**36118 Applied Natural Language Processing**

**Assignment 2 Part B - Group 43**

**Categorizing Depressive Tweets**

Name: Able Varghese

Email: ablevarghese@student.uts.edu.au

Student ID: 24712198

# 

# Project objectives and scope

The goal of this project is to examine Twitter data by looking at tweets that contain popular hashtags for depression. The goal is to group these tweets into various subjects or themes in order to obtain insight into the motivations behind individuals' adoption of these hashtags. The study uses a data-driven methodology to mine the enormous quantity of publicly accessible Twitter data for useful information. This study attempts to help us better understand people's experiences and reasons for talking about depression on social media by classifying the tweets based on their content.

The study of Twitter data primarily pertaining to tweets that have been labelled with well-known hashtags connected to depression is included in the project's scope. The study seeks to classify the tweets based on their content and obtain insights into the factors that influence users' involvement with these hashtags. The scope comprises gathering a sizable volume of Twitter data, cleaning and pre-processing the data, choosing pertinent hashtags for research, and using the proper subject classification approaches. In order to gain a greater understanding of the users' motives and experiences connected to depression, the analysis will concentrate on finding similar themes or subjects within the tweets. Neither the project nor any qualitative interviews or surveys include direct communication with Twitter users. The findings and insights gained from the analysis of the collected data will contribute to the existing knowledge in the field of mental health discussions on social media platforms, particularly in relation to depression.

# Dataset Description:

The dataset used in this project consists of a collection of tweets that have been labelled according to their sentiment with respect to depression. The dataset was created by extracting tweets from Twitter using a set of keywords related to depression, and then manually annotating the tweets for sentiment by human annotators. The first column named "message to examine" contains the text of the tweet, and the second column named "label (depression result)" contains a binary label (0 or 1) that indicates whether the tweet expresses depression or not.

**Key Characteristics of the Dataset:**

**Size**: The dataset comprises a significant number of tweets, providing ample data for training and evaluation purposes.

**Tweet Text**: Each tweet is represented as a text string, reflecting the original content posted by users on the Twitter platform.

**Labels**: The dataset includes labels assigned to each tweet as either ‘1’ or ‘0’, indicating whether they express depression or not respectively.

**Topic Diversity**: The tweets cover a wide range of topics, reflecting the diversity of discussions and opinions found on Twitter.

**Noise and Variability**: Similar to real-world Twitter data, the dataset may contain noise, misspellings, abbreviations, emoticons, and other linguistic variations commonly found in user-generated content.

**Dataset Usage and Pre-processing**:

To ensure accurate analysis, the dataset will undergo pre-processing steps to handle noise, remove irrelevant elements (e.g., URLs, usernames, hashtags), and standardize the text representation. Additionally, tokenization, stemming, and lemmatization will be applied to normalize the tweet text further.

It is important to note that the specific details of the dataset, such as the number of tweets, distribution of sentiment labels, or specific pre-processing steps applied, will be included in the project report based on the actual dataset used and its characteristics.

The availability of this dataset enables researchers and practitioners to develop and evaluate sentiment analysis models, explore sentiment trends, and gain insights into public opinions on various topics as expressed on Twitter.

# NLP methods and techniques

**1.** **Data Cleaning:**

Data cleansing, also known as data cleaning or data pre-processing, refers to the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. It is an essential step in preparing the data for analysis and ensuring its quality and reliability. Data cleansing involves several techniques and procedures aimed at improving the integrity and usability of the dataset.

The main objectives of data cleansing include:

* Removing Duplicate Entries: Identifying and eliminating duplicate records or observations that may distort the analysis or lead to biased results.
* Removing stop words: Stop words are words that are frequently used in a language but have little meaning, such as "the", "a", "an", "in", "of", and so on. Remove stop words in data cleansing refers to removing these frequently occurring words from a text corpus in order to reduce noise and improve the accuracy of natural languages processing tasks such as text classification or sentiment analysis.
* Removing punctuation & special characters: Removing punctuation and special characters can also help to keep unwanted noise or errors out of the data, such as when text is copied from a source that includes formatting or special characters that aren't relevant to the task at hand.
* Removing links: Links or URLs can introduce noise into a text corpus because they may not contain information relevant to the task at hand. Furthermore, some natural language processing techniques may treat URLs as separate words or entities, which can reduce the analysis's accuracy.
* Tokenization: Tokenization refers to the process of dividing a text corpus into smaller units known as tokens. Depending on the needs of the natural language processing task at hand, these tokens can be individual words, phrases, or even individual characters. Tokenization is a crucial step in many NLP jobs and is typically used as a pre-processing step before using methods like sentiment analysis or topic modelling.
* Lemmatization is the process of condensing words to their simplest or dictionary form, sometimes referred to as a lemma. In order to group comparable words together and simplify the text corpus, this technique comprises eliminating word inflections and variations like tense, case, and gender.
* Managing Conflicting Data: To maintain data quality and dependability, it is important to spot and address any anomalies in the dataset, such as competing values or contradicting information.
* Analysts may reduce possible biases, increase the correctness of analytical results, and guarantee the validity and dependability of conclusions by doing data cleaning.

**2.** **Exploratory Analysis:**

Exploratory Analysis: An essential element in the data analysis process is exploratory analysis, commonly referred to as exploratory data analysis (EDA). To get a general grasp of the dataset's key features, patterns, and relationships, it entails inspecting and visualising the data. Exploratory analysis aids in locating key elements, spotting anomalies, exposing possible patterns, and directing decisions on additional research and modelling.

The following are the primary goals of exploratory analysis:

Visualising Data: To analyse correlations, trends, and distributions within the collection, visual representations such word clouds, histograms, scatter plots, and box plots are often created.

Feature Selection: Determining the applicability and significance of various features or variables in light of the objectives of the study, which aids in the direction of dimensionality reduction approaches or feature selection.

**3.** **Sentiment analysis:**

Sentiment analysis, commonly referred to as opinion mining, is a computer approach used to identify and categorise the sentiment represented in a text, such as reviews, social media postings, or client feedback. Understanding and quantifying the emotional tone, personal opinion, or attitude expressed in the text—whether it be favourable, negative, or neutral—is the aim of sentiment analysis.

The following steps are commonly included in the sentiment analysis process:

Pre-processing of text: The text data goes through pre-processing stages include tokenization, which separates the text into individual words or tokens, eliminating punctuation and stop words, conducting stemming or lemmatization to break down words to their simplest forms, and deleting often used terms with little significance.

2. Feature Extraction: From the pre-processed text, pertinent features or properties are extracted. These characteristics may include particular words, n-grams (a run of related words), or other language components that convey mood.

3. Sentiment Classification: The text is divided into many sentiment groups using rule-based algorithms. lexicon-based techniques that rate the emotion of words using word lists or sentiment dictionaries.

4. Sentiment Scoring: Based on the existence and polarity of the sentiments observed, sentiment scores are applied to the entire text after identifying the sentiment of specific words or phrases. This may entail adding up the sentiment scores of various words or calculating an overall sentiment score using more complex algorithms.

5. Interpretation and Analysis: The sentiment analysis results are interpreted and analysed to gain insights into public opinion, customer satisfaction, or overall sentiment trends. Visualizations, statistical analyses, or comparative studies can be used to further analyse and interpret sentiment patterns.

**4.** **Topic Modelling**

Topic modelling is a computational technique used to uncover latent topics or themes within a collection of documents or texts. It is a popular method for exploratory analysis of large text datasets and can assist in organizing, understanding, and extracting meaningful information from unstructured text data.

In the process of topic modelling, we involved the following steps:

1. Document-Term Matrix: A document-term matrix is created, where each row represents a document or text, and each column represents a unique term or word in the entire corpus. The matrix captures the frequency of terms within each document.
2. Topic Modelling Algorithm: Topic modelling algorithms, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), are applied to the document-term matrix. These algorithms aim to identify the underlying topics by discovering patterns of co-occurring words in the documents.
3. Topic Extraction: The topic modelling algorithm assigns a probability distribution to each document, indicating the likelihood of the document belonging to different topics. Each topic is represented by a distribution of words, indicating the words most associated with that topic.
4. Model Evaluation: Various metrics, such as coherence scores or perplexity, can be used to evaluate the quality of the topic model. These metrics assess the coherence and interpretability of the extracted topics.
5. Interpretation and Analysis: The resulting topics can be examined and interpreted by reviewing the most representative words associated with each topic. Visualizations, such as word clouds or topic proportion distributions, can aid in understanding the relationships between topics and their prevalence in the dataset.

**5. Word Embedding**

A common method in Natural Language Processing (NLP) for representing words as dense vectors in a high-dimensional space is word embedding. Machines can better grasp and handle textual data because it captures the semantic and contextual links between words. Here is a thorough description of word embedding in action:

1. Tokenization: The first stage is to separate the text into tokens, or groups of words. Tokenization is the name of this procedure.
2. The process of creating a vocabulary involves gathering all distinctive terms from the corpus of text. Each word in the lexicon is given a special index or identification.
3. Word Representation: Each word in the lexicon is represented by a vector with set dimensions. Word embedding techniques try to give comparable vectors to words with related meanings.
4. A word embedding model may be trained using a variety of approaches. Word2Vec is one well-liked algorithm. The model learns to predict the context (neighbouring words) of a target word after being trained on a huge corpus of text data.
5. A context window is established around each target word. The quantity of nearby words that are taken into account for context depends on the size of the window. These nearby terms aid in conveying the meaning and usage of the word in various situations.
6. Each word in the lexicon is given a vector representation as the model learns to represent it. The vectors are taught to recognise semantic links, including analogies and word similarity.
7. Similar words are placed closer to one another in a vector space, which is used to order the word vectors. The semantic links between words are preserved by the vector space.
8. Transfer Learning: The word vectors may be used for a variety of NLP tasks, including sentiment analysis, text classification, and machine translation, once the word embedding model has been trained. It is possible to adjust the pre-trained word vectors or utilise them as input features for later models.

Word embedding enables NLP models to make use of the semantic linkages between words, enhancing their comprehension of text and enabling more precise and efficient language processing tasks by representing words as dense vectors in a high-dimensional space.

**6. Text Classification**

Natural language processing (NLP) tasks frequently require text classification, which involves classifying textual data into preset categories or classes. Numerous applications, including sentiment analysis, spam detection, subject categorisation, and others, make extensive use of it. A thorough explanation of the text categorisation procedure is provided below:

1. Feature Extraction: The textual input must be transformed into numerical characteristics that computer learning models can comprehend in order to do text classification. Several methods, including Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word Embeddings (such as Word2Vec, GloVe), and more sophisticated methods like BERT (Bidirectional Encoder Representations from Transformers), are used to do this. These methods capture the crucial information about words or phrases in the texts by representing the textual data in a vectorized representation.
2. Splitting the Dataset: Training and testing sets are created from the preprocessed data. The testing set is used to gauge how well the trained model performs on new data, whereas the training set is used to train the text classification model.
3. Model selection: Select an appropriate deep learning or machine learning model for classifying text. For text categorization, some popular models are Naive Bayes, Support Vector Machines (SVM), Logistic Regression, Random Forest, and Neural Networks (such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN)).
4. Model Training: Using the training dataset, the chosen model is trained. The input characteristics (textual data) and the related labels or categories are analyzed by the model while trained to identify patterns and correlations.
5. Evaluation of the Model: The model's performance is assessed following training using the testing dataset. Depending on the particular needs of the text classification job, a variety of assessment measures may be utilised, including accuracy, precision, recall and F1 score
6. Model Deployment and Prediction: The model may be used to generate predictions on fresh, unexplored text data after it has been trained and assessed. The trained feature extraction techniques are applied to the preprocessed textual input in the deployed model to predict the class or category of the input text.

# Findings

1. Sentiment Distribution: The analysis of the dataset reveals the distribution of sentiments expressed in the tweets. It provides insights into the percentage of positive, negative, and neutral tweets, indicating the overall sentiment trends within the dataset. This information helps in understanding the sentiment landscape and the prevalence of different sentiment categories.

2. Popular Topics and Sentiments: By analyzing the dataset, it is possible to identify the popular topics or themes that elicit strong sentiments. This can provide valuable insights into the subjects that Twitter users feel strongly about, enabling a better understanding of public opinion on specific issues or events.

3. Sentiment Fluctuations: The study of the dataset might reveal changes in sentiment over time or between various themes. It makes it possible to spot patterns or trends in sentiment, such as fluctuations in public opinion during particular occasions or movements in sentiment towards particular topics. These perceptions aid in monitoring sentiment dynamics and comprehending the variables affecting sentiment shifts.

4. Model Performance: The assessment of the sentiment classification model sheds light on how well it performs in identifying the sentiment of tweets. To evaluate the performance of the model, metrics like accuracy, precision, recall, and F1-score may be generated. This data aids in evaluating the robustness and dependability of the sentiment categorisation model.

5. Model Challenges: The assessment procedure may highlight certain issues or limits with the sentiment categorisation model. For instance, the model could perform worse in some situations when categorising tweets that contain irony or sarcasm. Understanding these restrictions is essential for correctly understanding the sentiment analysis findings and identifying potential areas for future development.

6. Feature Importance: During the evaluation, it is feasible to pinpoint the key terms or characteristics that are crucial for accurately classifying emotion. This knowledge aids in identifying the textual elements that have a significant impact on sentiment predictions and can offer helpful hints about the sentiment drivers in the dataset.

7. Comparative Analysis: If multiple sentiment classification models or techniques are evaluated, it allows for a comparative analysis of their performance. This analysis can reveal the strengths and weaknesses of different approaches and assist in selecting the most suitable technique for sentiment analysis in similar datasets or future projects.

8. Practical Applications: The insights gained from the data, model, and evaluation have practical applications in various domains. For example, businesses can utilize the sentiment analysis results to gauge customer sentiment towards their products or services, enabling them to make informed decisions and tailor their strategies accordingly.

The first part of the code is preparing the data for an LDA model (Latent Dirichlet Allocation) on tweets related to trust, where the label is not depression. The LDA model is used to identify the topics that are being discussed in these tweets. Two topics are identified, one of which seems to be related to positive emotions, while the other topic is about love, new experiences, and having fun.

The second part of the code is using Word2Vec to create word embeddings for the word "depression" and visualizing the embeddings using a scatter plot. Word2Vec is a method for representing words as vectors in a high-dimensional space, where similar words are closer together. The visualization helps to see how similar or dissimilar words are to the word "depression" in this vector space.

For topic modelling, the results show that the logistic regression classifier has the highest accuracy score, precision score, recall score, and F1 score on the training set compared to SVM and Naive Bayes classifiers. However, the classification report shows that the model has relatively low performance in the minority class (label 1), indicating that the model may not generalize well to new data. It is recommended to further evaluate the model's performance on the validation set and potentially adjust the classification threshold or use techniques like oversampling or undersampling to balance the classes.

On both the training and validation datasets, the code produces confusion matrices and classification reports for three different models (logistic regression, support vector machine, and multinomial Naive Bayes).

The confusion matrices for each model show the number of true positives, false positives, true negatives, and false negatives, which can be used to evaluate the model's performance. The classification reports include additional metrics such as precision, recall, and F1-score, which can be used to evaluate the models' performance in greater depth.

The accuracy scores assigned to each model on the validation set indicate how well each model generalises to new, previously unseen data.

Overall, it seems like the code is exploring different techniques for analyzing mental health-related Twitter data, including topic modelling and word embeddings.

In conclusion, the analysis of the dataset, performance evaluation of the sentiment classification model, and insights derived from them contribute to a deeper understanding of sentiment trends, popular topics, sentiment fluctuations, and the strengths and limitations of the model. These insights have practical implications for various domains and pave the way for further research and applications in sentiment analysis.

# Project outcomes and value added

*The outcomes of applying* ***data cleansing*** *techniques to the dataset used in this project included the following:*

1. **Improved Data Quality**: Data cleansing helped to enhance the quality of the dataset by removing inconsistencies, errors, and irrelevant information. It eliminated noise, such as spelling mistakes and grammatical errors, which could impact the accuracy of the sentiment analysis results.
2. **Standardized Data Format**: Data cleansing involved standardizing the format of the data. This process ensured that the tweets in the dataset are uniformly structured and followed a consistent pattern. Standardization simplifies subsequent processing steps, such as feature extraction and sentiment classification.t.
3. **Enhanced Data Interpretability**: Cleansed data is easier to interpret and analyse. By eliminating noise and inconsistencies, the sentiment analysis results became more reliable and interpretable.
4. **More Accurate analysis: Data cleaning results in a better dataset that makes it possible to conduct more precise analysis. The final dataset offered a clearer and more representative sample of tweets by eliminating noise and standardising data. In turn, this led to more accurate model predictions and deeper perceptions of public opinion.**

**In conclusion, using data cleansing techniques on the dataset resulted in higher-quality data, a more consistent format, the elimination of duplicates, greater interpretability, and more precise analytical findings. These findings strengthened the basis for later research and improved our comprehension of the trends in the Twitter dataset.***The outcomes of applying* ***sentiment analysis*** *on the given dataset include the following:*

1. **Sentiment Classification: Each tweet is categorised into one of the sentiment categories—positive, negative, or neutral—after sentiment analysis is performed on the dataset. This categorisation helped us make sense of the emotions each tweet communicated.**
2. **Sentiment Distribution: An overview of the sentiment distribution within the dataset was supplied by the sentiment analysis findings. This data is displayed statistically or visually, such as the proportion of good, negative, and neutral tweets. It provided information on the dataset's general sentiment patterns.**
3. **Sentiment Trends: Sentiment patterns and trends were discovered by examining the sentiment of the tweets over time. These are helpful for monitoring sentiment shifts in relation to certain occasions, subjects, or trends. It gave information about the changes and evolution of sentiment.Identification of Highly Positive or Negative Tweets**:
4. Sentiment analysis identified tweets that have exceptionally positive or negative sentiment. These tweets were highlighted as notable instances, potentially revealing important opinions, experiences, or sentiments expressed by users in the dataset.
5. **Insights into User Sentiment**: The sentiment analysis outcomes offered insights into the sentiment of Twitter users regarding the topics or themes covered in the dataset. By understanding the sentiment expressed by users, it became possible to gain a deeper understanding of their attitudes, opinions, and emotions related to specific subjects.
6. **Validation of Sentiment Analysis Techniques**: Applying sentiment analysis on the dataset allowed us for the evaluation and validation of the chosen sentiment analysis techniques or algorithms. The accuracy and performance of the sentiment analysis model were assessed based on known sentiments within the dataset.
7. **Potential Applications**: The outcomes of sentiment analysis can have various applications. For example, businesses can use the sentiment analysis results to understand customer sentiment towards their products or services. Researchers can utilize the findings to analyse public opinion on specific topics or track sentiment trends. Policymakers can use the insights to gauge public sentiment related to specific policies or initiatives.

Overall, applying sentiment analysis on the dataset will provide a comprehensive understanding of the sentiment expressed in the tweets, enabling insights into user sentiment, sentiment distribution, trends, and notable instances of highly positive or negative sentiment. These outcomes can be valuable for both academic research and practical applications in domains such as marketing, public opinion analysis, and decision-making.

*The outcome of applying* ***topic modelling*** *on the given dataset can include the following:*

1. **Identification of Latent Topics**: Topic modelling algorithms, such as Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF), was applied to the dataset to identify latent topics within the tweets. These topics are underlying themes or subjects that emerge from the textual content of the tweets. The outcome was a set of identified topics that represent the major themes present in the dataset.
2. **Topic Distribution**: The analysis yielded information about the distribution of topics within the dataset. It provided insights into the prevalence and frequency of each topic and how they are represented across the tweets. This information was useful for understanding the dominant topics and their relative importance within the dataset.
3. **Topic Keywords**: Topic modelling techniques generated a list of keywords or terms associated with each identified topic. These keywords represented the key terms that are most strongly associated with a particular topic. The outcome was a collection of topic keywords that help summarize and describe the main content of each topic.
4. **Topic Coherence and Interpretability**: The project evaluated the coherence and interpretability of the generated topics. Coherence measures the semantic similarity between the keywords within a topic and can indicate how well-defined and meaningful the topic is. A higher coherence score suggests more coherent and interpretable topics, enhancing the usefulness and reliability of the outcome.
5. **Insights into Tweet Content**: By applying topic modelling, the project provided insights into the underlying themes and content distribution within the dataset. This helped in understanding the prevalent discussions, trends, and interests among Twitter users. The outcome contributed to a better understanding of the topics discussed on Twitter related to the dataset's domain.
6. **Practical Applications**: The outcomes of topic modelling had practical applications in various domains. For instance, marketers can use the identified topics to understand the interests and preferences of their target audience. Researchers can utilize the topic distribution to study the prevalence of specific themes or to explore relationships between topics and sentiment. The outcomes also aided in content recommendation systems, trend analysis, or social listening applications.

Overall, applying topic modelling to the dataset will yield insights into the latent topics present in the tweets, their distribution, and the associated keywords. These outcomes can contribute to a better understanding of the dataset's content, facilitate further analysis, and support decision-making processes in diverse fields.

# Challenges faced and solutions

In applied natural language models, feature or method selection is just as crucial as data cleaning. After cleaning the data, we must decide which NLP approaches should be explicitly used to analyze the dataset. The nature of the issue and the desired results will determine this choice.

Our team worked closely together to solve this problem and met to discuss the best NLP strategies for our particular requirements. We looked into several possibilities, including sentiment analysis, text classification, named entity recognition, and topic modeling. Depending on the precise project objectives, a technique might be chosen among those that give distinctive insights.

We may extract meaningful information from the text data and gain insightful knowledge by choosing and using suitable NLP approaches. To better grasp the data and draw conclusions that can be put into practice, this approach uses models and algorithms created to analyze and interpret human language patterns. We can optimize the analysis and get the required outcomes from our applied natural language model by combining data cleaning and wise procedure selection.

# Limitations and future steps

# Because the dataset used for this research is quite small, getting accurate and trustworthy findings will take a lot of work. Additionally, there is a problem with a class imbalance in the dataset, where the proportion of tweets identified as depressed is substantially lower than those marked as not depressed. In particular, the depressed class makes up just 22% of the overall weight, while the dominant class rules with 78%.

Future work will entail acquiring a more extensive and balanced dataset to solve these constraints and improve the study. We can collect a more representative sample of tweets by growing the dataset, increasing the precision and generalizability of the analysis's findings. To gather more information and improve the current dataset, it would be helpful to investigate different sources and techniques, such as using tools like Twitter scrapers.

Additionally, it is essential to designate a committed team member or domain expert to assess the performance and efficacy of the produced model to guarantee its validity and dependability. Metrics, including precision, recall, accuracy, and F1-score, are evaluated as part of this assessment process, and the model is iteratively improved depending on the input and insights gleaned.

Overall, we seek to improve the analysis and deliver more solid and persuading findings in the field of depression identification in tweets by addressing the concerns of data quantity and imbalance and combining other data sources and expert reviews.

# Project progress timeline with milestones achieved

* 18th April: Topic and dataset selection. Preliminary data is included in the PowerPoint template.
* Team meeting on April 22 to discuss the chosen subject and start the data cleansing process.
* 28th April: The data cleaning process is primarily finished after incorporating input and making the necessary improvements.
* 2nd May: Team meeting to discuss and choose the precise NLP methods to apply to achieve desired results. Data cleansing is complete.
* 7th May: Group meeting to discuss coding work updates. Specific NLP approaches, such as text categorization or topic modelling, were given to each team member.
* Code testing was done on May 13 and comments and learnings from the results were provided for documentation.
* The project report will be written on May 14 and 15, including thorough explanations of the methodology, findings, and outcomes.

# Personal Contribution

Being a part of team of students in their first semester, I had to put a lot of effort into understanding the project. But all the members equally contributed to the project. There were several meetings held, some face to face but most of it online, to understand and discuss the project and the challenges we faced. Although there were ups and downs in progress of the project, overall it was a good experience.

I contributed to the team with finding the right dataset, although everyone searched equally. I also helped my team during the project summary report where I shared my thoughts and opinions with my team regarding the project objectives and goals. During the execution phase, I developed the codes for Sentiment Analysis and Word embeddings and also helped others with any troubleshooting and debugging. I equally contributed in making the project report as well.

Since this is my first time working on a data science project, I am aware about my lack of knowledge and expertise. I was guided by my peers and mentors during lectures and other personal communications which helped me in doing my project. I definitely have a long way to go as I need to spend more time analyzing and understanding data and how to gain insights from it.

# References

1. Babu, N. V., & Kanaga, E. G. (2022). Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. *SN Computer Science; SN Comput Sci, 3*(1), 74. <https://doi.org/10.1007/s42979-021-00958-1>
2. Chakrabarti, H. (2021, Mar 30,). *Tokenising rants — the data*. <https://medium.com/analytics-vidhya/tokenising-rants-d61b1a0def80>
3. Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2016). *The language of mental health problems in social media*. <https://doi.org/10.18653/v1/W16-0307>
4. Jung, H., Park, H., & Song, T. (2017). Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals. *Journal of Medical Internet Research; J Med Internet Res, 19*(7), e259. <https://doi.org/10.2196/jmir.7452>
5. Kaggle. (2021). *Sentimental analysis for tweets*. <https://www.kaggle.com/datasets/gargmanas/sentimental-analysis-for-tweets?select=sentiment_tweets3.csv>
6. Li, J., Li, X., Wang, S., Li, J., & Sheng, Q. Z. (2016). Sentiment Analysis for Depression Detection on Social Networks. (Vol.10086, pp. 807-810). Springer International Publishing AG. <https://doi.org/10.1007/978-3-319-49586-6_59>
7. Mori, K., & Haruno, M. (2021). Differential ability of network and natural language information on social media to predict interpersonal and mental health traits. *Journal of Personality; J Pers, 89*(2), 228-243. <https://doi.org/10.1111/jopy.12578>
8. Oladipupo, F. (2022, Aug 19,). *How to scrape twitter data for sentiment analysis with python and power BI*. <https://techcommunity.microsoft.com/t5/educator-developer-blog/how-to-scrape-twitter-data-for-sentiment-analysis-with-python/ba-p/3593365>
9. Prieto, V. M., Matos, S., Álvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PloS One; PLoS One, 9*(1), e86191. <https://doi.org/10.1371/journal.pone.0086191>
10. ProjectPro. (2023, May 9,). *10 NLP techniques every data scientist should know*. <https://www.projectpro.io/article/10-nlp-techniques-every-data-scientist-should-know/415>
11. Tana, J., Shcherbakov, A., & Espinosa-Leal, L. (2022). Sentiment analysis of depression related discussions in the Suomi24 discussion forum. *Informaatiotutkimus, 41*(2). <https://doi.org/10.23978/inf.122663>
12. Wolff, R. (2021, October 20,). *Natural language processing (NLP): 7 key techniques*. <https://monkeylearn.com/blog/natural-language-processing-techniques/>
13. Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., & Luo, J. (2021). Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study. *JMIR Infodemiology; JMIR Infodemiology, 1*(1), e26769. <https://doi.org/10.2196/26769>

# Appendix

