kaufman, A. Martin, R. A. King, and D. Charney. 2001. Are child-, adolescent-, and adult-onset depression one and the same disorder? *Biological Psychiatry*, 49(12):980–1001.

Tingting Liu, Lyle Ungar, Brenda Curtis, Garrick Sherman, Kenna Yadeta, Louis Tay, Johannes Eichstaedt, and Sharath Chandra Guntuku. 2022. Head versus heart: social media reveals differential language of loneliness from depression. *npj Mental Health Research*, 1:16.

Kate Loveys, Jonathan Torrez, Alex B. Fine, Glendon L. Moriarty, and Glen A. Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *CLPsych@NAACL-HTL*.

Daniel Mark Low, Laurie Rumker, Tanya Talkar, John B Torous, Guillermo A. Cecchi, and Satrajit S. Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, 22.

Diana L. MacLean, S. Gupta, Anna Lembke, Christopher D. Manning, and Jeffrey Heer. 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

Douglas Maynard and Anssi Peräkylä. 2006. *Language and Social Interaction*, pages 233–257.

Juhi Mittal, Abha Belorkar, Venu Pokuri, Vinit Jakhetiya, and Sharath Chandra Guntuku. 2023. Language on reddit reveals differential mental health markers for individuals posting in immigration communities.

Elizabeth Murnane and Scott Counts. 2014. Unraveling abstinence and relapse: Smoking cessation reflected in social media. *Conference on Human Factors in Computing Systems - Proceedings*.

J. Ormel, M. VonKorff, T. B. Ustun, S. Pini, A. Korten, and T. Oldehinkel. 1994. Common mental disorders and disability across cultures. results from the WHO collaborative study on psychological problems in general health care. *JAMA*, 272(22):1741–1748.

V Patel, Ricardo Araya, Marcia Guerino De Lima, Ana Bernarda Ludermir, and Charles Todd. 1999. Women, poverty and common mental disorders in four restructuring societies. *Social science & medicine*, 49 11:1461–71.

Michael Paul and Mark Dredze. 2011. You are what your tweet: Analyzing twitter for public health. *Artificial Intelligence*, 38:265–272.

James W. Pennebaker and Lori D. Stone. 2003. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85 2:291–301.

Inna Loginovna Pirina and Çagri Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Conference on Empirical Methods in Natural Language Processing*.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Nairán Ramírez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. *Proceedings of the International AAAI Conference on Web and Social Media*.

Ps Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The university of maryland clpsych 2015 shared task system. pages 54–60.

Frances Rice, Lucy Riglin, Terri C. Lomax, E. Souter, Robert Potter, DJ Smith, Ajay K Thapar, and Ajay K Thapar. 2019. Adolescent and adult differences in major depression symptom profiles. *Journal of affective disorders*, 243:175–181.

H. Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook.

H. Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. pages 55–60.

Rachel E. Spector. 1991. Cultural diversity in health and illness. *Journal of Transcultural Nursing*, 13:197 – 199.

Shelley E. Taylor and J D Brown. 1988. Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103 2:193–210.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. Reddust: a large reusable dataset of reddit user traits. In *International Conference on Language Resources and Evaluation*.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

World Health Organization. 2017. *Depression fact sheet*.

Satyanarayana Yatam and T. Raghunadha Reddy. 2014. Author profiling: Predicting gender and age from blogs, reviews & social media. *International journal of engineering research and technology*, 3.