Human Centered Computing in Mental Health: Twitter Sentiment Analysis

By

Fardin Ahmed

CS 4395 – Senior Project Report

In partial Fulfillment of a Bachelor's of Science in Computer Science

Department of Computer Science and Engineering Technology

University of Houston – Downtown

Fall 2023

Faculty Advisor:
Dr. Dvijesh Shastri

Dvijesh Shastri

Digitally signed by Dvijesh Shastri
DN: cn=Dvijesh Shastri, o=University of Houston - Downtown, ou=Department of Computer Scinece and Engineering Technology, email=shastrid@uhd.edu, c=US
Date: 2023.11.29 11:41:59 -06'00'

Course Instructor: Dr.
Ling Xu

Table of Contents

Abstract

Taking on an exploration of a branch of Human-Centered Computing in sentiment analysis and topic modeling within the dynamic Twitter landscape, this research project seeks to unravel the intricate array of emotions, concerns, and trends embedded in user-generated content. Utilizing a multifaceted approach that leverages natural language processing and machine learning, put together in a software, the study sets clear objectives to gain profound insights into the digital realm.

The journey begins with meticulous data collection from diverse Twitter sources. Rigorous processing eliminates noise, establishing a robust foundation for subsequent analyses. Effective data management ensures seamless analysis by organizing and structuring datasets.

The application of sentiment analysis techniques proves powerful for unraveling emotional meanings within millions of tweets, providing a comprehensive view of prevailing sentiments. Today’s modern machine learning enhances the accuracy and depth of sentiment analysis results, allowing for diverse categorization beyond simple positive or negative labels. Continuous refinement of models adapts to evolving linguistic trends, providing a contextually aware interpretation of online expressions. This multifaceted approach enriches comprehension of the emotional tapestry in social media, with implications for applications from brand monitoring to public opinion analysis.

In the realm of topic modeling, Scikit-Learn's toolkit uncovers latent topics within Twitter data. Machine learning methodologies further refine exploration, offering a detailed understanding of prevalent themes in users’ content.

Prioritizing user interaction, a user-friendly interface enables dynamic exploration of results whether it’s used by the user themselves or by appropriate analytics. Challenges encountered during the project are candidly discussed, shedding light on intricacies in sentiment analysis and topic modeling.

In summary, this research not only enriches academic discourse but also holds implications for social awareness and mental health advocacy. The abstract encapsulates a comprehensive journey through data collection, processing, sentiment analysis, and topic modeling, paving the way for future research and innovation. Additionally, the project underscores the relevance of Twitter sentiment analysis in Human-Centered Computing, offering valuable insights for user behavior feedback and interaction analysis.

Motivation Behind Project

The choice of this project stems from deep concerns for the often-neglected issue of mental health. In society, the distinction between physical health and mental health is stark and obvious. While neglecting physical health typically impacts the individual, overlooking mental health can have repercussions that extend beyond the individual to the broader community.

Physical health issues arising from poor diet or lack of exercise primarily harm the individual's well-being. Mental health, on the other hand, possesses the potential to affect those in the individual's vicinity. The implications range from harm to personal relationships to, in extreme cases, posing a threat to public safety, as seen in tragic incidents like mass shootings.

The media's tendency to sideline discussions on mental health until a catastrophic event occurs perpetuates the stigma associated with it. Instances such as mass shootings or the tragic suicide of public figures often serve as catalysts for brief, intense discussions on mental health. However, these conversations are reactive, occurring after irreparable and extreme damage has been done.

One saddening example is the events in Uvalde, Texas, a tragic incident unfolded as the community grappled with the repercussions of neglected mental health. Despite mounting concerns about a high school student named Alex, whose subtle cries for help went unnoticed, interventions were delayed, leading to a devastating school shooting that claimed innocent lives. This heart-wrenching event underscored the consequences of ignoring mental health issues, emphasizing the urgent need for communities to prioritize awareness, understanding, and timely support to prevent silent struggles from escalating into irreparable harm.

Human society faces challenges in early detection of individuals with mental health issues due to social stigmatization and the tendency of those affected to conceal their struggles. The human tendency to overlook or dismiss obvious signs compounds the difficulty in identifying potential issues.

Here shows the principles for employing technology, particularly sentiment analysis on social media, as a proactive tool in addressing mental health concerns. Unlike humans who can be manipulated away from the truth, computers operate without biases, diligently analyzing vast amounts of data in real-time. They are not susceptible to the deception or self-denial that often accompanies mental health issues.

Modern technology has proven effective in early detection and prevention in various domains, such as the early detection of cancer. Applying a similar principle, the project seeks to leverage technology to identify potential mental health concerns before they escalate into serious issues, thereby preventing harm to both individuals and their communities.

The motivation behind this project is rooted in the belief that technological advancements can play an important role in reshaping the narrative around mental health. By employing technology to identify and intervene in potential mental health issues, we aspire to contribute to a society where mental well-being is prioritized, stigma is dismantled, and individuals receive timely support on their journey to mental health.

Acknowledgements

The conceptualization and execution of Human-Centered Computing in Mental Health demanded meticulous guidance from professors, advisors, online research, and tutorial resources.

This interdisciplinary project, straddling both computer science and mental health, presented unique challenges that were overcome with the generous support of several individuals.

Foremost among these contributors is Carolina Hernandez, my girlfriend and best friend, who has played a pivotal role in shaping my understanding of mental health. Carolina, a dedicated 5th-grade NAC (Newcomer Academy) teacher, specializes in science education. Her commitment extends beyond subject matter expertise, as she works with students who have recently arrived in America, many of whom possess limited English proficiency. Each of these students brings a unique set of challenges, considering their diverse backgrounds, which may impact their mental well-being, particularly at such a young age. Carolina graduated with a Bachelor of Arts in Interdisciplinary Studies from the University of Houston Downtown. Our shared life experiences and struggles have significantly deepened our awareness of mental health realities, especially in children and what they face growing up. This shared understanding has fostered a strong commitment between us to support each other and extend assistance to those grappling with mental health concerns. I am profoundly grateful for her invaluable contributions to both my personal and professional growth.

Aisha Khan, a close friend, and peer with a major in Psychology, has also played a pivotal role. Through shared life experiences, we explored the shaping of mental health by traditions and the imperative to break the silence surrounding mental health concerns. Aisha is currently pursuing her dream of becoming a surgeon, a testament to her dedication and knowledge in various mental health issues.

My peers and guides, Vijay Chaudhary, Pujan Acharya, and Nahian Sharif provided indispensable assistance in shaping ideas and methodologies related to computer science, including the development of this project outline. Vijay and Pujan graduated with a Bachelor of Science in Computer Science from the University of Houston Downtown, marking their commitment to excellence. Nahian will soon graduate alongside.

My everlasting pillar of support is my mother, Bilquis Ahmed, a certified respiratory therapist with over 20 years of experience in HCC hospitals. Her insights into the mental processes of patients, particularly during the challenges of the COVID era, have been invaluable. She imparted crucial life lessons, emphasizing the importance of understanding and respecting the mental health concerns of others.

Lastly, I extend my gratitude to Dr. Dvijesh Shastri, my mentor and supporter for this senior project. As my professor during the Senior Seminar class in Spring 2023, Dr. Shastri's guidance, feedback, and insights have played a pivotal role in refining my presentation skills and enhancing my ability to communicate efficiently. His mentorship has been instrumental in shaping my journey as a spokesperson and academic professional.

Chapter 1: Description of Project

1.1 Introduction

In our modern society, the influence of social media platforms on shaping public opinion has become dominant. This study hunts into the complex domain of sentiment analysis on Twitter data, with a specific emphasis on recognizing patterns, trends, and valuable insights associated with mental health. Twitter's pervasive nature as a communication tool makes it a rich repository for examining public sentiment on various topics, including those related to mental health.

The digital age has witnessed a paradigm shift where individuals increasingly use social media platforms for self-expression and communication. As among a primary platform, Twitter encapsulates a broad spectrum of human emotions, providing a distinctive lens through which to examine mental health concerns. Users frequently leverage Twitter to share their thoughts, experiences, and emotions, extending to matters concerning mental well-being.

A crucial surface of this exploration involves the identification of expressions indicative of mental health concerns, with a specific focus on depression. The informal and blunt nature of tweets affords users the ability to authentically express their feelings. Consequently, the platform emerges as a valuable reservoir for discerning linguistic cues, emotional expressions, and patterns associated with individuals undergoing or alluding to depression.

The endeavor to comprehend how users express themselves in ways that signal mental health concerns extends beyond a mere exercise in data analysis. It represents a vital step towards cultivating a supportive digital environment. The potential for early detection of signs related to depression through social media channels holds promise for ease timely interventions, support systems, and resources for individuals in need. By deciphering the refined language embedded within tweets, the objective is to contribute to an expanding body of research aimed at bridging the gap between online expression and mental health awareness.

This report travels on a delicate exploration, navigating the multifaceted landscape of sentiments on Twitter, with a dedicated emphasis on expressions related to mental health, particularly depression. As we search into the intricacies of users' contents, the aspiration is to illuminate not only the sentiments expressed but also to recognize and respond to the fine calls for support embedded within the digital dialogue.

This study aligns with the principles of Human Centered Computing, offering insights for user behavior feedback through a twitter notification, either through the Twitter app or mobile push notifications, and analysis in either clinical studies or law enforcements. The utilization of Twitter sentiment analysis contributes to the understanding of users' mental well-being, promoting a human centric approach in the digital realm.

1.2 Background

Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a field within natural language processing that involves evaluating and interpreting subjective information expressed in text. The objective is to discern the sentiment or emotion conveyed by the author, whether it be positive, negative, or neutral. As the digital era unfolds, sentiment analysis has become an invaluable tool for gauging public opinion, understanding consumer attitudes, and exploring the emotional undertones within vast datasets.

Importance of Sentiment Analysis

Sentiment analysis, a multifaceted tool with applications ranging from business to healthcare, including mental health awareness and in-depth social studies, plays an important role in shaping decision-making processes and understanding the dynamics of human expression.

In the corporate environment, businesses utilize sentiment analysis to gain profound insights into customer satisfaction and brand perception. By dissecting users' contents across online platforms, companies can unravel not only overall sentiment but also specific aspects that drive consumer opinions. This understanding empowers businesses to adapt their strategies, enhance products or services, and cultivate a more resonant brand image. Whether it's gauging reactions to a new product release or assessing responses to marketing campaigns, sentiment analysis provides a valuable lens through which businesses can refine their approach and stay attuned to customer needs.

In the field of social sciences, sentiment analysis serves as a dynamic tool for figuring public sentiment on critical issues, tracking trends, and identifying potential areas of concern. By analyzing the sentiments expressed in social media discussions, forums, and public discourse, researchers can discern collective attitudes, uncover emerging themes, and track the evolution of social perspectives. This information is instrumental in informing public policies, addressing social concerns, and fostering a deeper understanding of the prevailing sentiments within diverse communities.

Sentiment analysis has found a significant application in healthcare, particularly in the context of mental health awareness and research. By analyzing textual data related to mental health shared on social media platforms, researchers can identify linguistic cues, emotional patterns, and expressions indicative of individuals undergoing mental health challenges. This insight not only contributes to a better understanding of mental health issues but also paves the way for early detection, intervention, and the development of targeted support systems. The integration of sentiment analysis in mental health studies holds promise for creating more responsive and empathetic healthcare strategies tailored to the needs of individuals facing mental health challenges.

Human-Centered Computing

Human-Centered Computing (HCC) is an interdisciplinary field that focuses on the design and implementation of computing systems centered around human needs, behaviors, and capabilities. It emphasizes creating technology that is accessible, user-friendly, and enhances the overall human experience. Examples of HCC include user interface design, accessibility features in software and hardware, and the development of interactive technologies like smartwatches and virtual reality devices. In the context of sentiment analysis, HCC relies on understanding human emotions, behaviors, and communication patterns to tailor computing systems that better serve and interact with users. The insights gained from sentiment analysis contribute to creating human-centric digital environments, shaping technology to align with users' emotional and communicative needs. This integration enhances user experience, facilitates more meaningful interactions, and promotes a supportive digital ecosystem. Sentiment analysis, within the realm of HCC, becomes a valuable tool for analyzing the emotional cues of users, providing feedback, and adapting computing systems to better cater to the diverse spectrum of human emotions and expressions.

Literature:

Numerous studies have delved into the intricacies of sentiment analysis, each contributing to the evolving landscape of this field. Research explores different methodologies, from rule-based systems to machine learning approaches, seeking to enhance the accuracy and efficiency of sentiment classification. Literature also emphasizes the challenges associated with sentiment analysis, such as context-dependent language, sarcasm, and the evolving nature of internet language.

My exploration into sentiment analysis builds upon this rich body of literature, focusing on its application in deciphering sentiments related to mental health within the expansive realm of Twitter data. By leveraging the insights gained from existing research, this project aims to contribute to the ongoing dialogue on sentiment analysis applications, particularly in areas of societal importance such as mental health awareness.

1.3 Objectives

Embarking on the objective of solving the intricate sentiments expressed on Twitter, our project is propelled by a concise set of well-defined objectives, focusing on mental health awareness, specifically securing in on indicators of depression. Each objective is crafted to align with our overarching mission, ensuring a comprehensive exploration of the digital dialogue.

Develop a Precision-Crafted Algorithm:

Our foremost objective and advantage of this project is to design and implement a sentiment analysis algorithm tailored to accurately identify tweets conveying sentiments related to mental health, specifically focusing on indicators of depression. Precision is key, aiming to extract meaningful insights from the vast ocean of user-generated content on Twitter.

Measure Success:

To gauge the efficacy of our sentiment analysis algorithm, we aim for a measurable accuracy rate of 75% or higher. Strict testing and validation processes, employing various machine learning models such as Naive Bayes, SVM, and Logistic Regression, will ensure the reliability and consistency of our classification results.

Balance Ambition with Feasibility:

Anchoring our ambitions, we leverage machine learning models and natural language processing techniques to analyze a dataset of a minimum 100,000 tweets across various trends. This ensures computational feasibility without compromising the depth of our analysis.

Ensure Relevance:

Every objective is intricately woven into the fabric of our mission. We commit to aligning sentiment analysis results with existing mental health frameworks and literature, providing contextually relevant insights that contribute meaningfully to the discourse surrounding mental well-being.

Temporal Commitment:

Time-bound objectives are the backbone of our project. We set a timeframe of four months for both studies in mental health and the causes, and the completion of the development and testing phase, allowing for continuous refinement based on ongoing feedback and advancements in sentiment analysis methodologies.

Forge Collaborative Partnerships:

Beyond algorithmic success, our objectives extend to establishing partnerships with mental health organizations, professionals, peers, articles on Reddit and Quora regarding users' stories, and those with significant life experiences. This collaborative approach aims to validate the project's impact and promote ongoing collaboration in promoting mental health awareness.

Data Sourcing for Sustainable Analysis:

To address financial constraints associated with leveraging social media APIs, we adopt an alternative approach, strategically utilizing historical Twitter data obtained through ethical and legal means. This allows for a consistent and sustainable stream of information, ensuring robust sentiment analysis without incurring excessive costs.

Contribute to Ongoing Initiatives:

Our relevance extends to contributing to ongoing research and initiatives led by mental health organizations and various online platforms. We envision our project becoming an integral part of broader efforts aimed at enhancing mental health support mechanisms.

Timely Milestones:

Establishing partnerships and gathering initial feedback by my mentor will occur within the first four months, followed by regular weekly check-ins and updates to ensure the continual refinement and impact of our project.

Chapter 2: Project Framework

2.1 Data Collection

Source of Twitter Data

The primary and agreeable source of our Twitter sentiment analysis project lies in the rich patchwork of data chosen from Kaggle.com. Kaggle, a prominent platform for datasets and data science competitions, provides an expansive collection of datasets spanning various domains. This project strategically leverages Kaggle's diverse range of datasets, which vary in size from 10,000 to over a million tweets in workable CSV files, to conduct an in-depth exploration of user sentiments, with a specific focus on expressions related to mental health within the dynamic Twitter community.

Inclusivity Across Twitter Data Trends

Mental health-related expressions are not confined to specific topics; they permeate discussions across diverse trends and themes on Twitter. Our datasets encompass a broad spectrum of subjects, including sports, politics, climate change, COVID-19, and more. This inclusivity ensures that the sentiment analysis extends beyond niche topics, capturing genuine user expressions related to mental well-being across countless contexts.

Analysis Approach

The decision to extract mental health-related expressions from datasets spanning various trends aligns with our commitment to conducting a comprehensive analysis. By researching diverse topics, we aim to uncover different ways in which users express their sentiments, providing an overall understanding of the intersection between mental health awareness and diverse online conversations.

Ethical Considerations

Ethical considerations will remain vital. Our approach clings to ethical data usage practices, and compliance with Kaggle's terms of service is maintained rigorously. The datasets used are carefully selected to exclude any personally identifiable information, safeguarding user privacy and upholding the highest standards of data protection.

2.2 Data Processing

Our approach to identifying tweets signifying mental health concerns, particularly those related to depression, involves a multi-step process. This section outlines the algorithmic procedures deployed for efficiently categorizing and analyzing tweets within each dataset considering every dataset is formed differently.

Detection of Likely Tweet Column

Before researching into categorization, we need to consider that each dataset obtained will be formatted differently. The algorithm begins by identifying the column in each dataset that most likely contains tweet text. This dynamic algorithm assesses columns based on criteria such as average text length, non-empty percentage, and uniqueness of values. By iteratively evaluating these criteria, the algorithm determines the most likely tweet column, ensuring adaptability across diverse datasets.

Categorization and Analysis of Tweets

Once the tweet column is identified, the algorithm proceeds to categorize and analyze tweets based on predefined categories related to depressive expressions. The categories surround a spectrum of emotional states, self-esteem, isolation, suicidal thoughts, coping mechanisms, physical symptoms, relationships, negative thoughts, hopelessness, and mental health awareness.

Aligning relevant and Contextual Tweets

To ensure relevance, the categorized tweets are aligned with existing mental health frameworks and literature, providing contextually relevant insights that contribute meaningfully to the discourse surrounding mental well-being.

2.3 Sentiment Analysis

Sentiment analysis endeavors to uncover the emotional distinctions embedded within tweets, with a particular focus on mental health-related expressions. This area provides a comprehensive overview of the methodology and tools employed in this sophisticated process.

Methodology Overview

The sentiment analysis methodology comprises several key steps, each designed to extract meaningful insights from the vast array of user-generated content on Twitter.

1. Raw Tweet Processing

The first step involves processing raw tweets, encompassing the entire spectrum of user expressions on diverse topics. This processing phase includes text cleaning, removal of irrelevant information, and normalization to ensure consistency across datasets.

2. Categorization of Depressive Expressions

Utilizing a predefined set of categories related to mental health, tweets are categorized based on the presence of specific phrases or keywords associated with depressive expressions. This categorization forms the foundation for subsequent sentiment analysis.

3. Sentiment Score Assignment

Each tweet undergoes sentiment analysis using the TextBlob library, which assigns a sentiment score ranging from -1 (most negative) to 1 (most positive). This score quantifies the overall emotional tone conveyed by the tweet.

4. Creation of Filtered CSV Files

The classified category, sentiment score, and the original tweet text are recorded in new filtered CSV files. These files serve as curated repositories of tweets expressing sentiments related to mental health, providing a structured dataset for further analysis.

Tools Utilized

1. TextBlob Library

The sentiment analysis process heavily relies on the TextBlob library, a powerful tool for processing textual data. TextBlob's sentiment analysis module enables the extraction of sentiment scores, contributing to the understanding of emotional tones within tweets.

2. Python Programming Language

The entire sentiment analysis workflow is implemented using the Python programming language, leveraging its versatility and extensive libraries for natural language processing and data analysis.

Unification of Filtered Datasets

To consolidate the insights derived from various datasets, the filtered CSV files are unified and merged into a comprehensive dataset. This unification process ensures a holistic representation of depressive expressions across diverse Twitter trends and topics. This unified dataset serves as a valuable resource for subsequent analyses, providing a consolidated perspective on mental health-related sentiments expressed on Twitter.

2.4 Data Management

In sentiment analysis, effective data management lays the foundation for a systematic and organized exploration of Twitter datasets. Our approach to data management within Jupyter Notebook environment is characterized by its dynamic creation of directories, efficient organization of raw and filtered datasets, and the implementation of TXT file apprehension to track dataset processing status. This comprehensive strategy ensures transparency, traceability, and systematic processing throughout the sentiment analysis workflow.

Dynamic Folder Creation

Our Jupyter Notebook environment dynamically generates folders to house raw, filtered, and combined datasets. This approach ensures a structured and organized workspace for the sentiment analysis project. The following directories are dynamically created:

- Raw Dataset Folder: Stores original Twitter datasets obtained from Kaggle.com.
- Filtered Dataset Folder: Houses datasets that have undergone sentiment analysis, categorization, and classification.
- Combined Dataset Folder: Contains the fused dataset resulting from the unification of filtered datasets.

TXT File Apprehension

To enhance the control and traceability of dataset processing, we employ two TXT files:

- available_datasets.txt: Lists the names of raw datasets retrieved from Kaggle.com.
- filtered_tweets_records.txt: Maintains a record of datasets that have undergone sentiment analysis and categorization.

These TXT files serve as checkpoints to identify datasets that have not yet been processed. The integration of TXT file apprehension allows for a systematic comparison, ensuring that datasets are processed only if they have not previously undergone sentiment analysis.

Combined Dataset Categorization

The combined dataset is further organized based on classified categories. Dynamic creation of folders for each category facilitates systematic storage of categorized CSV files. This categorization enhances accessibility and simplifies subsequent analyses, allowing for focused exploration of specific mental health-related themes.

Sustainability and Transparency

In addition to organizational efficiency, our data management strategy underscores sustainability and transparency. The integration of TXT file apprehension not only enhances the systematic processing of datasets but also upholds ethical considerations by preventing redundant analyses of previously processed datasets.

In essence, our dynamic and comprehensive data management approach sets the stage for a thorough exploration of sentiments expressed on Twitter, particularly focusing on mental health-related expressions. Through structured organization, systematic tracking, and categorization, our project ensures a robust and transparent sentiment analysis workflow.

2.5 Machine Learning

Our exploration into sentiment analysis extends beyond just categorization. It crawls into the complex and endless domain of machine learning, employing a complex approach to uncover patterns, balance, and characteristics within mental health-related Twitter datasets. The machine learning phase involves the segregation of datasets, examination of tweet characteristics, and the application of diverse models for sentiment classification.

Segregation of Datasets

The combined dataset is methodically segregated into categories corresponding to various mental health issues. This segmentation enables a distinct analysis, allowing us to gain insights into specific themes and expressions related to distinct aspects of mental health.

Analysis of Tweet Characteristics between Positive and Negative Sentiments

Before searching into machine learning models, we conduct a thorough examination of the balance and characteristics of positive and negative tweets within each dataset. This preliminary step ensures a comprehensive understanding of the dataset's composition and paves the way for informed model selection.

Application of Machine Learning Models

Our approach involves the utilization of prominent machine learning models for sentiment classification:

- Logistic Regression: Leveraged for its simplicity and efficiency in binary classification tasks.
- Naive Bayes: Applied for its probabilistic nature and effectiveness in text classification.
- Support Vector Machines (SVM): Employed for its robustness in handling high-dimensional data and versatility in classification tasks.

Performance Assessment Metrics

To calculate the efficacy of our machine learning models, we employ a range of performance assessment metrics:

- Accuracy: Measures the overall correctness of the classification.
- Precision: Evaluates the accuracy of positive predictions.
- Recall: Assesses the ability of the model to capture all positive instances.
- F1 Score: Coordinate precision and recall, providing a balanced metric.

Additionally, we measure True Positive (TP) and True Negative (TN) rates, along with False Positive (FP) and False Negative (FN) rates. This refined evaluation offers a comprehensive understanding of the model's performance across diverse dimensions.

Dataset Variants

Our machine learning endeavors extend to three distinct dataset variants:

- Individual Kaggle Datasets: Analyzing sentiment within each individual filtered dataset.
- Combined Dataset: Exploring sentiments within the overall combined dataset, providing a holistic perspective.
- Classified Categorized Datasets: Focusing on sentiment patterns within categorized datasets, shedding light on specific mental health-related themes.

Our machine learning phase integrates versatility and depth, unraveling sentiment patterns within the varied landscape of mental health-related tweets on Twitter. Through careful model selection, performance assessment, and dataset segmentation, we strive to contribute meaningful insights to the evolving discourse on mental health awareness.

2.6 Visualization

In the visual exploration of sentiment within the realm of mental health-related tweets, our project employs compelling visualization techniques to distill patterns, emphasize themes, and communicate refined insights. The distinct use of word clouds and horizontal bar charts offers a rich assortment of visual representations, providing stakeholders with an intuitive understanding of the sentiments pervading Twitter's digital landscape.

Word Clouds

Word clouds serve as dynamic and engaging snapshots, visually representing the frequency and prominence of words within the corpus of tweets. Through the application of word clouds, we encapsulate the essence of sentiments, allowing users to discern prevalent themes and focal points briefly. The size and prominence of words directly correlate with their frequency, creating an impactful visual narrative of the prevailing sentiments.

Horizontal Bar Graphs

Horizontal bar charts emerge as powerful tools for dissecting sentiment categories and their relative prevalence. Through categorization and visualization, we unravel the distribution of sentiments across various mental health themes. Each bar represents a specific sentiment category, and its length correlates with the frequency or intensity of occurrences. This methodical representation enables a comparative analysis, shedding light on the prominence of different sentiments within the dataset.

Insightful Visual Narratives

The collaboration of word clouds and horizontal bar charts promotes a visually compelling narrative, enhancing the interpretability of sentiment patterns. Analysts can navigate through the visual landscape, discerning key sentiments, identifying prevalent expressions, and gaining an exact understanding of the emotional contours within the Twitter datasets. This visual journey not only aids in comprehending sentiment distributions but also serves as a valuable tool for raising awareness and facilitating meaningful discussions surrounding mental health.

2.7 Topic Modeling

In the pursuit of unraveling the intricate environment of sentiments expressed on Twitter, the incorporation of Topic Modeling becomes a vital component of our project framework. Topic Modeling is a specialized technique employed for uncovering latent topics within a corpus and analyzing document similarity. As a fundamental aspect of our sentiment analysis workflow, we seamlessly integrate the powerful Scikit-Learn library to extract meaningful topics from the vast variety of user-generated content.

Process Overview

The incorporation of Scikit-Learn into our workflow involves a systematic process designed to enhance our understanding of tweet themes and trends. This toolkit plays a crucial role in the identification and extraction of both common and detailed topics. To ensure a focus on mental health-related discussions, we utilize depressive words and phrases, which were used earlier to filter the datasets, as additional stop words, tailoring the analysis to align with the project's overarching mission.

Revealing Latent Topics

Scikit-Learn facilitates the extraction of latent topics within tweets, providing a comprehensive view of the underlying themes. By applying advanced algorithms, we can identify topics that may not be immediately apparent through traditional sentiment analysis. This approach adds a layer of depth to our analysis, enabling us to uncover subtle prevalent trends in the digital dialogue surrounding mental health.

Enhancing Comprehension

The benefits of incorporating Topic Modeling extend beyond the identification of themes. It enhances our comprehension of the reasons behind mental health issues. Through the extraction of meaningful topics, we gain insights into the prevalent concerns, discussions, and sentiments that contribute to the broader narrative. This deeper understanding is instrumental in contextualizing the sentiment analysis results and providing a more detailed perspective on the dynamics of mental health discourse on Twitter.

2.8 User Interface

In crafting a user interface, the project seamlessly integrates software engineering principles to deliver a robust, flexible, and intuitive platform within the familiar confines of Jupyter Notebook. This interface acts as a gateway to the intricate landscape of sentiment analysis, machine learning, and data exploration, providing users with the tools to navigate and derive insights effortlessly.

Flexibility and Efficiency in Jupyter Notebook

The utilization of Jupyter Notebook serves as the foundational framework for our project, offering flexibility in handling diverse tasks seamlessly. Backend operations, including the integration of essential libraries and tools for data analysis, are efficiently managed within this environment. Jupyter Notebook's interactive nature promotes a dynamic and iterative analytical process, enabling users to explore, experiment, and refine their approach in real-time.

Intuitive Options for Dataset Analysis

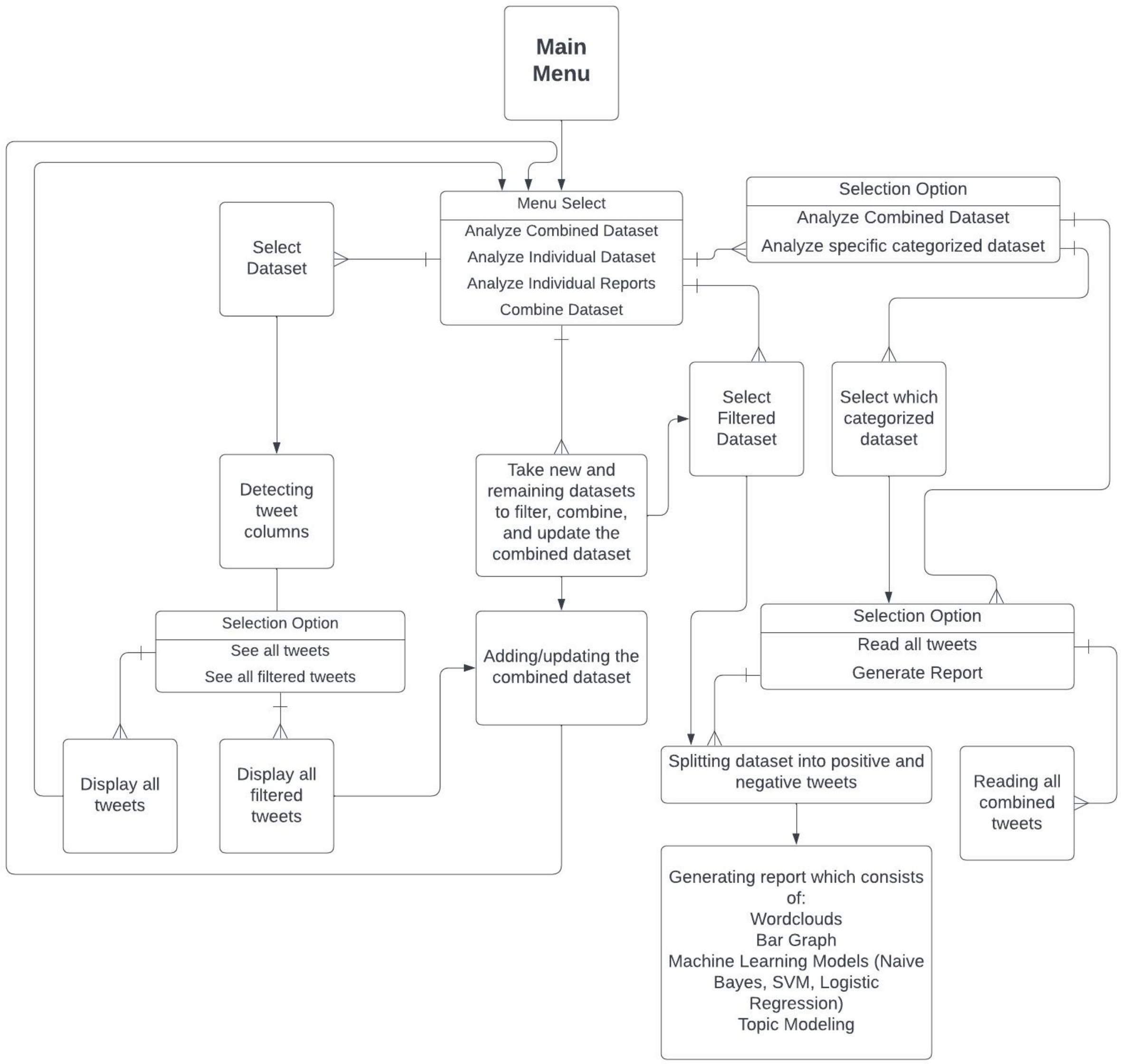
Our user-friendly interface prioritizes ease of use, offering a range of options for dataset analysis. Users can effortlessly choose datasets for analysis, opt for raw or filtered datasets, and witness the automatic combination of filtered datasets. The system's intelligent design prompts users when encountering an unanalyzed dataset, providing the choice to include it in the combined dataset, enhancing the completeness of the analytical landscape.

Dynamic Report Generation

The user interface extends beyond mere analysis, incorporating dynamic report generation capabilities. Users can generate comprehensive reports on the combined dataset, individual categorized datasets, or even individual datasets. These reports encompass machine learning insights, visualizations, and topic modeling outcomes, providing a broad overview of the sentiments encapsulated within the data.

Seamless Integration of Machine Learning and Visualizations

The software engineering underpinning the user interface facilitates seamless integration with machine learning models and visualization tools. Users can effortlessly navigate through machine learning reports, interpret visualizations, and dig into the established landscape of sentiment analysis. This integration not only streamlines the analytical process but also empowers users to derive meaningful insights with minimal complexity.



This diagram shown above incorporates a user interface diagram, illustrating the seamless flow of directions within the software. This diagram serves as a visual guide, providing a clear representation of the user interface's functionality and enhancing the overall user experience throughout the sentiment analysis process.

Chapter 3: Results

3.1 Filtering Raw Tweets

The initial step involved filtering raw tweets from a mixed-content dataset to identify users expressing sentiments signifying of mental health issues. Through sentiment analysis and categorization, tweets were classified into various depressive struggles. Examples of filtered tweets from everyday life include:

```
--------------------------------------------------
Tweet 528224:
i hate my life and want to die

Sentiment Score: -0.80
Classified Categories: Suicidal Thoughts and Self-Harm
--------------------------------------------------
Tweet 457600:
just kinda upsets me that everyones parents still blame me for trying to kill myself when i was fucking 11 tbh!! we were all
hurting and fucking babies practically and somehow that was me? it was my fault ?

Sentiment Score: -0.77
Classified Categories: Suicidal Thoughts and Self-Harm
--------------------------------------------------
Tweet 470897:
men would rather be in a place without service so as to not receive my text about a crazy cooking video i saw than go to the
rapy!

Sentiment Score: -0.75
Classified Categories: Treatment and Coping
--------------------------------------------------
Tweet 378873:
no bc i’m eating skittles rn and that requires every inch of my mouth muscles…cant believe justin is struggling like this!!!
like i’m sick to my stomach

Sentiment Score: -0.71
Classified Categories: Emotional State and Struggles
--------------------------------------------------
Tweet 23319:
hungry but loss of appetite and sick 👎

Sentiment Score: -0.71
Classified Categories: Physical and Behavioral Symptoms
--------------------------------------------------
Tweet 217536:
i think i need anger management therapy

Sentiment Score: -0.70
Classified Categories: Treatment and Coping
--------------------------------------------------
```

The subsequent analysis expanded beyond a specific theme, focusing on an overall dataset comprising tweets from everyday life. This broader scope allowed for a comprehensive exploration of sentiments related to mental health struggles in various contexts. Users expressed a spectrum of emotions, reflecting everyday challenges and concerns. This diverse dataset provided a diverse understanding of the intersections between mental health and daily life, laying the foundation for a comprehensive and deep examination.

Tweets from users with predicted mental health concerns have been isolated for further examination. This subset of data will undergo comprehensive machine learning and in-depth analysis to extract distinct patterns and identify potential correlations between mental health expressions and specific discourse elements across diverse topics presented in everyday conversations.

3.2 Machine Learning Outputs

Word Clouds

To gain a visual understanding of the prevalent themes within the filtered tweets covering various everyday life topics, word cloud analysis was conducted on both positive and negative sentiment categories. Word clouds provide an intuitive representation of the most frequently occurring words, offering insights into the key topics and sentiments expressed by users.

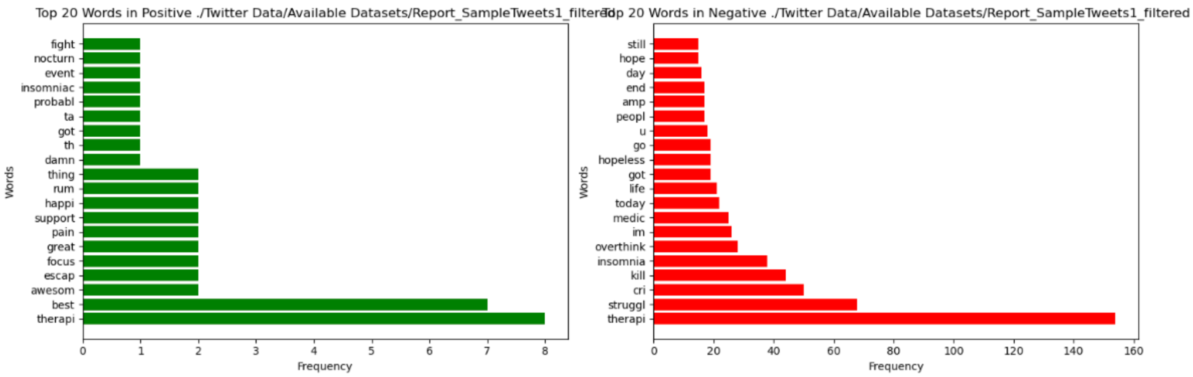
In the subsequent sections, we present word clouds generated from positive and negative sentiment tweets across mixed content. These visualizations unveil the prominent words associated with each sentiment category, exposing the prevailing discourse on social media. By identifying the significant terms, we aim to capture the essence of user sentiments and pinpoint specific language patterns that contribute to the overall sentiment landscape.

The word cloud analysis revealed distinct patterns in the language used by users across mixed content in expressing sentiments. Common themes in negative sentiment tweets included words like "struggle" and "cry," emphasizing emotional challenges. Conversely, positive sentiment tweets featured uplifting terms such as "love," "happy," "good," and "better." Notably, the negative word cloud contained words like "bad," "scream," and "worst," along with various swear words, indicating a more intense and emotionally charged language. These findings underscore the diverse emotional expressions within the online discourse, reflecting a spectrum of sentiments from struggles and challenges to positive and uplifting experiences.

Horizontal Bar Graphs

In conjunction with the word cloud analysis, horizontal bar charts were employed to systematically present impactful words prevalent in the mixed content of tweets. Unlike word clouds, which provide an intuitive but unstructured view, horizontal bar charts offer a more organized display of the relative frequencies of significant terms associated with both positive and negative sentiments.

The subsequent sections unveil horizontal bar charts corresponding to positive and negative sentiments found in the mixed content of tweets. These charts aim to highlight specific words that carry considerable weight in influencing sentiment across various topics within the online discourse. By structuring the data in this way, the horizontal bar charts contribute to a broad understanding of the language dynamics within the mixed content, providing valuable perception into prevalent themes and sentiments expressed by users.



This analysis of horizontal bar charts reveals notable patterns within the mixed content of tweets related to mental health and climate change. In the negative sentiment category, impactful words such as "kill," "hopeless," and "cry" stand out, indicating the presence of distressing language. Conversely, the positive sentiment category features words like "best," "awesome," "great," "happi," and "support," suggesting a more optimistic tone within certain discussions. Additionally, common terms like "therapi" and "insomnia" appear across both sentiments, emphasizing the interconnectedness of mental health and climate change discussions. These findings provide a structured overview of the prevailing language dynamics, offering comprehension into the sentiments expressed by users in the context of mixed-content tweets.

Machine Learning Models

Machine learning models, including Logistic Regression, Naive Bayes, and SVM, were utilized for sentiment classification. Performance assessments, such as accuracy, precision, recall, F1 score, and True/False Positive rates, were measured to evaluate the effectiveness of the models.

```
Naive Bayes Accuracy: 0.72
Naive Bayes Precision: 0.88
Naive Bayes Recall: 0.21
Naive Bayes F1-Score: 0.34
Naive Bayes Confusion Matrix:
TN: 411, FP: 6, FN: 167, TP: 44
Naive Bayes True Positive Rate (TPR): 0.21
Naive Bayes False Positive Rate (FPR): 0.01
Naive Bayes True Negative Rate (TNR): 0.99
Naive Bayes False Negative Rate (FNR): 0.79

SVM Accuracy: 0.79
SVM Precision: 0.88
SVM Recall: 0.45
SVM F1-Score: 0.60
SVM Confusion Matrix:
TN: 404, FP: 13, FN: 116, TP: 95
SVM True Positive Rate (TPR): 0.45
SVM False Positive Rate (FPR): 0.03
SVM True Negative Rate (TNR): 0.97
SVM False Negative Rate (FNR): 0.55

Logistic Regression Accuracy: 0.79
Logistic Regression Precision: 0.84
Logistic Regression Recall: 0.46
Logistic Regression F1-Score: 0.60
Logistic Regression Confusion Matrix:
TN: 399, FP: 18, FN: 113, TP: 98
Logistic Regression True Positive Rate (TPR): 0.46
Logistic Regression False Positive Rate (FPR): 0.04
Logistic Regression True Negative Rate (TNR): 0.96
Logistic Regression False Negative Rate (FNR): 0.54
```

As shown in the example above, the machine learning models were trained and evaluated on the same mixed-content dataset, yielding distinct performance metrics. Naive Bayes exhibited a precision of 88%, indicating a high accuracy of positive predictions, but struggled with recall at 21%, leading to a lower F1-Score of 34%. SVM and Logistic Regression both achieved an accuracy of 79%, with SVM demonstrating a higher recall at 45%, resulting in a superior F1-Score of 60%. Logistic Regression displayed balanced precision and recall at 84% and 46%, respectively. The confusion matrices provided detailed insights into the models' true positive, true negative, false positive, and false negative predictions. SVM stood out with a balanced trade-off between precision and recall, showcasing its effectiveness in classifying tweets with potential mental health concerns.

Topic Modeling

Topic modeling using Latent Dirichlet Allocation (LDA) was employed to uncover latent topics within the tweets. The identified topics shed light on overarching themes and trends, providing deeper insights into the reasons behind mental health issues.

```
Topics from Latent Dirichlet Allocation (LDA):
Topic #1:
cri im year tire hopeless peopl love amp got medic
Topic #2:
cri im fuck therapi liter hard struggl go realli bro
Topic #3:
therapi cri struggl day fuck got today peopl kill hope
Topic #4:
insomnia cri struggl therapi kill im video still go peopl
Topic #5:
cri scream throw mental ill overthink shake therapi today struggl

Detailed topics and sources from Latent Dirichlet Allocation (LDA):
Topic #1:
cri im fuck rn therapi start scream look today day
Topic #2:
cri throw scream insomnia im therapi shake struggl sob peopl
Topic #3:
therapi cri overthink struggl got tire day go hopeless right
Topic #4:
cri ill therapi im happi liter love struggl someon peopl
Topic #5:
therapi cri struggl peopl ill actual im gon na watch
```

The Latent Dirichlet Allocation analysis of mixed-content tweets has revealed five distinctive topics capturing a spectrum of emotions and experiences. Topic #1 encompasses sentiments of weariness, hopelessness, and expressions of love, highlighting the emotional exhaustion users convey. In Topic #2, users' express frustration, emphasizing the challenges of therapy and the tangible struggles they face. Topic #3 dives into daily struggles, incorporating a mix of emotions and mentions of therapy and hope. Topic #4 revolves around the theme of insomnia, combining mentions of crying, struggle, and therapeutic efforts. Finally, Topic #5 encapsulates intense emotional reactions, featuring elements of screaming, throwing, and the ongoing struggle with mental health. This comprehensive analysis provides insights into the diverse emotional outcomes embedded within the mixed-content tweets, reflecting the multifaceted nature of users' mental health expressions on social media.

3.3 Challenges Encountered

The successful execution of this project is not without its share of challenges. This section exposes the obstacles encountered during various stages of the project and the strategies employed to overcome them.

Access to Live Tweets

One notable challenge surfaced in the endeavor to access live tweets directly from Twitter API. The primary hindrance revolved around the associated costs, as acquiring real-time data from Twitter demands a significant and expensive financial commitment. The voluminous nature of tweets generated every second, coupled with the accompanying API usage fees, led to the decision to pivot towards utilizing pre-existing datasets from Kaggle. This adjustment, while addressing the challenge, brought about a shift in the data collection approach.

	Free	Basic	Pro	Enterprise
Getting access	Get Started	Get Started	Get Started	Get Started
Price	Free	$100/month	$5000/month	
Access to Twitter API v2	✓ (Only Tweet creation)	✓	✓	
Access to standard v1.1	✓ (Only Media Upload and Login With Twitter)	✓ (Only Media Upload and Login With Twitter)	✓(Only Media, Help, Rate Limit, and Login with Twitter)	
Project limits	1 Project	1 Project	1 Project	
App limits	1 App per Project	2 Apps per Project	3 Apps per Project	
Tweet caps - Post	1,500	3,000	300,000	
Tweet caps - Pull	✗	10,000	1,000,000	
Filteres stream API	✗	✗	✓	
Access to full-archive search	✗	✗	✓	
Access to Ads API	✓	✓	✓	

Directory Management and Permission Limitations

Efficient directory management within the Windows operating system presented a unique set of challenges. Despite efforts to streamline the organization of folders and datasets within Jupyter Notebook, permission limitations emerged as a hurdle. Even when operating in administrator mode, certain directory management tasks encountered restrictions, impacting the seamless execution of the project workflow. Mitigating these challenges required meticulous troubleshooting and adjustments to the directory structure to ensure a smooth and uninterrupted analysis process.

Chapter 4: Future Work and Conclusion

4.1 Future work – The use of AI technology

As we venture into the horizon of future possibilities, the integration of AI algorithms emerges as a promising avenue for the enhancement of our analysis process. The potential automation of tasks, such as sentiment classification and tweet categorization, holds the key to expediting the analytical journey. The infusion of AI into our framework signifies an ideal shift towards efficiency, where repetitive tasks are automated, accelerating the pace of analysis.

In the realm of machine learning, the prospect of continuous learning and adaptation approaches large. Machine learning models, under the instructions of evolving algorithms, stand to refine and enhance the accuracy of sentiment analysis over time. This perpetual learning loop not only augments the precision of our analyses but also positions our project at the forefront of adaptive and intelligent data exploration.

Ethics, a guiding beacon in the evolving landscape of AI, will play an important role in shaping the path of mental health analysis. As AI advances, cohesion to ethical guidelines becomes tyrant, ensuring a delicate and considerate approach to sensitive areas within the realm of mental health. The future promises a harmonious coexistence between technological advancement and ethical principles, fostering a responsible and empathetic analytical framework.

The envisioned future strikes a balance between automation and user involvement, heralding an era where meaningful and interpretable analysis takes center stage. By correlating the efficiency gains of automation with the human touch of interpretation, our project aspires to navigate the evolving landscape of AI, ushering in a future where technology amplifies our capacity for insight without compromising the essence of empathetic understanding.

4.2 Conclusion

This research got onto a diverse exploration at the intersection of mental health discussions on social media within the realm of Human Centered Computing. Through the application of sentiment analysis, machine learning, and topic modeling, we navigated the complexities of mixed-content datasets, unraveling the various emotions embedded in user-generated content. The findings not only illuminated the complex relationship between individual mental well-being and broader environmental concerns but also underscored the real-world complexities of working with social media datasets within the domain of Human Centered Computing.

The challenges faced during this research, from access limitations to live Twitter data to tangled directory management, highlighted the importance of considering human-centered factors in technology-driven investigations. Despite these hurdles, the project stands as a testament to the potential of leveraging technology to uncover meaningful insights into the collective psyche of online communities, guided by the principles of Human Centered

Computing.

As we charted the emotional landscapes within tweets, filtered out raw content to isolate mental health-related expressions, and harnessed machine learning to categorize sentiments, the project unveiled an environment of human experiences. The word clouds, horizontal bar charts, and topic modeling endeavored to capture the essence of these experiences, showing both the struggles and positive narratives embedded in the digital conversations, ultimately contributing to the understanding of Human Centered Computing.

In undertaking this project within the framework of Human Centered Computing for the last 4 months alongside Dr. Shastri’s support, I found profound satisfaction in the belief that this work may serve as a valuable resource for researchers, mental health advocates, law enforcements, and policymakers in the future. The fusion of sentiment analysis and climate change discussions, viewed through the lens of Human Centered Computing, provides a unique perspective through which we can comprehend not only the global environmental narrative but also the deeply personal stories woven into it.

As the project concludes, I am left with a sense of accomplishment and the certainty that the insights derived from this work, grounded in the principles of Human Centered Computing, may contribute to the broader dialogue on mental health, the intricate ways they intersect in the digital realm, and better shaping mankind. It is my hope that these findings, enriched by the considerations of Human Centered Computing, promote a deeper understanding of the human experience within the digital landscape, ultimately paving the way for informed, empathetic, and effective interventions in both mental health and environmental advocacy.

References

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020, May 14). *Using topic modeling methods for short-text data: A comparative analysis*. Frontiers. https://www.frontiersin.org/articles/10.3389/frai.2020.00042/full

Clifton, A., Turkheimer, E., & Oltmanns, T. F. (2008, September 4). Personality disorder in social networks: Network position as a marker of interpersonal dysfunction. Social Networks. https://www.sciencedirect.com/science/article/pii/S0378873308000415

Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., & Torii, M. (2019, April 4). Extracting health-related causality from Twitter messages using Natural Language Processing - BMC Medical Informatics and Decision making. SpringerLink. https://link.springer.com/article/10.1186/s12911-019-0785-0

Ellwood, B. (2021, December 31). An analysis of Twitter posts suggests that people with depression show increased rumination on social media overnight. PsyPost. https://www.psypost.org/2021/12/an-analysis-of-twitter-posts-suggests-that-people-with-depression-show-increased-rumination-on-social-media-overnight-62306

Goyal, G. (2023, May 16). Twitter sentiment analysis using Python: Introduction & techniques. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/

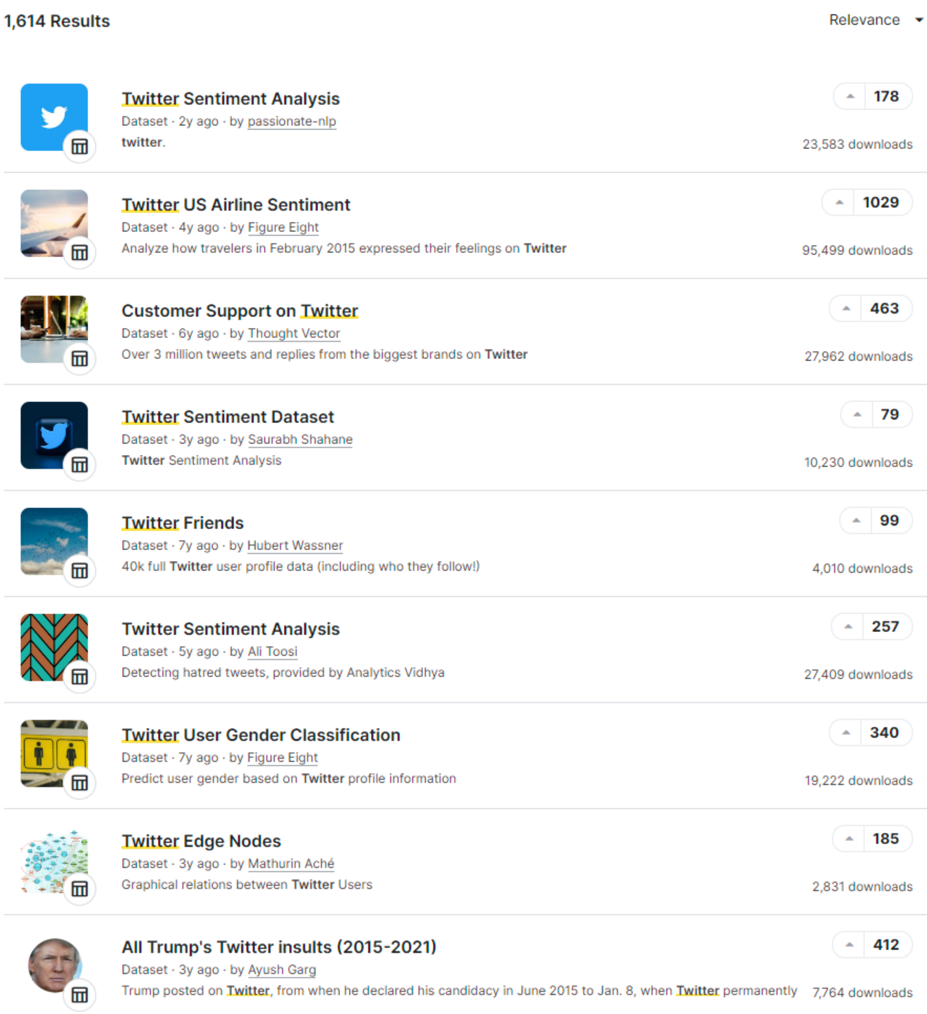
Löwe, B., Arditte, K. A., Goldberg, L. R., Coppersmith, G., Cepoiu, M., Wang, P. S., Seabrook, E. M., Neter, J., Tibshirani, R., Cortes, C., Choudhury, M. D., Radloff, L. S., Beck, A. T., Reece, A. G., Tsugawa, S., Schwartz, H. A., Resnik, P., Preotiuc-Pietro, D., & Pedersen, T. (2017, July 31). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences. https://www.sciencedirect.com/science/article/abs/pii/S2352154617300384

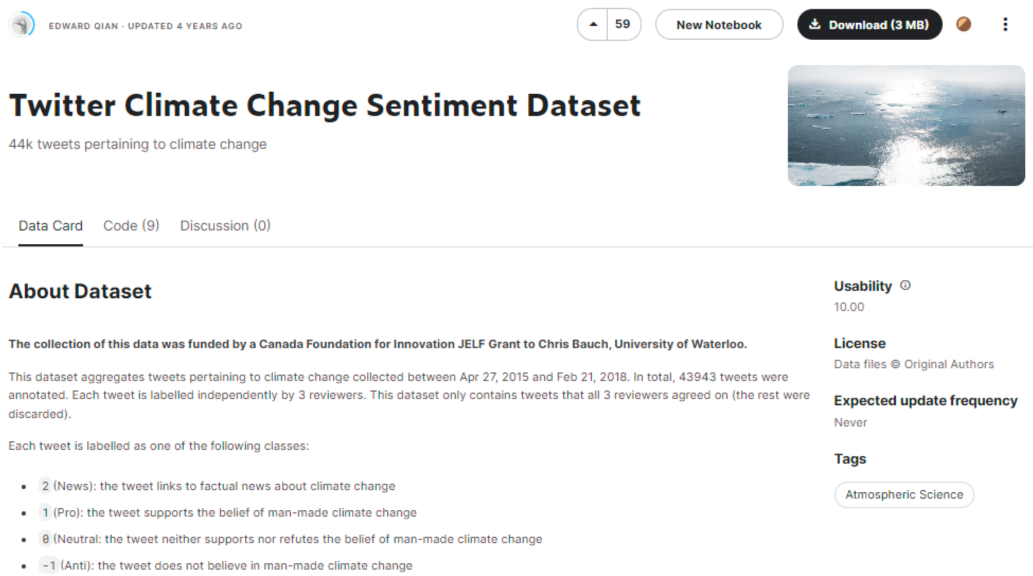
Singh, R., Subramani, S., Du, J., Zhang, Y., Wang, H., Miao, Y., & Ahmed, K. (2023). Antisocial Behavior Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning. EAI Endorsed Transactions on Scalable Information Systems, 10(4), e17. https://doi.org/10.4108/eetsis.v10i3.3184

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022, April 8). Natural language processing applied to Mental Illness Detection: A Narrative Review. Nature News. https://www.nature.com/articles/s41746-022-00589-7

Appendix

Data Collection through Kaggle





EDWARD QIAN · UPDATED 4 YEARS AGO

59 New Notebook Download (3 MB)

Twitter Climate Change Sentiment Dataset

44k tweets pertaining to climate change

Data Card Code (9) Discussion (0)

About Dataset

The collection of this data was funded by a Canada Foundation for Innovation JELF Grant to Chris Bauch, University of Waterloo.

This dataset aggregates tweets pertaining to climate change collected between Apr 27, 2015 and Feb 21, 2018. In total, 43943 tweets were annotated. Each tweet is labelled independently by 3 reviewers. This dataset only contains tweets that all 3 reviewers agreed on (the rest were discarded).

Each tweet is labelled as one of the following classes:

- `2` (News): the tweet links to factual news about climate change
- `1` (Pro): the tweet supports the belief of man-made climate change
- `0` (Neutral: the tweet neither supports nor refutes the belief of man-made climate change
- `-1` (Anti): the tweet does not believe in man-made climate change

Usability
10.00

License
Data files © Original Authors

Expected update frequency
Never

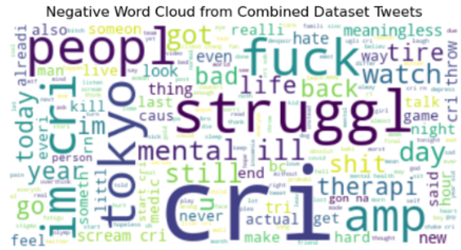
Tags
Atmospheric Science

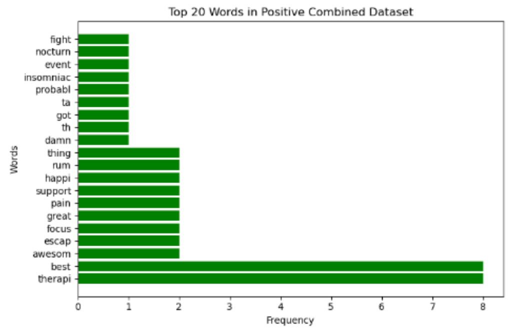
Code Snippets

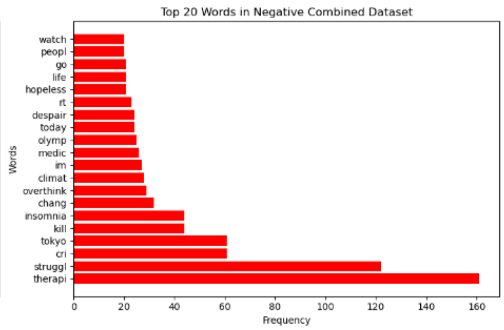
```
#==========================================================================================================================
# Preprocessing
#==========================================================================================================================

# Custom pronouns, function words, and simple verbs as additional stopwords
custom_pronouns = ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself',
                   'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself',
                   'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that',
                   'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
                   'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as',
                   'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through',
                   'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off',
                   'over', 'under', 'again', 'further', 'then', 'once', 'one', 'dont', 'time', 'ive', 'much',
                   'want', 'like', 'get', 'would', 'make', 'take', 'see', 'say', 'think', 'come', 'go', 'know', 'tell',
                   'ask', 'work', 'seem', 'try', 'call', 'need', 'use', 'find', 'give', 'show', 'hear', 'play', 'run',
                   'move', 'help', 'start', 'stop', 'write', 'be', 'become', 'begin', 'feel', 'bring', 'buy', 'put']


# Merge the lists and remove duplicates
combined_stopwords = list(set(custom_pronouns))

# Initialize the Porter Stemmer, WordNet Lemmatizer
porter = PorterStemmer()
lemmatizer = WordNetLemmatizer()

# Set of stopwords
stopwords = set(nltk.corpus.stopwords.words('english'))

# Text cleaning function with stemming, lemmatization, and removal of stopwords
def clean_text(text):
    if combined_stopwords is not None:
        stopwords.update(combined_stopwords)

    text = re.sub(r'https?://\S+|www\.\S+', '', text)  # Remove URLs
    text = re.sub(r'<.*?>', '', text)  # Remove HTML tags
    text = re.sub(r'[^a-zA-Z]', ' ', text)  # Remove non-alphabetic characters
    text = text.lower()  # Convert to lowercase

    # Tokenize into words using nltk's word_tokenize
    words = word_tokenize(text)

    # Remove stopwords, apply stemming, and lemmatization
    processed_words = [lemmatizer.lemmatize(porter.stem(word)) for word in words if word.lower() not in stopwords]

    # Join the processed words back into cleaned text
    cleaned_text = ' '.join(processed_words)

    return cleaned_text
```

```python
def detect_likely_tweet_column(df):
    # Create a dictionary to store the criteria for each likely tweet column
    likely_tweet_columns = {}

    # Iterate through the likely tweet columns
    for column in df.columns:
        print(f"Analyzing column: {column}")
        if not column.startswith("https://"):
            # Calculate the average length of text
            average_length = df[column].apply(lambda x: len(str(x))).mean()

            # Calculate the percentage of non-empty values
            non_empty_percentage = (df[column].apply(lambda x: isinstance(x, str) and len(str(x).strip()) > 0).sum() / len(df

            # Calculate the percentage of unique values
            unique_percentage = (len(set(df[column])) / len(df)) * 100

            likely_tweet_columns[column] = {
                'average_length': average_length,
                'non_empty_percentage': non_empty_percentage,
                'unique_percentage': unique_percentage
            }

    # Determine the most likely tweet column based on criteria
    if likely_tweet_columns:
        weights = {
            'average_length': 1,
            'non_empty_percentage': 2,
            'unique_percentage': 3
        }

        # Calculate a score for each column based on the weighted criteria
        for column, criteria in likely_tweet_columns.items():
            score = sum(weights[criterion] * criteria[criterion] for criterion in criteria)
            likely_tweet_columns[column]['score'] = score

        # Find the column with the highest score
        most_likely_tweet_column = max(likely_tweet_columns, key=lambda x: likely_tweet_columns[x]['score'])

        # Return the most likely tweet column
        return most_likely_tweet_column

    else:
        return None
#==============================================================================================================================
#  Topic Modeling
#==============================================================================================================================

def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic #{topic_idx + 1}:")
        print(" ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]]))

def perform_topic_modeling(df, num_topics):
    texts = [word_tokenize(text) for text in df['Cleaned_Text']]

    # Vectorize the text data using TF-IDF
    vectorizer = TfidfVectorizer()
    X_tfidf = vectorizer.fit_transform(df['Cleaned_Text'])

    # Train the LDA model
    lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)
    lda_model.fit(X_tfidf)

    # Print the topics
    print("Topics from Latent Dirichlet Allocation (LDA):")
    if hasattr(vectorizer, 'get_feature_names_out'):
        feature_names = vectorizer.get_feature_names_out()
    else:
        feature_names = vectorizer.get_feature_names()
    display_topics(lda_model, feature_names, 10)

def perform_topic_modeling_trends(df, num_topics, depressive_words=None):
    # Tokenize the text data
    texts = [word_tokenize(text) for text in df['Cleaned_Text']]

    depressive_stop_words = define_depressive_stop_words()

    # Define the vectorizer with optional depressive words as stopwords
    vectorizer = TfidfVectorizer(stop_words=depressive_stop_words)

    # Vectorize the text data using TF-IDF
    X_tfidf = vectorizer.fit_transform(df['Cleaned_Text'])

    # Train the LDA model
    lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)
    lda_model.fit(X_tfidf)

    # Print the topics
    print("Detailed topics and sources from Latent Dirichlet Allocation (LDA):")
    if hasattr(vectorizer, 'get_feature_names_out'):
        feature_names = vectorizer.get_feature_names_out()
    else:
        feature_names = vectorizer.get_feature_names()
    display_topics(lda_model, feature_names, 10)
```

Topic Modeling in additional datasets








```
Naive Bayes Accuracy: 0.74
Naive Bayes Precision: 0.86
Naive Bayes Recall: 0.28
Naive Bayes F1-Score: 0.42
Naive Bayes Confusion Matrix:
TN: 470, FP: 11, FN: 178, TP: 68
Naive Bayes True Positive Rate (TPR): 0.28
Naive Bayes False Positive Rate (FPR): 0.02
Naive Bayes True Negative Rate (TNR): 0.98
Naive Bayes False Negative Rate (FNR): 0.72

SVM Accuracy: 0.80
SVM Precision: 0.89
SVM Recall: 0.45
SVM F1-Score: 0.60
SVM Confusion Matrix:
TN: 467, FP: 14, FN: 135, TP: 111
SVM True Positive Rate (TPR): 0.45
SVM False Positive Rate (FPR): 0.03
SVM True Negative Rate (TNR): 0.97
SVM False Negative Rate (FNR): 0.55

Logistic Regression Accuracy: 0.80
Logistic Regression Precision: 0.87
Logistic Regression Recall: 0.49
Logistic Regression F1-Score: 0.63
Logistic Regression Confusion Matrix:
TN: 463, FP: 18, FN: 125, TP: 121
Logistic Regression True Positive Rate (TPR): 0.49
Logistic Regression False Positive Rate (FPR): 0.04
Logistic Regression True Negative Rate (TNR): 0.96
Logistic Regression False Negative Rate (FNR): 0.51

Topics from Latent Dirichlet Allocation (LDA):
Topic #1:
cri tokyo im olymp fuck gold kill hidilyn liter scream
Topic #2:
cri got therapi die peopl rock amp struggl without bottom
Topic #3:
therapi cri throw mental struggl overthink ill insomnia scream im
Topic #4:
cri therapi struggl got team tokyo amp today love mental
Topic #5:
cri struggl realli day therapi tire still oh got love

Detailed topics and sources from Latent Dirichlet Allocation (LDA):
Topic #1:
cri struggl watch tokyo therapi peopl love day back rn
Topic #2:
cri therapi struggl insomnia tire still right hopeless bro got
Topic #3:
cri im scream therapi throw overthink struggl look shake actual
Topic #4:
cri therapi im fuck struggl insomnia cute go video got
Topic #5:
cri tokyo olymp gold ill hidilyn today therapi struggl diaz
```