

Loneliness Detection from Social Media: A Text Analytics Approach

Winston Fan
West Senior High School
winstonfan1256@gmail.com

Jonathan Fan
Yale University
jonathan.fan@yale.edu

Jiyuan Yu
University of Iowa
jiyyu@uiowa.edu

Min Zhang
Communication University of
China
mzhang@cuc.edu.cn

Qianzhou Du
University of Science and
Technology of China
qianzhoudu@ustc.edu.cn

Ling Tong, Weiguo Fan
University of Iowa
{ling-tong, weiguo-
fan}@uiowa.edu

Abstract

In an era where digital interactions significantly influence our social interactions, understanding how loneliness is expressed online becomes paramount. This study delves into the linguistic representation of loneliness on Reddit, utilizing techniques in natural language processing (NLP) and machine learning. By employing frequency-based, similarity-based, and association-based methods, a unified lexicon was generated, demonstrating promising performance in classifying loneliness-related posts. The identification and validation of 536 most impactful entries from this lexicon underscore their predictive power and genuine relevance as markers of loneliness. This research advances our comprehension of loneliness within digital contexts and underscores the potential of computational methods in detecting and addressing loneliness online. The identified lexicon lays the groundwork for AI-driven mental health interventions, underscoring the significance of language in understanding and addressing loneliness in the digital age.

Keywords: Loneliness, NLP, healthcare, lexicon, machine learning, design science

1. Introduction

Loneliness, characterized by a distressing experience associated with a perceived lack of social connection, has emerged as a significant public health concern globally. This pervasive issue impacts individuals of all ages and demographics, leading to serious mental and physical health implications. The subjective nature of loneliness, coupled with its substantial implications for mental and physical health,

underscores the necessity of addressing this issue comprehensively (Fox, 2019). Over the past few years, social media has expanded exponentially, with 72% of United States (US) adults using some form of it. This is especially true among young people, with 84% of those aged between 18-29 using it in 2021 (Auxier, 2021). While social media platforms offer significant opportunities for individuals seeking friendships and connections, they can paradoxically lead to feelings of exclusion and a sense of being less popular. With the recent Covid-19 pandemic came a lack of social interaction, and for those isolated, social media was the only form of interaction they had. The pandemic led to peak loneliness levels with 52% of Americans reporting feeling lonely and 80% of young people under the age of 18 reported feeling lonely. (Zauderer, 2023) We can see this increase in loneliness levels by looking at the growth in the r/lonely subreddit on Reddit where users from across the world share their own experiences in regard to loneliness and offer guidance to others. The subreddit surged from under

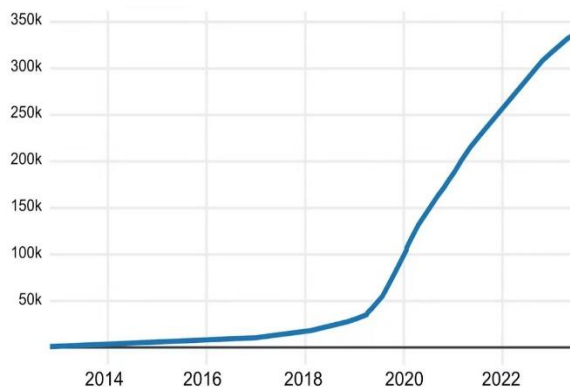


Figure 1. Increase in subreddit loneliness users (Venkatesh, 2023)

50,000 members to 350,000 during the pandemic as seen in Figure 1 (Venkatesh, 2023). This significant rise in membership highlights a growing trend of individuals seeking online communities to express and share their experiences with loneliness.

In the digital age, the ubiquity of online forums and social media platforms presents a novel opportunity for the detection and analysis of loneliness. These platforms, where users frequently and openly share their emotions and experiences, provide a valuable dataset for detecting indicators of loneliness. The anonymous and often supportive environment of these platforms can encourage more open and honest sharing of personal struggles, including feelings of loneliness. This phenomenon provides a unique vantage point for researchers aiming to understand and address loneliness through computational means.

The potential of utilizing online forums and social media posts for loneliness detection lies in their ability to provide real-time, authentic insights into individuals' emotional states and social interactions. Unlike traditional survey methods, which can be restricted by recall bias and social desirability bias, online platforms capture spontaneous expressions of emotion, offering a more accurate reflection of the user's current state. This approach not only facilitates a broader understanding of loneliness dynamics but also aids in early detection, enabling timely intervention (Cacioppo & Patrick, 2008). Current literature explores various methods for detecting loneliness in online texts, including machine learning algorithms and natural language processing (NLP) techniques. These methods analyze textual features, sentiment, and linguistic patterns to infer the emotional state of the users (Ahmed et al., 2022). By analyzing these digital traces, we can find patterns and trends that might otherwise go unnoticed.

However, existing approaches face limitations, such as the challenge of accurately interpreting the nuanced expressions of loneliness. Furthermore, the reliance on complex models can hinder the scalability and accessibility of these detection methods (Burnap et al., 2015). In response to these challenges, the development of a lexicon specific to loneliness presents a promising alternative. A lexicon-based approach involves identifying and cataloging words and phrases commonly associated with expressions of loneliness, offering a more straightforward and interpretable means of analysis (Fan et al., 2005; Khoo et al., 2017). Lexicons can be easily updated and adapted to different contexts, making them more flexible and user-friendly. Additionally, they can be implemented with less computational power, making the approach more accessible to a broader range of users and applications.

Literature on the application of lexicons for loneliness is still nascent but promising. Studies have shown that lexicons can effectively capture the emotional valence of texts, suggesting their potential for loneliness detection (Pennebaker et al., 2007). By building on these foundational insights, this study aims to develop a comprehensive lexicon of loneliness derived from Reddit posts. Reddit, with its diverse user base and wide-ranging discussions, provides an ideal source of data for this research. By analyzing the language used in Reddit posts, we can identify common terms and phrases that are indicative of loneliness. This approach not only addresses the limitations of current methods but also contributes to the existing body of literature by providing a scalable and interpretable tool for loneliness detection.

For the following sections, section 2 provided the research context and detailed the data collection and annotation process, laying the foundation for subsequent analysis. Section 3 outlined the lexicon generation method, emphasizing the importance of a lexicon-based approach for overcoming the limitations of existing methods. Section 4 presented the results of the lexicon analysis, demonstrating its efficacy in identifying markers of loneliness in online discourse. Lastly, section 5 offered conclusions and discussions drawn from the study, highlighting the potential for AI-driven interventions and the significance of language in understanding and addressing loneliness in the digital age.

2. Research Context and Data

2.1. Research Context

Lonely subreddits serve as fertile grounds for uncovering lexicons related to loneliness due to the richness of content they offer. Within these forums, individuals share personal narratives, reflections, and engage in discussions about their experiences of loneliness. This diverse array of user-generated content provides researchers with abundant material for identifying and analyzing vocabulary specific to loneliness. The dynamic linguistic environment of these subreddits fosters creative language use, contributing to the emergence of a nuanced lexicon specific to loneliness. Users often employ inventive expressions and metaphors to articulate their feelings, enriching the lexicon with diverse linguistic nuances. This linguistic diversity reflects the multifaceted nature of loneliness and offers researchers valuable insights into its various dimensions. Moreover, the anonymity afforded by Reddit serves as a catalyst for open and honest expression. Users feel free to share their

innermost thoughts and experiences without fear of judgment, resulting in candid and introspective discussions. This authenticity of expression contributes to the authenticity of the linguistic data available for study, allowing researchers to gain a deeper understanding of the intricacies of loneliness and its expression in digital contexts.

2.2. Data Collection and Annotation

Reddit's platform provides researchers with convenient access to extensive datasets via its public APIs, facilitating streamlined data collection and analysis. Leveraging the API, we conducted a thorough crawl of the r/lonely subreddit (top 1% by size with over 373,000 members), amassing a substantial dataset comprising approximately 1,000,000 posts. While Reddit provides valuable real-time, authentic narratives, it also presents challenges. The platform is known for the presence of bots, engagement farming, and fake stories, which can introduce noise and affect the reliability of the data. However, larger subreddits like r/lonely are moderated by experienced, credible volunteers, which helps filter out a significant portion of inauthentic content. While these moderation efforts do not eliminate all risks, they do enhance the reliability of the data used in this research. By acknowledging these challenges and the mitigations in place, we believe that the data still provides a valuable foundation for capturing loneliness.

Within the Lonely subreddit, users engage in discussions pertaining to various aspects of loneliness. Our focus lies in identifying posts where individuals articulate their personal experiences of loneliness and extracting relevant lexicons. To effectively filter such posts, we adhere to three criteria (Bonsaksen et al., 2023). Firstly, loneliness must originate from perceived deficiencies within the user's social relationships. This could manifest as feelings of disconnection, alienation, or a lack of meaningful interaction with others. Secondly, we recognize that loneliness is a highly subjective experience, distinct from objective isolation. While someone may be physically alone, they may not necessarily feel lonely if they perceive their solitude as peaceful or enjoyable. Conversely, individuals can experience profound loneliness even when surrounded by others if they feel disconnected or misunderstood. Understanding this distinction allows us to appreciate the complexity of loneliness and its varied manifestations. Lastly, loneliness is characterized by its unpleasant and distressing nature, signifying emotional discomfort, and yearning for connection. By applying these criteria, we aim to pinpoint posts that authentically convey users' experiences of loneliness.

We conducted a random sampling of 8,000 posts and implemented a binary labeling system to annotate the data, where '1' denoted the presence of loneliness, while '0' indicated its absence. This systematic approach ensured clarity and consistency in identifying posts that conveyed feelings of loneliness. To enhance the reliability of our annotations, four annotators with diverse perspectives participated in the annotation process to ensure comprehensiveness and balance. Inter-rater reliability was assessed using the overall kappa score (McHugh, 2012), yielding a high score of 0.72, validating the robustness of our labeling methodology. In cases of disagreement, collaborative discussions and additional reviewer input were utilized to achieve consensus, ensuring accuracy and consistency in the labeling process.

3. Lexicon Generation Methods

Frequency-Based Method: we compute TFIDF (Term Frequency-Inverse Document Frequency) scores for unigrams and bigrams in the posts, ranks them by their scores, and selects the top 2000 highest-scoring n-grams. TFIDF is a widely used technique in natural language processing for identifying important words or phrases in a corpus. It works by evaluating how frequently a term appears in a specific document relative to its frequency across all documents in the corpus, thereby highlighting terms that are significant within individual contexts but not ubiquitous across the entire dataset. In our approach, we first compute the term frequency (TF) for each unigram and bigram within individual posts. This measures how often a word or phrase occurs in a specific post, providing a sense of its prominence in that context. Next, we calculate the inverse document frequency (IDF) for each term, which assesses how common or rare a term is across all posts. The product of these two measures, the TFIDF score, allows us to identify terms that are not only frequent in specific posts but also distinctive across the entire collection of posts. By calculating TFIDF scores for both unigrams (single words) and bigrams (pairs of consecutive words), we capture a broader range of linguistic features. Unigrams can highlight individual words that are significant, while bigrams can reveal common phrases or combinations of words that may have special meaning. In the context of loneliness detection, this method is particularly effective. Words and phrases that are uniquely associated with loneliness are likely to have higher TFIDF scores because they are used frequently in posts about loneliness but are less common in other types of posts. Therefore, by selecting the top 2000 highest-scoring unigrams and bigrams, we can isolate the key linguistic features that are most indicative of loneliness.

This selection process is crucial for building an accurate and efficient lexicon for loneliness detection. It ensures that our model focuses on the most relevant and telling terms, improving its ability to recognize and analyze expressions of loneliness in new posts. This method enhances the precision and reliability of our loneliness detection framework, contributing to more effective identification and understanding of loneliness in digital communication.

Similarity-Based Method: we start with four seed words related to loneliness ('lonely', 'loneliness', 'isolation', 'lonesomeness'). These seed words provide a foundational reference point for our analysis. The process begins by tokenizing posts into words, computes word embeddings using a BERT model (Devlin et al., 2019). Next, we calculate the cosine similarity between each word in the posts and the seed words. Cosine similarity measures the degree of similarity or relatedness between the word embeddings. We then rank all the words based on their similarity to the seed words and select the top 2,000 words with the highest similarity. Word embeddings generated by a BERT model, capture semantic relationships between words by representing them in a high-dimensional vector space. Words in the posts that are similar to these seed words in the embedding space are likely to share semantic characteristics associated with loneliness, even if they are not explicitly mentioned in the seed list. By identifying and selecting words with high cosine similarity to the seed words, we can pinpoint words that are semantically related to loneliness. This approach allows us to capture a broader spectrum of language associated with loneliness, thus enhancing our understanding of how loneliness is expressed and discussed in various contexts.

Association-Based Method: We begin by identifying unigrams in the posts. We then compute the association between each unigram and the posts classified as related to loneliness versus those classified as unrelated to loneliness using the chi-square test. The chi-square test is a statistical method employed to determine the association between two categorical variables, in this case, the occurrence of each unigram in loneliness-related posts compared to its occurrence in non-loneliness-related posts. By comparing the frequency of each unigram across these two categories, we can assess whether the unigram is significantly associated with loneliness. Words are then ranked based on their chi-square scores, and the top 2000 unigrams with the highest scores are selected for further analysis. The chi-square test provides a quantitative measure of how strongly each unigram is associated with loneliness. Unigrams that exhibit a significantly higher occurrence in loneliness-related

posts compared to non-loneliness-related posts are considered to be indicative of loneliness.

This method allows us to systematically identify words that are statistically associated with loneliness within the dataset. By selecting words with high chi-square scores, we ensure that the identified words are not only relevant but also have a robust statistical association with loneliness. This approach helps in capturing the nuances of language that signify loneliness, providing a comprehensive understanding of how loneliness is linguistically expressed. The use of chi-square scores as a criterion for selection ensures that the identified words are not arbitrarily chosen but are supported by a solid statistical foundation, thereby enhancing the reliability and validity of the findings.

4. Results

To gauge the efficacy of the three lexicons generated through Frequency-based, Similarity-based, and Association-based methods, we developed a series of XGBoost models. Each of these models progressively incorporated an increasing number of top-ranked entries, starting from the top 100 entries and extending up to the top 2000 entries, with a step size of 100. The predictive performance of these models was evaluated by classifying posts into loneliness or non-loneliness categories, utilizing the F1 score as the primary metric for assessment. The results revealed that the Frequency-based method achieved its highest F1 score of 0.7514 with the top 1500 entries. The Similarity-based method demonstrated its peak performance with an F1 score of 0.7512, utilizing the top 1900 entries. Meanwhile, the Association-based method achieved an F1 score of 0.7458 with the top 1000 entries (Table 1).

Table 1.

Lexicons	Frequenc y-based	Similarit y-based	Associati on-based	Unified
F1 scores	0.7514	0.7512	0.7458	0.7538

Considering the distinctive methodologies of each approach, combining them promises to yield a comprehensive range of linguistic markers that signify loneliness. By integrating the lexicons derived from the top-performing XGBoost models—specifically, the Frequency-based method with the top 1500 entries, the Similarity-based method with the top 1900 entries, and the Association-based method with the top 1000 entries—a unified lexicon comprising 2888 unique entries was obtained. This consolidated lexicon was

then utilized to train a new XGBoost model aimed at classifying posts into loneliness versus non-loneliness categories, achieving an improved F1 score of 0.7538. Analysis of feature importance revealed that the optimal model utilized 536 out of the 2,888 entries (Table 1). This indicates that a select subset of the entries plays a crucial role in enhancing the model's performance. The identification of these 536 entries as significant suggests that they contain key information critical for distinguishing between loneliness and non-loneliness posts. This selective utilization underscores the value of these specific entries in contributing to the model's accuracy and robustness, demonstrating their unique relevance within the broader set of available features.

To further evaluate the utility of the 536 entries, additional features were incorporated to discern whether they provide redundant information or offer unique insights. Two types of features were considered: traditional features commonly employed in literature, encompassing topics, sentiment/emotions, readability, and lexical diversity, among others, and article embeddings generated by a BERT model, which capture contextual nuances. Despite the integration of these supplementary features, the performance was notably enhanced by the incorporation of the 536 entries. This significant improvement suggests that the lexicon captures additional, valuable information that aids in the accurate classification of loneliness versus non-loneliness posts (see Table 2). The enhanced performance highlights the lexicon's unique contribution, providing insights that are not fully covered by traditional features or advanced contextual embeddings alone. This finding underscores the importance of the 536 entries in enriching the feature set and enhancing the model's precision in distinguishing between loneliness and non-loneliness content.

Table 2.

	Traditional features	Traditional features and article embeddings
Without 536 entries	0.7410	0.7520
With 536 entries	0.7781	0.7861

To ascertain whether the identified lexicon serve as genuine markers of loneliness or merely represent fortuitous linguistic patterns aiding predictive performance, a cluster analysis was conducted using embeddings of the 536 entries. The Elbow method determined that four clusters were appropriate for the

analysis. These clusters each highlight different facets and contexts of loneliness, shedding light on how loneliness is experienced and understood across a spectrum of emotions, situations, and social contexts. Below is a summarization of each cluster along with their key entries.

Cluster 1: addresses the complex emotional and social layers of loneliness, encompassing relationships, personal feelings, and societal influences. Key entries include "family," "emotions," "relationship," "trust," "anxiety," "friendship," "social," "isolation," and "love." The presence or absence of family connections plays a crucial role in feelings of loneliness. This cluster reflects the nature of internal emotional experiences and external social factors, illustrating how deeply loneliness can be rooted in both personal and societal contexts. For example, the presence or absence of family connections plays a crucial role in feelings of loneliness, highlighting the significance of familial relationships. Emotions encompass a wide range of states associated with loneliness, including sadness, despair, and hopelessness, capturing the internal turmoil individuals often experience. The quality and nature of personal relationships, such as romantic partnerships, friendships, and familial bonds, are central to understanding loneliness, as they influence one's sense of connectedness and belonging. Trust is a fundamental part of meaningful relationships, and its absence can lead to feelings of alienation and loneliness. Loneliness is often related with anxiety, particularly social anxiety, creating a cycle that hinders social interactions and worsens isolation. Friendships are critical to social life and can significantly affect one's experience of loneliness, with supportive friends providing a buffer against feelings of isolation. The broader social context, including societal norms, expectations, and the impact of social networks, also plays a significant role. Isolation, both physical and social, is a key factor in the experience of loneliness, examining different forms from living alone to feeling disconnected in a crowd. Love, in its many forms, is a powerful counterforce to loneliness, providing emotional support and a sense of belonging, while the pain of unrequited love or loss can deepen feelings of loneliness.

This cluster covers a broad spectrum, from the internal experiences of loneliness and anxiety to the external factors that can contribute to these feelings, such as social isolation and the impact of social media.

Cluster 2: focuses on time and significant life events that can lead to feelings of loneliness. Key entries are "home," "hometown," "months," "career," "college," "years," and "school." This cluster suggests that loneliness can be tied to specific periods in one's life, such as transitions and changes, reflecting on past

experiences, or looking forward to the future. For instance, “home” and “hometown” evoke feelings of nostalgia and a sense of belonging that might be missed when one is away, highlighting how physical locations associated with different life stages can influence loneliness. The term “months” implies temporal aspects, indicating that certain times of the year, perhaps holidays or anniversaries, can trigger loneliness. The entries “career” and “college” reflect major life transitions and milestones that often come with significant social and environmental changes, such as starting a new job or moving to a new city for higher education. These periods are marked by the disruption of established social networks and the challenge of forming new connections, which can heighten feelings of isolation. Similarly, “years” can denote the passage of time and the cumulative impact of life events on one’s sense of loneliness, whether through long-term isolation or the progressive loss of close relationships. “School” represents another critical period, particularly for younger individuals, where social dynamics and peer relationships play a vital role in emotional well-being.

This cluster illustrates how significant life transitions, whether they involve moving to a new place, starting a new phase of education, or entering the workforce, can profoundly affect one’s experience of loneliness. It highlights the importance of understanding the temporal and event-specific contexts that contribute to loneliness, emphasizing the need for support systems during these critical periods to mitigate the impact of loneliness.

Cluster 3: delves into more intense feelings of loneliness, often coupled with negative self-perception and personal struggles. Key entries include “panic,” “scared,” “hatred,” “insecurities,” “breakup,” “depressing,” “loneliness,” “hate,” and “sad.” This cluster captures the deep, internal struggles associated with loneliness, encompassing a range of mental health issues, feelings of inadequacy, and the impact of negative experiences. For instance, “panic” and “scared” reflect the acute anxiety and fear that can go with profound loneliness, often worsening feelings of isolation and helplessness. The entries “hatred” and “hate” point to the intense negative emotions that can arise from loneliness, including self-loathing and resentment towards others, which can further isolate individuals from potential support networks. “Insecurities” highlights the feelings of inadequacy and low self-esteem that often go with loneliness, creating a vicious cycle where negative self-perception inhibits social interaction, leading to deeper isolation. The term “breakup” signifies a specific and impactful event that can trigger intense loneliness, reflecting the emotional turmoil and sense of loss that follow the end of a

sincere relationship. Similarly, “depressing” and “sad” encapsulate the pervasive feelings of sorrow and hopelessness that are central to the experience of loneliness, often overlapping with clinical depression and other mental health conditions. The entry “loneliness” itself underscores the central theme of this cluster, emphasizing the profound sense of disconnection and isolation that defines the experience.

This cluster underscores the severity of the internal battles associated with loneliness, highlighting the intersection of loneliness with broader mental health issues and personal struggles. It illustrates how loneliness can intensify negative self-perception and worsen feelings of inadequacy, creating significant emotional distress. The presence of terms like “panic” and “breakup” indicates that loneliness is often accompanied by acute emotional pain and significant life disruptions. By capturing these intense emotional and psychological dimensions, this cluster emphasizes the need for comprehensive mental health support and interventions to address the underlying issues that contribute to and result from profound loneliness.

Cluster 4: highlights the actions people take to address or cope with loneliness and the challenges they face in doing so. Key entries are “makes,” “interested,” “harder,” “ready,” “opportunity,” “trying,” “help,” “playing,” “hope,” and “talk.” It reflects on the efforts to make connections, the desire for improvement, and the challenges in finding meaningful relationships or overcoming personal obstacles. For instance, the entry “makes” suggests the initiatives and steps individuals take to foster social interactions and create opportunities for connection. “Interested” indicates the importance of mutual interest and engagement in forming and maintaining relationships, highlighting the active role that curiosity and attentiveness play in overcoming loneliness. The term “harder” acknowledges the difficulties and increased effort required to navigate social interactions and build meaningful connections, especially for those already struggling with feelings of isolation. “Ready” and “opportunity” reflect the readiness and openness to new experiences and relationships, emphasizing the potential for growth and change despite the challenges. “Trying” captures the ongoing efforts and perseverance individuals demonstrate in their quest to overcome loneliness and build social bonds. The entry “help” underscores the significance of seeking and providing support, whether through professional means like therapy or through personal networks of friends and family. “Playing” indicates engaging in activities, hobbies, or social games as a means of distraction or a way to meet new people and develop a sense of belonging. “Hope” reflects the optimism and forward-looking attitude that individuals maintain despite their

struggles, emphasizing the importance of a positive outlook in coping with loneliness. Finally, “talk” highlights the fundamental role of communication in addressing loneliness, whether it involves talking about one’s feelings with others or engaging in social conversations to foster connections.

This cluster illustrates the active strategies and hopeful perspectives people adopt to manage loneliness. It emphasizes the importance of persistence, openness to new opportunities, and the critical role of support and communication in overcoming the challenges of loneliness. By focusing on these proactive elements, the cluster underscores the dynamic and multifaceted efforts required to combat loneliness and highlights the resilience and determination of individuals striving to improve their social well-being.

5. Conclusion and Discussion

This study employed three distinct methodologies to identify lexicons indicative of loneliness in Reddit posts. Each of the lexicons derived from these methods exhibited promising performance in classifying loneliness posts versus non-loneliness posts. Upon consolidating the lexicons generated by these methods, the highest achieved F1 score was 0.7538, highlighting the robustness of the combined approach. Subsequent feature importance analysis of the best-performing model using the unified lexicons identified the 536 most impactful entries. The 536 most impactful entries underwent further evaluation for their predictive power by integrating additional features. Despite the inclusion of supplementary features, integrating the 536 entries notably enhanced performance, indicating their capture of vital information essential for precise classification. To determine the genuine nature of the identified lexicon as markers of loneliness, a cluster analysis was conducted using embeddings of the 536 entries. The Elbow method determined that four clusters were appropriate for the analysis, each revealing distinct facets and contexts of loneliness. This clustering analysis reaffirmed the validity and comprehensiveness of the lexicon, shedding light on how loneliness is expressed across various emotional, social, and situational dimensions.

The study’s implications are manifold, extending to the development of AI-driven interventions that could proactively identify and address loneliness on digital platforms. Clustering entries in the lexicon into different categories reveals the varied ways loneliness is expressed, enhancing the potential for tailored support in online communities. For instance, understanding the specific emotional and situational triggers of loneliness can help in crafting more

personalized interventions, thereby improving user engagement and well-being. Furthermore, this research underscores the importance of linguistic cues in psychological assessment. The findings suggest that language-based markers can be instrumental in developing novel diagnostic tools and therapeutic strategies.

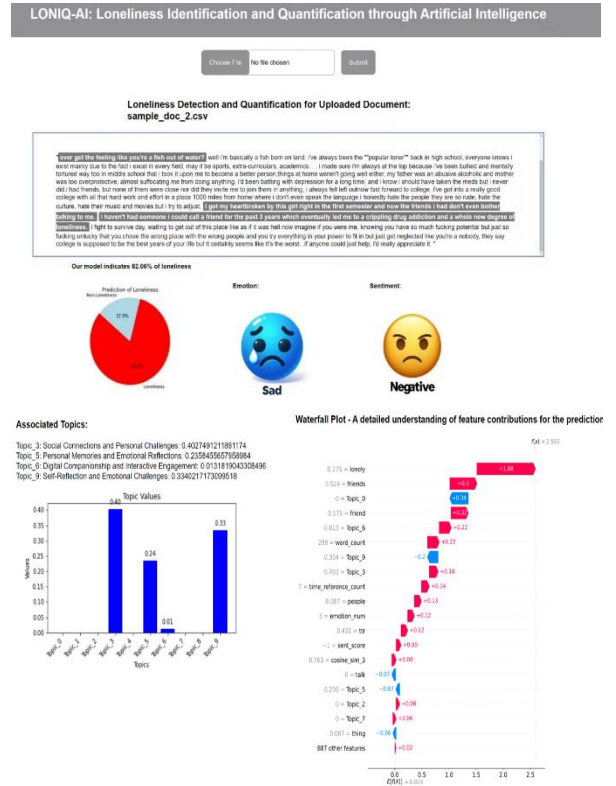


Figure 2. Dashboards of the Web-based loneliness detection and quantification system.

Implementing an XGBoost model that combines default features with the Lexicon feature set, we can also develop a web-based system for the detection and quantification of loneliness in new posts. This system utilizes the strengths of the XGBoost algorithm, known for its predictive accuracy and efficiency, along with the tailored lexicon that encapsulates key terms associated with loneliness. As new posts are input into the system, it analyzes the text and provides insights into the levels or aspects of loneliness expressed within. As we can see from Figure 2, a variety of dashboards within web-based loneliness detection and quantification system, offer distinct insights into loneliness from various perspectives. These dashboards present different types of analyses or visualizations based on the data extracted from new posts. This includes metrics like loneliness prediction, emotions/sentiment analysis, topic analysis, feature importance in prediction decision making and sentence

highlighting (The sentences which has high frequency of loneliness-related words are highlighted).

By providing a multifaceted view, these dashboards enable users to understand and interpret the aspects of loneliness in a comprehensive and nuanced manner, catering to different analytical needs or areas of interest. This could serve as a valuable tool for social media platforms, mental health professionals, or community support groups to identify and offer support to individuals who may be expressing feelings of loneliness.

Validation of these findings in clinical and therapeutic contexts may pioneer new approaches to mental health diagnostics and intervention. By emphasizing the critical role of linguistic cues, this study highlights the potential for more nuanced and accurate psychological assessments, contributing to better mental health outcomes. Integrating these insights into digital mental health platforms could lead to more effective monitoring and support for individuals experiencing loneliness, using AI to offer prompt and contextually proper interventions. Future research should consider the dynamic nature of language, exploring the temporal and cultural fluidity of the loneliness lexicon and its translation across various digital ecosystems. This could involve analyzing how expressions of loneliness evolve over time and differ across cultural contexts, ensuring the lexicon remains relevant and effective in diverse settings.

In summary, the study provides a comprehensive approach to finding and understanding the linguistic markers of loneliness, proving their utility in both digital and clinical settings. It paves the way for future research and applications that harness the power of language to improve mental health diagnostics and interventions, highlighting the profound impact of integrating linguistic analysis into psychological support frameworks.

6. References

- Auxier, B. (2021, April 7). Social Media Use in 2021. Pew Research Center: Internet, Science & Tech. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>. Is Social Media Making You Lonely? (2018). Psychology Today. <https://www.psychologytoday.com/us/blog/modern-mentality/201810/is-social-media-making-you-lonely>
- Ahmed, E., Xue, L., Sankalp, A., Kong, H., Matos, A., Silenzio, V., & Singh, V. K. (2022). Predicting Loneliness through Digital Footprints on Google and YouTube. *Electronics*, 12(23), 4821. <https://doi.org/10.3390/electronics12234821>
- Bonsaksen, T., Ruffolo, M., Price, D., Leung, J., Thygesen, H., Lamph, G., Kabelenga, I., & Geirdal, A. Ø. (2023). Associations between social media use and loneliness in a cross-national population: do motives for social media use matter? *Health Psychology and Behavioral Medicine*, 11(1). <https://doi.org/10.1080/21642850.2022.2158089>
- Burnap, P., Rana, O., Avis, N. J., Williams, M. L., Housley, W., Edwards, A., Morgan, J., & Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, 95, 96–108. <https://doi.org/10.1016/j.techfore.2013.04.013>
- Cacioppo, J. T., & Patrick, W. H. (2008). Loneliness: human nature and the need for social connection. *Choice Reviews Online*, 46(03), 46–1765. <https://doi.org/10.5860/choice.46-1765>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. <https://arxiv.org/pdf/1810.04805v2>
- Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems*, 40(2), 213–233. <https://doi.org/10.1016/j.dss.2004.02.003>
- Fox, B. (2019). Loneliness and social media: A qualitative investigation of young people's motivations for use, and perceptions of social networking sites. In *Springer eBooks* (pp. 309–331). https://doi.org/10.1007/978-3-030-24882-6_16
- Khoo, C. S. G., & Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4), 491–511. <https://doi.org/10.1177/0165551517703514>
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 276–282. <https://doi.org/10.11613/bm.2012.031>
- Miller, G. A. (1995). WordNet. *Communications of the ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count (LIWC2007).
- Venkatesh, S. (2023, October 27). *SMPNet: A Novel Approach to Combat Loneliness on Social Media*. Medium; Medium. <https://medium.com/@coolsumukh/smpnet-a-novel-approach-to-combat-loneliness-on-social-media-788abaec5f28>
- Zauderer, S. (2023, July 23). *49 Loneliness Statistics: How Many People are Lonely?* Crossrivertherapy.com; Cross River Therapy. <https://www.crossrivertherapy.com/research/loneliness-statistics>