UNVEILING SENTIMENT: A COMPREHENSIVE API-DRIVEN WORKFLOW FOR SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA

Ghanath V [1], Kholoud Al Nazzawi [2,] and Kasonde Chewe [3]

[1]Drexel University, Philadelphia, PA

DSCI 511: Data Acquisition and Pre-Processing

Instructor: Dr. Alejandro Erick Trofimoff

06/17/2023

Abstract [2]

This study investigates the impact of social media on mental health, specifically focusing on stress analysis, PTSD, anxiety and several other conditions. Datasets from diverse sources, including Kaggle and Reddit, were collected to examine the relationship between social media posts and depression levels. The study employs semantic analysis and develops a computational model for text analysis. The findings reveal insights into the potential risk factors associated with social media usage and their implications for mental health. The models employed in the analysis achieve varying prediction accuracies across different label classes, ranging from 20% to 78%. This research contributes to the understanding of mental health in the context of social media and provides valuable insights for researchers, psychotherapists, and social media users.

Keywords: social media, sentiment analysis, mental health, depression, computational model

In the digital era, the pervasive influence of social media on individuals' lives has prompted extensive research into its impact on mental health. Specifically, stress analysis in relation to social media usage has emerged as a critical area of investigation. This study aims to explore the complex relationship between social media usage and mental health, with a particular focus on understanding the potential risk factors associated with depression. The datasets utilized for this research were sourced from reputable platforms such as Kaggle, Reddit, and various other social media outlets through APIs.

Numerous studies have shed light on the detrimental effects of social media on mental well-being, particularly among young individuals. It has been found that excessive social media usage can contribute to a higher prevalence of mental health issues, including depression, anxiety, and post-traumatic stress disorder (PTSD) (Smith et al., 2018; Twenge et al., 2020). Moreover, cyberbullying, which is prevalent on social media platforms, has been identified as a significant risk factor for mental health problems, particularly among adolescents (Kowalski et al., 2014).

While the impact of social media on young people's mental health has received substantial attention, it is crucial to recognize that adults are also susceptible to its negative consequences. Research has indicated that excessive social media usage among adults can lead to feelings of loneliness, decreased self-esteem, and heightened depressive symptoms (Primack et al., 2017; Verduyn et al., 2017). Additionally, the comparison and self-presentation aspects of social media can contribute to stress and anxiety among adult users (van den Eijnden et al., 2016).

Understanding the complex dynamics between social media usage and mental health is of utmost importance from both academic and psychological perspectives. This study aims to contribute to the existing literature by conducting an in-depth analysis of the relationship between social media messages and depression levels. By examining the linguistic patterns and expressions commonly found in social media posts related to mental health, this research endeavors to enhance our understanding of the intricate interplay between social media usage and mental health outcomes.

Moreover, the development of computational models for text analysis holds significant promise in facilitating mental health research in the context of social media. These models can equip researchers and psychotherapists with powerful tools to identify, monitor, and address mental health concerns in a timely and effective manner. The ultimate goal is to provide actionable insights and recommendations for promoting mental well-being in the digital age.

In this report, we present a comprehensive analysis of stress in social media and its implications for mental health, drawing upon robust datasets obtained from various sources. We aim also to clearly present our data pipeline through all steps of acquisition, pre-processing, enrichment and analysis. The findings from this research have the potential to inform interventions, policies, and support systems to mitigate the negative impact of social media on mental health.

Potential Users and Applications [1, 2, 3]

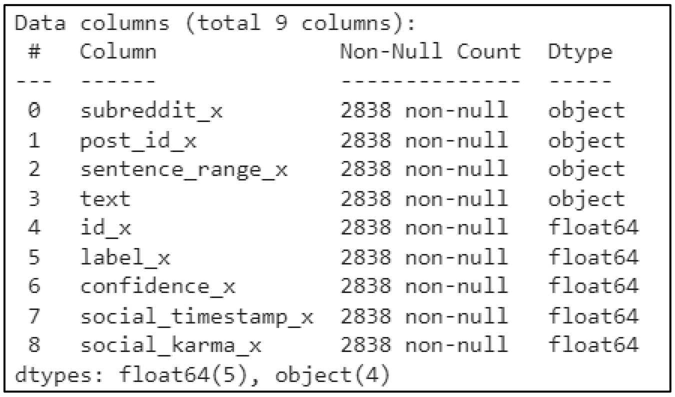
Our study has both academic and practical commercial benefit including tools that may help users quickly diagnose underlying mental health issues before seeking professional health. These computational models can be thought of as a pre-screening tool and not a replacement for trained professionals because actual psychiatric analysis requires multi-modal factors that have complex interactions. In this regard potential users maybe psychologists, neural science students and academics, social media companies and general social media users.

Team Member and Contributions [1, 2, 3]

Team Members	Contributions
Kholoud Al Nazzawi (ka974@drexel.edu) - MS Data Science (1st year), Programming background includes C++, Python, PHP, HTML, and CSS. For this team project, I aspire to utilize my expertise to make a valuable contribution to the project while also helping with project management.	My main contribution was in the field of data cleaning, where I focused on identifying and removing unnecessary data from the dataset. Through a systematic analysis, I determined variables that were not relevant, and eliminated duplicate records. This process resulted in a streamlined and more accurate dataset, enhancing the quality and effectiveness of subsequent analyses and modeling.
Ghanath V (gv374@drexel.edu) – MS Cyber Security (1st year), Programming background includes C, Java, Python and MATLAB. For this team project, I hope to contribute my skills in Project Management and also improve my knowledge in Data Mining concepts.	Made significant contributions to the project by handling data cleaning tasks such as standardization and visualization. I ensured the data was consistent and compatible by applying standardization techniques and addressed missing values. Additionally, I created informative visualizations using Python libraries, enabling better understanding of the dataset. I also played a key role in documenting our data cleaning processes, ensuring clarity and transparency for future reference.
Kasonde Chewe (kc3745@drexel.edu) – MS Bioinformatics (1st year), my programming background includes Python, MATLAB and C/C++. For this team project, I am hoping to contribute skills using packages such as NumPy, Pandas, Matplotlib, data processing, normalization and aid in machine learning and neural network implementation using keras or TensorFlow using pre-existing models to simplify project.	I assisted with preprocessing stages including data cleaning and merging steps. My main contributions were creating figures and illustrations, peer review and model selection, evaluation and optimization in the post data acquisition analysis stages. I also aided in managing project structure dependencies and directories and command line handling.

Motivation for Dataset Selection and Importance of Study:

The selection of this data set is related to the research subject of measuring various mental health indicators from social media posts. Stress is a critical factor in mental health and is crucial to identifying risk factors for depression. The dataset is a valuable resource and supplies an in-depth understanding of the patterns, themes, and languages used in social media posts on mental health, thus contributing to a deeper understanding of the topic. Below is a snippet of our data dictionary, attached in the file data_dictionary.csv. Our dataset is 199.7 MBs in size. The RangeIndex includes 2838 entries, from 0 to 2837 with 9 columns.



```
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   subreddit_x         2838 non-null   object
 1   post_id_x           2838 non-null   object
 2   sentence_range_x    2838 non-null   object
 3   text                2838 non-null   object
 4   id_x                2838 non-null   float64
 5   label_x             2838 non-null   float64
 6   confidence_x        2838 non-null   float64
 7   social_timestamp_x  2838 non-null   float64
 8   social_karma_x      2838 non-null   float64
dtypes: float64(5), object(4)
```

Figure 1: Dataset information including number of Column Names, Number of Entries, Non-Null Counts Data Types (Dtypes). Object includes string data types in alphanumeric characters with UTF-8.

The study of social media's impact on mental health is a topic of significant importance. There is ample evidence to suggest that social media can have negative effects on young people, including children and teenagers, which in turn can lead to various mental health issues such as depression, anxiety, and post-traumatic stress disorder (PTSD), especially due to cyberbullying. While not as well studied, the negative impact on adults is also a cause for concern.

Therefore, the primary aim of this research is to explore the connection between mental health diseases and social media usage. Through an in-depth analysis of this relationship, the study aims to enhance the academic and psychological comprehension of this subject and offer valuable knowledge for the public that uses social media. Furthermore, the study seeks to create computational models for text analysis that can be employed by researchers and psychotherapists. The primary goals of this research project include finding answers to the following questions:

- Does excessive social media contribute to depression?
- What is the definition of excessive use of social media?
- What are the keywords and phrases that people use when faced with mental health issues?
- Does social media usage impact young people more than older people?
- Are certain platforms more harmful to mental health than others?
- What are the consequences of mental health issues related to social media usage?
- Can we produce recommendations based on our findings to help tackle this challenge?

Gantt chart -- workflow

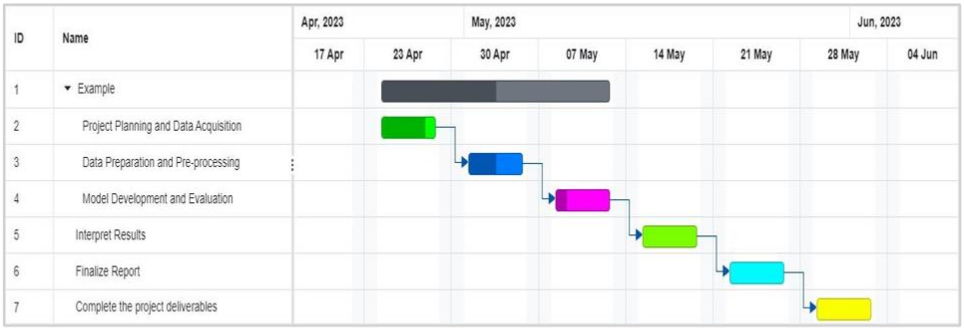


Figure 2: Gnatt Chart highlighting proposed timeline for final project submission. Note that times in yellow extend beyond June 04 to account for holidays and academic examinations.

Table 1: Workflow Protraction and Tasks for Processing and Analysis

Week 4	• Define research question and goals. • Identify data sources and acquire data from Reddit API. • Perform initial data exploration and cleaning. • Set up collaboration tools (e.g., GitHub, Trello).
Week 5	• Conduct detailed data cleaning, pre-processing, and feature engineering. • Perform exploratory data analysis (EDA) to gain insights from the data. • Address any issues with missing data, duplicate data, and data quality.
Week 6	• Select appropriate machine learning models based on project goals. • Split data into training and testing sets. • Train and tune machine learning models using Scikit-learn library. • Evaluate models based on performance metrics such as accuracy, precision, recall, and F1-score.
Week 7	• Interpret and present the results of the machine learning models. • Visualize key findings using Matplotlib and Seaborn libraries. • Draw conclusions and insights from the data analysis. • Prepare for the final report and presentation.

Week 8	• Finalize the report and presentation. • Review and revise the project deliverables. • Prepare for team presentation and feedback session.
Week 9	• Complete the project deliverables. • Present the project findings to the team and stakeholders. • Address any questions or feedback from the team. • Submit the final project report and presentation.

Project Processing Chart

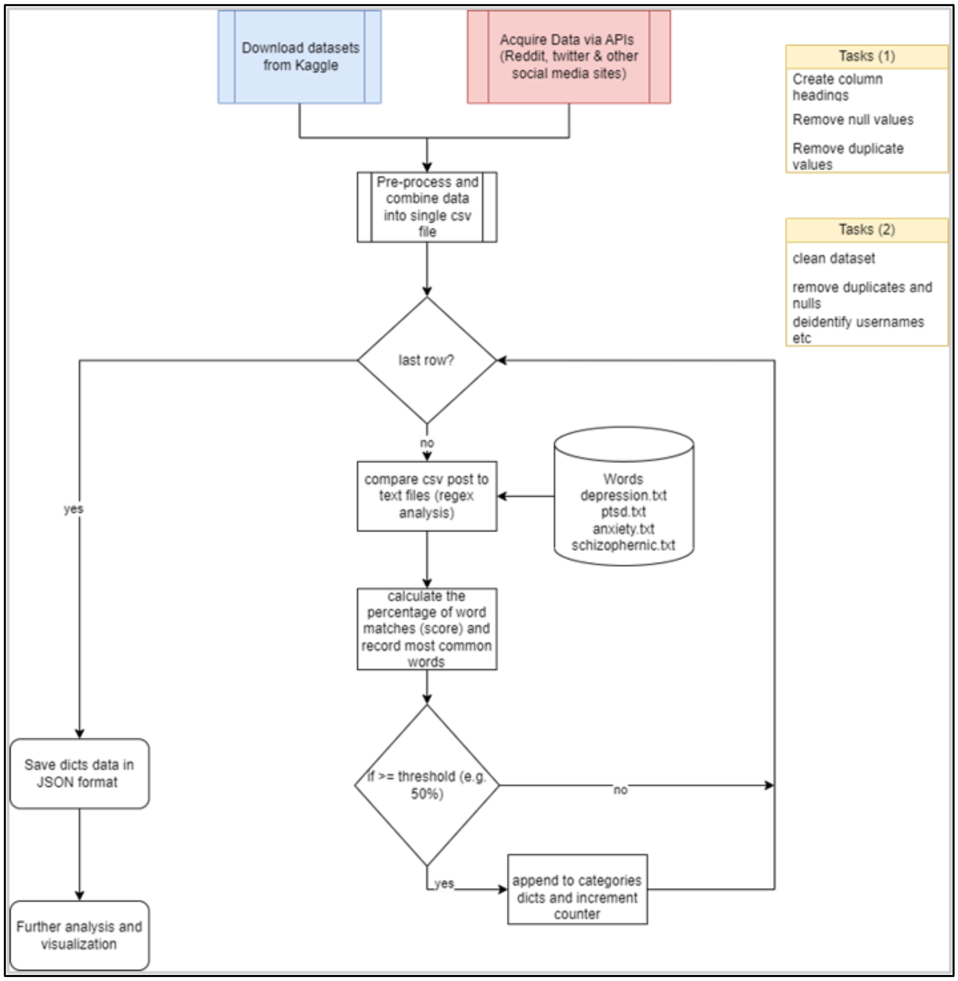

Download datasets from Kaggle
Acquire Data via APIs (Reddit, twitter & other social media sites)
Tasks (1)
Create column headings
Remove null values
Remove duplicate values
Pre-process and combine data into single csv file
Tasks (2)
clean dataset
remove duplicates and nulls
deidentify usernames etc
last row?
yes
no
compare csv post to text files (regex analysis)
Words
depression.txt
ptsd.txt
anxiety.txt
schizophernic.txt
calculate the percentage of word matches (score) and record most common words
if >= threshold (e.g. 50%)
no
yes
append to categories dicts and increment counter
Save dicts data in JSON format
Further analysis and visualization

Figure 3: This chart illustrates the processing of the project, starting with the initial data cleaning stage, followed by the modeling phase where the data is compared and analyzed to determine if an individual has depression or not. Finally, the processed data is carefully examined and finalized, ensuring accurate and reliable results for identifying depression in individuals.

Acquistion Methods

The data for this project was obtained from two different sources: Reddit and Kaggle. The Reddit data was acquired using the Reddit API, while the Kaggle dataset was downloaded directly from the Kaggle website. The Reddit data provides real-time posts and comments from various subreddits, while the Kaggle dataset contains pre-collected data on posts from different subreddits.

The Reddit data was retrieved by making API requests to gather posts and associated metadata such as subreddit, post ID, text content, label, confidence level, social timestamp, and more. The data collection process involved specifying the desired subreddits and applying filters to retrieve relevant posts. Appendix A1, shows the code snippet from "Group-Project-Sentiment-Anlaysis.ipynb", shows manual retrieval (requests) of reddit data and the resulting JSON aggregated with various mental health tags.

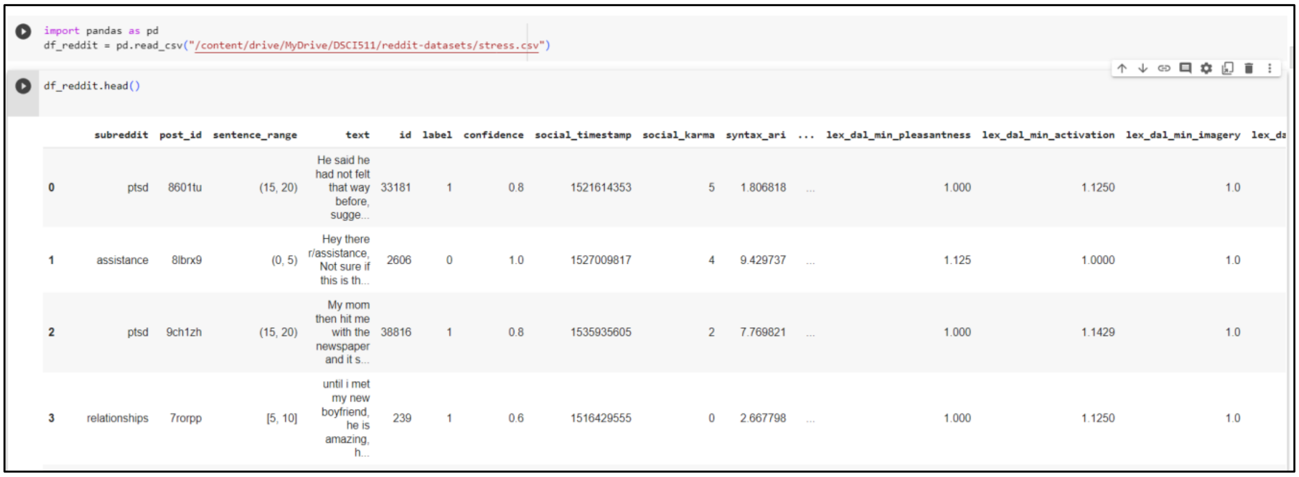
The Kaggle dataset, on the other hand, was downloaded as a CSV file containing multiple columns, including subreddit, post ID, text content, label, confidence level, social timestamp, and various other attributes related to the posts. Combining these two datasets provides a diverse and comprehensive collection of post data from various subreddits, allowing for a broader analysis and exploration of the topics covered.

Reddit

Reddit is a popular social news aggregation, web content rating, and discussion platform. It is composed of numerous communities called "subreddits" where users can post, comment, and interact with others who share similar interests. Each subreddit focuses on a specific topic, ranging from technology and entertainment to health and personal experiences. Users on Reddit can submit posts consisting of text, links, images, or videos to initiate discussions within a particular subreddit. These posts can be upvoted or downvoted by other users based on their quality and relevance. The most popular and engaging posts rise to the top of the subreddit's feed, while less engaging ones remain lower or become less visible.

Reddit provides a vast amount of user-generated content and diverse perspectives on countless topics. The platform is known for fostering active communities where people can seek advice, share personal stories, engage in debates, or simply connect with others who share their interests. In the context of the project, the Reddit dataset used as a source likely contains posts collected from various subreddits. These posts provide valuable insights into different subjects, such as mental health, relationships, anxiety, or any other topic represented by the subreddits included in the dataset. The data set includes information like the subreddit name, post content, associated labels or classifications, timestamps, user engagement metrics (e.g., upvotes, comments), and linguistic features.

Below is a sample of the raw data from Reddit datasets:
Reddit dataset:



```
import pandas as pd
df_reddit = pd.read_csv("/content/drive/MyDrive/DSCI511/reddit-datasets/stress.csv")
```

```
df_reddit.head()
```

	subreddit	post_id	sentence_range	text	id	label	confidence	social_timestamp	social_karma	syntax_ari	...	lex_dal_min_pleasantness	lex_dal_min_activation	lex_dal_min_imagery	lex_da
0	ptsd	8601tu	(15, 20)	He said he had not felt that way before, sugge...	33181	1	0.8	1521614353	5	1.806818	...	1.000	1.1250	1.0	
1	assistance	8lbrx9	(0, 5)	Hey there r/assistance, Not sure if this is th...	2606	0	1.0	1527009817	4	9.429737	...	1.125	1.0000	1.0	
2	ptsd	9ch1zh	(15, 20)	My mom then hit me with the newspaper and it s...	38816	1	0.8	1535935605	2	7.769821	...	1.000	1.1429	1.0	
3	relationships	7rorpp	[5, 10]	until i met my new boyfriend, he is amazing, h...	239	1	0.6	1516429555	0	2.667798	...	1.000	1.1250	1.0	

Figure 4: Sample snippet of Reddit dataset used in project

Here is a breakdown of the columns in the sample data:

- **subreddit:** This column represents the name of the subreddit to which the post belongs. It indicates the specific community or topic under which the post was made.
- **post_id:** This column contains the unique identifier associated with each post. It helps in distinguishing between different posts within the dataset.
- **sentence_range:** This column represents the range of sentences within the post where the specific text is derived from. It provides insights into the location of the extracted text within the original post.
- **text:** This column contains the actual text content of the post. It includes the information, story, question, or any other message shared by the user.
- **label:** This column indicates the label or classification assigned to each post. In the provided sample data, the label seems to represent binary sentiment classification, where 1 might represent a positive sentiment and 0 a negative sentiment.
- **confidence:** This column denotes the confidence level associated with the assigned label. It provides an indication of the certainty or reliability of the assigned sentiment label.
- **social_timestamp:** This column represents the timestamp when the post was made on Reddit. It provides information about the time and date of the post creation.
- **social_karma:** This column contains the social karma or the total number of upvotes received by the social media post. It reflects the post's popularity or engagement level within the Reddit community.
- **syntax_ari:** This column represents the Automated Readability Index (ARI) score associated with the post's syntax. The ARI score is a measure of the readability of the text.
- ... (additional columns): The sample data provided includes several other columns that are not explicitly mentioned. These columns likely contain various linguistic, sentiment-

related, or statistical features extracted from the text, such as lexical diversity, sentiment scores, and social engagement metrics.

This Reddit dataset provides a collection of posts from different subreddits, allowing researchers and data scientists to explore and analyze user-generated content from various topics and communities. The information in this dataset can be utilized for tasks such as sentiment analysis, topic modeling, linguistic analysis, or studying user behavior within specific subreddits.

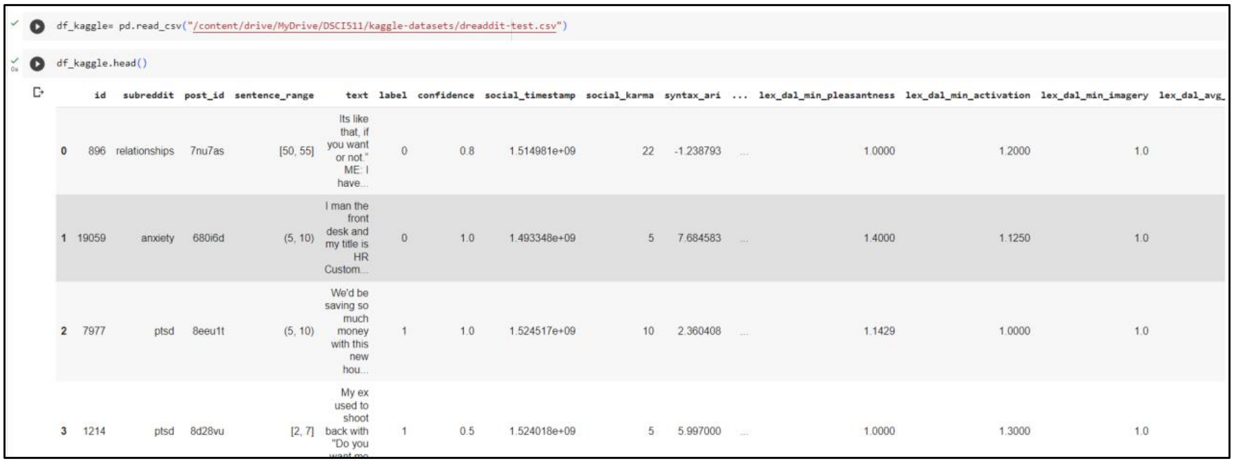
Kaggle

Kaggle, on the other hand, is a popular online community and platform for data science and machine learning enthusiasts. It serves as a hub for datasets, competitions, and collaborative data-driven projects. Kaggle hosts a wide range of datasets contributed by users, researchers, and organizations from various domains. Data scientists and researchers can access and download datasets from Kaggle for exploration, analysis, and modeling purposes. These datasets cover diverse topics, including but not limited to social media, healthcare, finance, climate, and sports. Kaggle datasets often come in structured formats, such as CSV files, which facilitate easy data manipulation and analysis.

In the context of the project, the Kaggle dataset used contains information related to posts or discussions like those found on Reddit. The dataset may consist of columns such as post IDs, subreddits, post content, labels or classifications, timestamps, user engagement metrics, and additional features like sentiment analysis scores, readability measures, or linguistic attributes.

Below is a sample of the raw data from Kaggle datasets:

Kaggle dataset:



```
df_kaggle= pd.read_csv("/content/drive/MyDrive/DSCI511/kaggle-datasets/dreaddit-test.csv")
```

```
df_kaggle.head()
```

	id	subreddit	post_id	sentence_range	text	label	confidence	social_timestamp	social_karma	syntax_ari	...	lex_dal_min_pleasantness	lex_dal_min_activation	lex_dal_min_imagery	lex_dal_avg_
0	896	relationships	7nu7as	[50, 55]	Its like that, if you want or not." ME: I have...	0	0.8	1.514981e+09	22	-1.238793	...	1.0000	1.2000	1.0	
1	19059	anxiety	680i6d	(5, 10)	I man the front desk and my title is HR Custom...	0	1.0	1.493348e+09	5	7.684583	...	1.4000	1.1250	1.0	
2	7977	ptsd	8eeu1t	(5, 10)	We'd be saving so much money with this new hou...	1	1.0	1.524517e+09	10	2.360408	...	1.1429	1.0000	1.0	
3	1214	ptsd	8d28vu	[2, 7]	My ex used to shoot back with "Do you	1	0.5	1.524018e+09	5	5.997000	...	1.0000	1.3000	1.0	

Figure 4: Sample of dataset from Kaggle

Here is a breakdown of the columns in the sample data:

- **id:** This column represents the unique identifier associated with each post or data entry in the dataset.

- **subreddit:** This column indicates the name of the subreddit to which the post belongs. It helps identify the specific community or topic.
- **post_id:** This column contains the unique identifier associated with each post. It distinguishes between different posts within the dataset.
- **sentence_range:** This column represents the range of sentences within the post where the specific text is derived from. It provides insights into the location of the extracted text within the original post.
- **text:** This column contains the actual text content of the post, similar to the Reddit dataset.
- **label:** This column indicates the label or classification assigned to each post, similar to the Reddit dataset. It may represent sentiment, topic, or another categorization.
- **confidence:** This column denotes the confidence level associated with the assigned label, providing an indication of the reliability of the assigned classification.
- **social_timestamp:** This column represents the timestamp when the post was made on Reddit, capturing the time and date information.
- **social_karma:** This column contains the social karma or upvote count received by the post, reflecting its popularity or engagement level within the Reddit community.
- **syntax_ari:** This column represents the Automated Readability Index (ARI) score associated with the syntax of the post's text, indicating its readability.
- ... (additional columns): Similar to the Reddit dataset, there are likely additional columns in the Kaggle dataset that contain linguistic, sentiment-related, or statistical features extracted from the text, offering further insights into the posts and discussions.

The Kaggle dataset provides a structured representation of Reddit posts or discussions, allowing researchers and data scientists to perform various data analysis and modeling tasks. The dataset can be used for tasks such as sentiment analysis, classification, text mining, or building predictive models to gain insights into user behavior, sentiment patterns, or other aspects of online discussions.

Distribution

To ensure that the study and data benefits more researchers and users of social media, after data cleaning, saving the data in a CSV file format that makes it easily accessible and compatible with various data analysis tools and platforms. These platforms include open access on google colabs, and Github where the datasets can be readily and quickly downloaded using an open-source license.

Cleaning {Standardization, Joining, Null values}

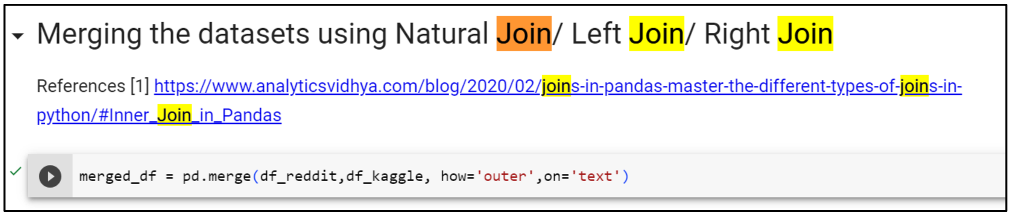
Cleaning the data is an essential step in the data analysis process. It involves preparing the data for further analysis by addressing issues such as missing values, standardizing data formats, and joining different datasets. In this code snippet, the following cleaning steps are performed:

Standardization

Standardization involves transforming data into a common format to ensure consistency and comparability. In the provided code, standardization is not explicitly mentioned, but it can be applied as needed. Standardization could include converting data to a specific data type, converting units, or transforming variables to a common scale. Additional standardization steps that were considered were the size of each sample. When looking at different mental health conditions we considered making each label have an equal number of entries.

Joining Datasets

The code snippet performs dataset merging using an outer join operation. The merging is done based on the 'text' column, which is assumed to be present in both datasets. The merging process combines the data from the Kaggle dataset and the Reddit dataset, aligning the rows based on the matching 'text' values. This allows for a more comprehensive analysis by incorporating data from multiple sources.



Merging the datasets using Natural Join/ Left Join/ Right Join

References [1] https://www.analyticsvidhya.com/blog/2020/02/joins-in-pandas-master-the-different-types-of-joins-in-python/#Inner_Join_in_Pandas

```
merged_df = pd.merge(df_reddit,df_kaggle, how='outer',on='text')
```

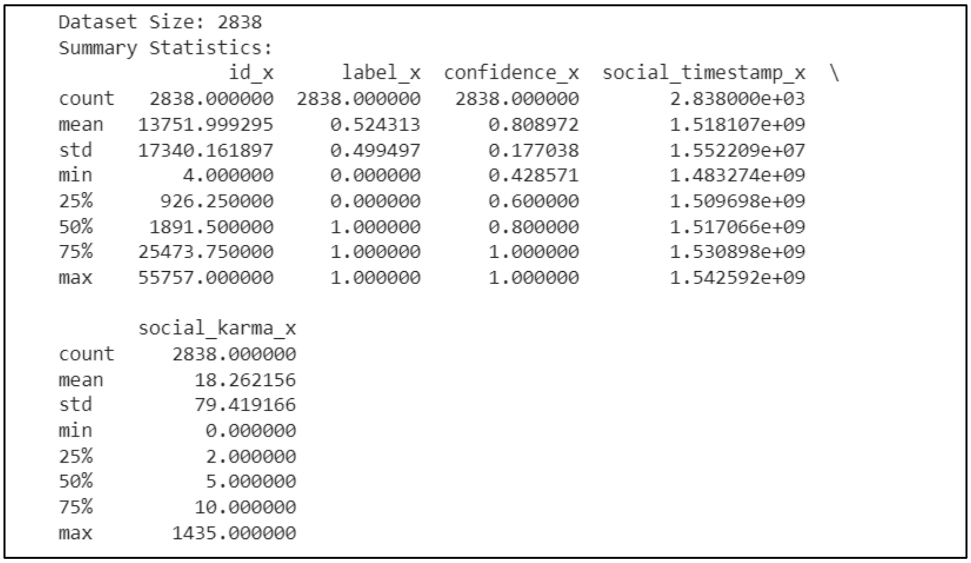
Figure 5: Code snippet of simple outer join operation method in ***panda's*** *module*

Handling Null Values

Null values refer to missing or undefined data in the dataset. In the code snippet, the function drop_null_rows() are used to drop rows that contain null values from the merged data frame. This step ensures that only complete and reliable data is retained for further analysis.

Post Processing Visualization and Statistics

Our resulting merged dataset was reduced from a total of 5000 merged entries down to 2838 entries suggesting that our reddit and Kaggle dataset had similar social media posts. The social media confidence_x field shows high confidence for over 75% of the entries. Additionally, the slow standard deviation (0.177) shows that there is limited variability in the merged datasets training labels. We are therefore sure that our data is of high quality regardless of limited size and can proceed without extensive normalization, although in future projects getting evenly distributed labeling classes is highly recommended for learning semantic features.



```
Dataset Size: 2838
Summary Statistics:
              id_x      label_x  confidence_x  social_timestamp_x  \
count   2838.000000  2838.000000   2838.000000        2.838000e+03
mean   13751.999295     0.524313      0.808972        1.518107e+09
std    17340.161897     0.499497      0.177038        1.552209e+07
min        4.000000     0.000000      0.428571        1.483274e+09
25%      926.250000     0.000000      0.600000        1.509698e+09
50%     1891.500000     1.000000      0.800000        1.517066e+09
75%    25473.750000     1.000000      1.000000        1.530898e+09
max    55757.000000     1.000000      1.000000        1.542592e+09

       social_karma_x
count     2838.000000
mean        18.262156
std         79.419166
min          0.000000
25%          2.000000
50%          5.000000
75%         10.000000
max       1435.000000
```

Figure 6: Merged dataset statistics for numerical data types

Post-Processing Results and Analysis[3]

Semantic analysis refers to the process of extracting meaning from text, enabling computers to understand and interpret sentences, paragraphs, or whole documents. This is achieved by scrutinizing grammatical structure and identifying the relationships between individual words in a specific context (Jurafsky & Martin, 2019). After pre-processing our dataset of social media posts, we proceeded to evaluate several machine learning and natural language processing (NLP) models. Among the most prominent models, such as Naïve Bayes Classifiers (NBC), Support Vector Machines (SVM), Transformers, and Convolutional Neural Networks (CNNs), we observed wide-ranging degrees of accuracy (Aggarwal & Zhai, 2012). This may have been because of limitations in size of our dataset (~ 3000 entries), where accurate training of deep neural networks can require on the order of 10^6 to 10^9 parameters. These and other challenges will be discussed in detail in the next section. Overall, our model aims to identify semantic patterns related to mental health issues in text data extracted from social media posts. In this report, we highlight the results from SVM and TinyBERT, a smaller and more efficient version of the BERT transformer model (Turc et al., 2019).

Support Vector Machine (SVM) Model

SVM is a robust and versatile model with the capacity to generate complex decision boundaries, which can lead to more accurate classifications (Cortes & Vapnik, 1995). For our data, the SVM model achieved an accuracy of 50%. Although this is a significant result, the performance varied widely across the different classes, as shown by the precision, recall, and F1-score metrics in the report below. For instance, while the model achieved an F1-score of 0.65 for class 7, it achieved a score of 0 for class 4. This discrepancy may be due to the imbalance in the dataset. We suggest optimizing our dataset by making sure that each class (mental health) label is equal and large in terms of sample size. For instance, using 100,000 – 200, 000 samples of text for each label category. The results of our SVM model are seen below.

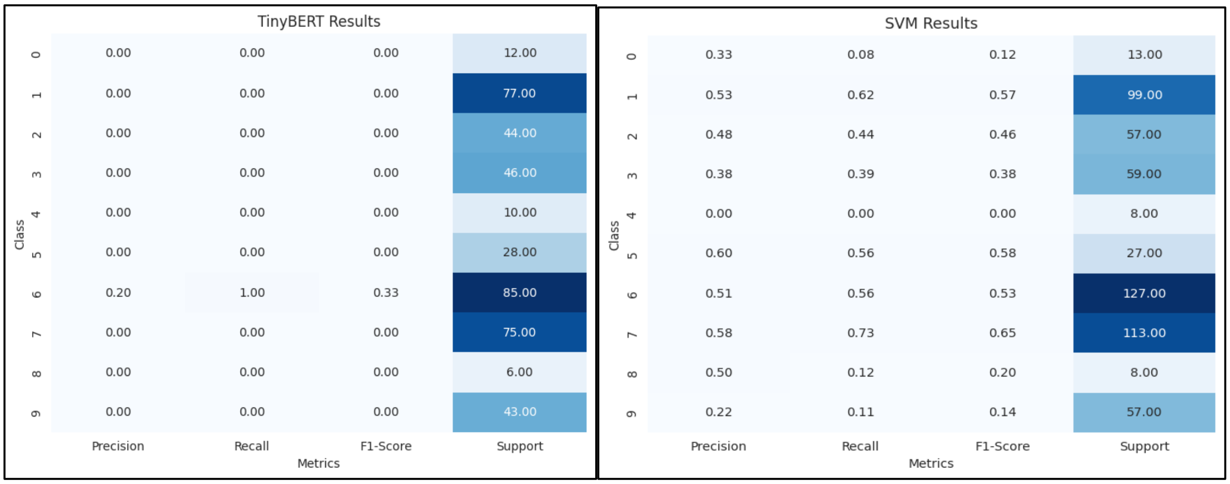


Figure 7) TinyBERT achieved an accuracy of 20% on the validation set, with low precision (ranging from 0% to 20%), recall (ranging from 0% to 20%), and F1-scores (ranging from 0% to 33%) across all classes. These results indicate the difficulty of effectively capturing semantic patterns related to mental health issues in social media text data using TinyBERT. B) SVM achieved a higher accuracy of 50% on the validation set, with varied precision (ranging from 0% to 58%), recall (ranging from 0% to 73%), and F1-scores (ranging from 0% to 65%) for different classes. The SVM model showed better performance in certain classes compared to TinyBERT, but overall, it still faced challenges in accurately predicting sentiment from social media posts.

Bidirectional Encoder Representations from Transformers (BERT)

BERT, and its smaller counterpart TinyBERT, are transformer-based models capable of handling a wide range of NLP tasks (Devlin et al., 2018; Jiao et al., 2020). Unlike SVM, these models consider the full context of each word by looking at the words that come before and after it in a sentence. Despite its sophisticated capabilities, TinyBERT achieved an accuracy of only 20% on our dataset. The model performed particularly poorly on most of the classes, with precision, recall, and F1-score metrics equal to zero for nine out of ten classes. These results suggest that, despite its smaller size and efficiency, TinyBERT might not be the optimal choice for our specific task given the current dataset and model configuration.

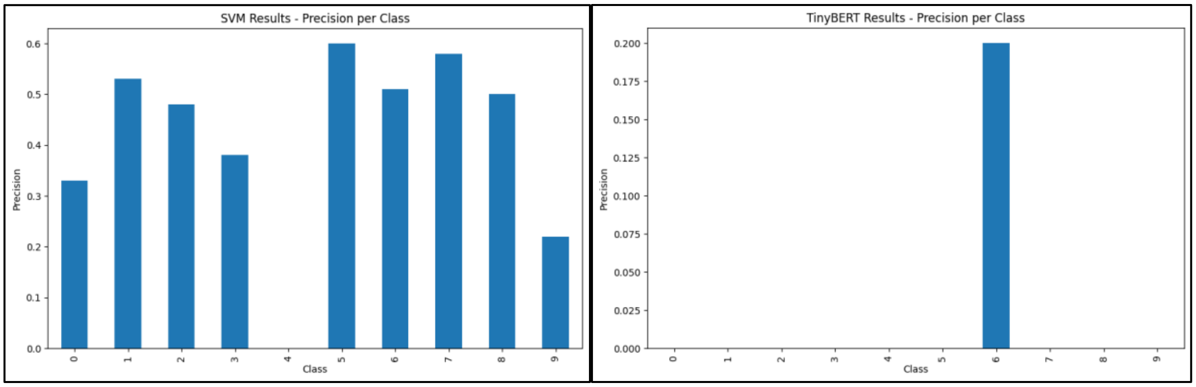


Figure 8: TinyBERT and Support Vector Machine (SVM) statistics from model training. The data here supports the results seen in the above figure showing the poor performance of TinyBert as compared to SVM.

Custom Model Build

To address the challenges posed by limited data and poor performance of deep neural network models, our approach for semantic analysis drew inspiration from the work of Zouaq et al. (2014). They proposed a modular pipeline that incorporated cutting-edge techniques such as state-of-the-art dependency parsing, logical analysis based on a dependency-based grammar, and semantic annotation. Notably, they employed the upper-level ontology SUMO and the WordNet lexicon as standard resources for semantic annotation. By opting for a dependency grammar approach, they capitalized on the advanced capabilities and maturity of existing analyzers in the field. Building upon their framework, we developed a natural language tokenization model using the NLTK package, following a similar workflow.

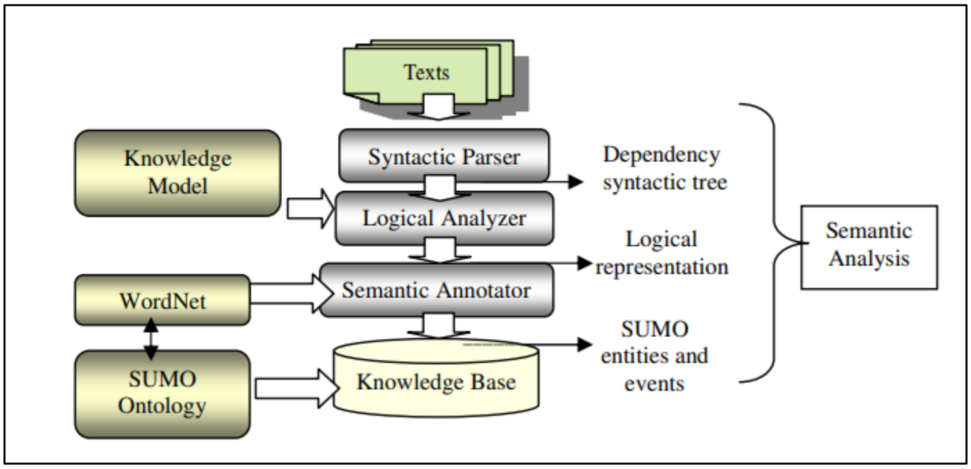
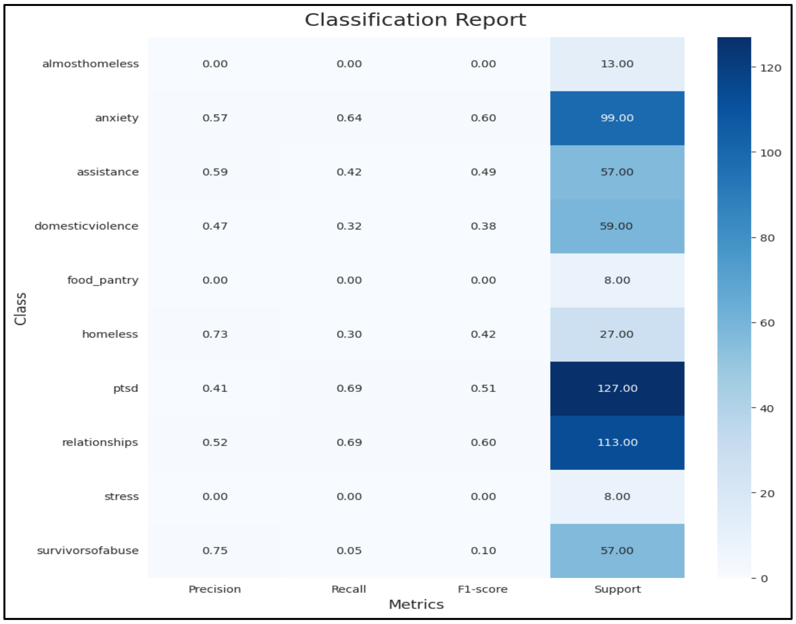


Figure 9: NLTK Sentiment Analysis Workflow modified from Zouaq et al. (2014).

Our custom-built model revealed significant improvement in model prediction accuracy with limited data as across of label classes all grouped especially the text related to ‘survivors of abuse’, from 22% to 75%. Additionally, the custom model showed better results without **Semantic**

Annotation using NLTK (wordnet). Additional considerations that are difficult to factor into the model are multi-language variability, colloquial shorthand commonly used, and the regular communication.



Class	Precision	Recall	F1-score	Support
almosthomeless	0.00	0.00	0.00	13.00
anxiety	0.57	0.64	0.60	99.00
assistance	0.59	0.42	0.49	57.00
domesticviolence	0.47	0.32	0.38	59.00
food_pantry	0.00	0.00	0.00	8.00
homeless	0.73	0.30	0.42	27.00
ptsd	0.41	0.69	0.51	127.00
relationships	0.52	0.69	0.60	113.00
stress	0.00	0.00	0.00	8.00
survivorsofabuse	0.75	0.05	0.10	57.00

Figure 10: Custom Model precision, recall, F1-score and support statistics

11. Challenges and Limitations

1) *Data Management*: The initial dataset contained numerous unnecessary columns, which made the data processing and cleaning steps more complex and time-consuming. Efficient data management was a key challenge, particularly in ensuring only relevant features were used in the sentiment analysis model.
2) *Data Integration:* We faced difficulties while integrating multiple datasets, mainly because not all columns were similar. The inconsistent structure across different datasets added to the complexity of data preprocessing and required additional work to harmonize the datasets into a consistent format.
3) *Data Access*: Accessing data from social media platforms like Reddit and Twitter turned out to be a challenging task. Since these platforms have paid APIs, our ability to access and retrieve data was limited, which hindered us from getting as much data as we would have wanted.
4) *Data Modification*: Modifying the data to suit the requirements of our model was another challenge. This included activities like dealing with missing values, encoding categorical variables, and scaling numerical variables.
5) *Size of the Dataset:* The size of the dataset posed another significant challenge. Training the sentiment analysis model on a large dataset is computationally expensive and time-consuming. We needed to balance the need for a large enough dataset to train a robust and accurate model with the computational resources required.

6) *Computational Limitations*: Our computational resources were limited by Google Colab's 12.7 GB RAM limit, which resulted in the environment crashing frequently. This constraint not only slowed down the model training process but also limited the size of the data that could be processed at once.
7) *Limited Expert Knowledge*: Finally, our team's limited knowledge in the field of sentiment analysis and mental health conditions impacted the project. Expertise in these areas is crucial for features selection, model design, interpretation of results, and the application of the sentiment analysis model. Despite the team's best efforts to educate themselves and make informed decisions throughout the project, the lack of expert knowledge was a limitation.

Conclusion

We aimed to acquire, pre-process and enrich social media post data in meaningful ways to understand how semantic text can be analyzed to extract meaningful mental health diagnosis. Our initial models had poor performance ranging from 0% to 65%. As result an alternative model was built using natural language tokenizers and built-in dictionaries with specialized semantic analyzers which performed very well in comparison averaging 50% accuracy for some classes and 75% for some classes such as “survivorsofabuse”. Whilst these results are much better overall there are several

Future Directions

- **Validation and Testing:** It is crucial to validate the models and analysis using external datasets or cross-validation techniques. Evaluating the generalizability of the findings and testing the robustness of the models can provide more confidence in the results.
- **Feature Engineering**: The current dataset may contain additional features that could enhance the predictive power of models. Exploring feature engineering techniques, such as creating new variables or transforming existing ones, could potentially uncover hidden patterns and improve model performance.
- **Larger datasets:** The use of commercial APIs may also aid us in generating much larger datasets. Additionally, we can incorporate more elaborate web scrapping methodologies such as ‘beautifulsoup4’ python modules.

References

Academic Papers:

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Classification Algorithms. In Mining Text Data (pp. 163-222). Springer.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain Separation Networks. arXiv preprint arXiv:1608.06019.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint Xiv:1810.04805.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. arXiv preprint arXiv:1909.10351.

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv preprint arXiv:1908.08962.

Zouaq, Amal & Gagnon, Michel & Ozell, Benoit. (2010). Semantic analysis using dependency-based grammars and upper-level ontologies. 1.

Research Studies:

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073-1137.

Primack, B. A., Shensa, A., Escobar-Viera, C. G., Barrett, E. L., Sidani, J. E., Colditz, J. B., & James, A. E. (2017). Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults. Computers in Human Behavior, 69, 1-9.

Smith, A., Anderson, M., & Kumar, M. (2018). Social Media Use in 2018. Pew Research Center.

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2020). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clinical Psychological Science, 8(6), 916-925.

Van den Eijnden, R. J., Lemmens, J. S., & Valkenburg, P. M. (2016). The social media disorder scale. Computers in Human Behavior, 61, 478-487.

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. Social Issues and Policy Review, 11(1), 274-302.

Code References:

Accessed on 06/02/2023
https://colab.research.google.com/drive/1-NgakmQG5dPnjiF5ZmcjmmojDQU2VVS-?usp=sharing

Accessed on 06/10/2023
https://colab.research.google.com/drive/1Xnd_KftFmFKRYdWDGeWG-35x4EtTU1rE?usp=sharing#scrollTo=vwmVF2msKGiI