Analysis For Nationa’s Image Using Twitter Sentiment Analysis

INFO 5810 SECTION 001

University of North Texas

PROJECT GROUP-8

TEAM:

HARSHAVARDHAN GYARALA	11437843(Leader)
SNEHA SAHITHI PAMARTHI	11514108
VAISHNAVI SIMHACHALAM	11525465
NIKHIL POLISETTI	11546616
VASHISTH SUKHADIYA	11520555
VAMSI ANDE	11518655

DATE: 11/27/2022

Introduction:

In this modern era, technology plays a vital role and is the key factor in producing lots of data. Now a day's internet has become the main communication platform from where people express their ideas and opinions. Social media like Twitter, Facebook, and Google are allowing people to connect with each other. These are sites where people share their ideas, express their views, talk about public issues, and a lot more than this. Some of them are useful and some are not. These sites allow people to communicate around the world.

Using sentiment analysis algorithms, social media data may be automatically examined to detect the intensity of expressed opinions. These algorithms have evolved over the past few years to look at more variables, such as a person's sentiment toward a topic or their emotions and have even merged text analytics with other inputs like multimedia analysis or social network analysis. Sentiment analysis is contextual text mining that recognizes and extracts subjective information from the source material. Keeping an eye on online discussions enables businesses to identify the social sentiment surrounding their brand, product, or service. Recent developments in machine learning have greatly enhanced text-mining methods. The inventive application of cutting-edge artificial intelligence methods can be a useful instrument for doing in-depth study.

We have applied sentiment analysis to the data through the Twitter Platform. We have obtained various kinds of data through the Twitter API developer account by using the Python programming language. Based on the words in the context, we have classified the tweets and tried to decide the emotion of the person. All the references, methodology, data collection, and data analysis are done and explained below. For this research, we used Python programming language and rapid miner to analyze the data. We have used Python to get the data from platforms like Twitter and used Rapid Miner to implement a few machine learning algorithms as we can obtain accurate results compared to previous research.

Research questions:

1. Is recession really a factor of unemployment and inflation which affects the nation?
2. How many people in the nation are paying attention to their mental health and taking action to get out of their current conditions?
3. A country's reputation is greatly influenced by the percentage of anti-nationals living there. What proportion do they control?

Research Purpose:

The goal of this research is to analyze text data using sentiment analysis through naive Bayes and other machine learning algorithms. We also want to research how people's mental health has changed. And our goal is to find to what extent social networking platforms such as Twitter influence current society including politics and decision-making of people. In this research, we can learn about people's sentiments and opinions regarding policies, goods, brands, and other related topics by gathering tweets and looking at the individual's emotions expressed in them.

Literature Review:

Understanding the consumers' attitudes regarding health insurances is the focus, "Using Social Media to Identify Consumers' Sentiments Toward Attributes of Health Insurance During Enrollment Season." They used the NRC Emotion Lexicon, which offers each word's polarity as well as its accompanying emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), to mine Twitter chats and analyze them using a dictionary-based technique. This study's key finding is that political preferences, prescription drug benefits, and provider networks worry consumers. Customers also have faith in medical professionals but worry about unforeseen circumstances. These findings imply that additional study is required to identify the root causes of consumer motivations in order to improve insurance products.

The Novel Coronavirus disease 2019 caused by the Sars-Cov-2 virus Has become a pandemic with a growing number of cases globally. In this duration, the people have to face challenges, although primarily infection disease with physical health implications, also affect mental health and wellbeing. Studies said that depression, anxiety disorders, substance abuse, increased suicidal tendencies, and PSTD. Twitter is the biggest social media platform for hosting an abundant number of user-generated posts. It is considered as a gold mine of data. At that time people tweeted aggressively on twitter. Most of the people started having negative tweets but with increasing time people shifted towards positive and neutral comments. In April 2020 most comments were positive and about winning against CoronaVirus. The tweets have been collected, pre-processed, and then used for text mining and sentiment analysis. The results of the study concludes that while the majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited worldwide.

The severity of both the new coronavirus outbreak has resulted in significant financial challenges, worry, anxiety, and future concerns. Social media can offer a platform for tracking the state of people’s mental well-being in local areas. Examining the evolving speech use on social networking sites can support conventional survey-based methodologies and offer fresh perspectives on the condition of a country or region throughout a public-health emergency.

Twitter posts can reveal shifts in a community’s psychological health amid a health emergency when mass polling could not be possible. Anxiety, stress, and solitude have diverged more from readings in 2019. Early detection of areas where psychological health is deteriorating can result in society treatments.

The fact that Twitter accounts must not include all demographic groups as well as the language-based calculations are dependent on a randomized 1% digital signal of tweets are both limitations of this study.

Our projections haven't been verified during the assessment period during the assessment period due to a lack of public polling. We plan to verify these models with industry-recognized polling data in subsequent research.

In summation, real-time analysis of posts made on destination social media platforms can provide light on recently discovered public concerns. Early detection of regional trends can help with resource allocation, focused public health initiatives, and improved readiness for the current and upcoming public health catastrophes.

Recent studies in sociologists and other disciplines, such as medicine and entertainment, have benefited from sentiment classification, which reveals developments in human emotions, especially via social media. The approaches used for trend analysis on twitter data are the same as those used for other domains. Future research might study the application of these strategies to rich language research, such as pattern recognition and disambiguation.

The spreading of pandemic influenza makes people vulnerable and anxious on a worldwide scale. This type of situation considers conversation essential when developing better preventive actions because it does not alter the following of specific cut-off dates.

Methodology:

Python has been used to analyze the sentiment of user-generated tweets. Effective libraries for Python include Tweepy and TextBlob.

The official Python application for the Twitter API is called Tweepy, and it allows users to access Twitter using both OAuth and the more recent Basic Authentication mechanism. OAuth is the sole method available to access the Twitter API as Twitter no longer accepts Basic Authentication. It provides access to the thoroughly explained Twitter API. With Tweepy, you can obtain an object and employ every technique that the verified Twitter API provides. Tweets, Users, Entities, and Places are the primary Model classes in the Twitter API. In Python, it's incredibly simple to navigate through data because each access produces a response in JSON format. Using tweepy we can retrieve tweets in our wanted time frame and context.

One of the most used NLTK-based Python modules for analyzing text data is TextBlob. Almost all actions required for Basic NLP can be accomplished using it as a framework (Natural Language Processing). Sentiment Extraction and Spelling Correction are two of the more sophisticated aspects of TextBlob.

Tokenization via TextBlob allows the text blocks to be broken up into various phrases and clauses. Reading between the lines is significantly simpler as a result. In phrases, the noun is typically employed as an Entity. Additionally, it is among the most crucial NLP tools for dependency parsing. Using TextBlob, several nouns are taken out of a sentence. TextBlob can be used to tag sentences with various components of speech.

Acquiring data relating to research questions can be achieved by following methods. Large data is always imbalanced. So, Data cleaning and preprocessing plays an important role.

Using TextBlob we can filter our tweets related to covid and can acquire data related to the research. We can perform operations on this data in Rapidminer which is very essential to analyze Big data. Sentiment Analysis is the platform technique we are going to perform on all the datasets which we acquire and a classifier named Naïve Bayes is performed on this type of data to produce results. Will provide some labels relating to depression which are termed negative in sentiment analysis. Using Tokenization we can label those features and perform naïve bayes before covid timeframe and after covid timeframe. Will compare both the results and can answer our research question did covid affect people mentally or not.

A large weightage for Nation’s image analysis can be given to country people reacting for social purposes. Previously build algorithms will be used to acquire related data for this research statement. User tweets who have text label features relating to human rights, environment, gender equality, religious freedom, property rights, trustworthy, well-distributed political power, racial equity, cares about animal rights, committed to climate goals and fight for justice will be Tokenized and pass through Naïve Bayes classifier to predict the output. Predicted output analysis can be done using Rapidminer and visualize the output data. If there are more positive quoted tweets then it will add huge weightage for Country’s reputation.

Happy or Sad seems simple in terms of words but it is difficult to extract one’s emotion behind a few lines of sentence. Happiness of the people is rated very high and resembles Happy people is equal to Happy country. This research question can predict and give accurate solutions for our research. Above used algorithms will be used to obtain required data. Using nltk we can preprocess data by removing special characters and stopwords, TFIDF vectorization is used to transform the data into a sparse matrix format. Cosine similarity function can be used to detect similar tweets regarding one’s emotion and can be labeled into two vectors passing happy and sad tweets of people. We will verify Nation’s Happy percentage using Data visualization features from Rapid Miner using Bar charts, Histogram and mosaic graph. Accurate output will be obtained by proceeding with the proposed methodology and can differentiate people's happiness rate before and after pandemic too.

DATA COLLECTION AND CLEANING

Twitter provides a device with the means to collect data through its API (API). The streaming mechanism collects the input data from Tweets and performs any mandatory analyzation, percolation, or aggregation before adding the results to a data repository. The HTTP dealing mechanism asks the storage service for the user's query's response.

HTTP uses GET method requests and, in contrast to ATOM, can produce results that have been modified using Java Structural Object Syntax. Python scripts were created to communicate with the streaming API and get associated data. These scripts gather information based on keywords and output to different entities, such as dates, location, race, languages, hyperlinks, text, and so on.

DATA COLLECTION:

Twitter is a largest social media platform to get millions of data but to acquire all the desired tweets one should have a Twitter API developer account. We can extract tweets using python library tweepy.

There are authentication keys and tokens to perform the API. Below you can find our snippet used to collect data. We used the Twitter API Elevated account, which allows us to access up to 2 million tweets from the current date, to gather data from Twitter. But data extraction takes a long time. Data on recission, mental health, depression, and data for the United States have all been gathered. Although we applied for an academic search Twitter account, it was denied. so have collected the recent data using elevated account tokens. We used tweepy library to collect our data. We pulled 1,000 tweets every 20 minutes during this time-consuming procedure, and we have 1.5–2 million tweets for our project.

DATA CLEANING:

Gathering millions of samples could lead to imbalanced data and biased data. We need to first check if there are any null values in the data and drop it if there are any.

We can import libraries like regex in python and apply to the data to remove any special characters Like #,@,! and emojis if any. Converting all the text into lower case letters is important to tackle imbalancement. Splitting the text into a single word will make our work much easier thankfully regex function has that feature to split into words.

Using NLTK we can download “stopwords” and can be applied to the data. To have unique words so that it does not affect the outcome which acts as an outlier for our research.

Below you can find a sample code-snippet which can be used for the data cleaning process.

```
X1 = “ text data which you want to perform”

#1. Remove non-Letters

X1 = X1.apply(lambda x: (re.sub("[a-zA-Z]", "", str(x))))
```

```
# 2. Convert to Lower case, split into individual words

X1= X1.apply(lambda x: str(x).lower())
words=X1.apply(lambda x: x.split())

#3. create stopwords using NLP

stops = set(stopwords.words ("english"))

# 4. removed stop words from the dataset

meaningful_words =  pd.Series (words).apply(lambda x: [item for item in x if item not in stops])
meaningful_words = meaningful_words.str.join(" ")
```

Remaining data cleaning for replacing missing values and duplicate samples was done in rapid miner

EXPLORATORY DATA ANALYSIS :

1. **Defining The Problem:**

    Identifying the problem is the first stage in any data analysis procedure. This is sometimes referred to as the "Problem Statement" in the context of data analytics.
    Creating a hypothesis and planning how to test it is necessary to define the purpose. Start by determining what business issue you're attempting to resolve. Although it seems simple, it can be much more difficult than it appears. Senior management of your company, for instance, can raise the question, "Why are we losing customers?" However, it's possible that this doesn't address the root of the issue. To properly outline a problem, a data analyst must thoroughly understand the business and its objectives.

2. **Collecting the Data:**

    Once your objective has been established, you must devise a strategy for collecting and aggregating the required data. A key part of this is selecting the data you need. These could be quantitative (numeric) data, such as sales figures, or qualitative (descriptive), such as customer reviews. All data goes into one of three categories: first-party data, second-party data, or third-party data. Examine each one independently. In these we need Categorical data to perform sentiment Analysis. There are numerous tools available to you once you've developed a data strategy.
    Regardless of your sector or area of expertise, we need a data management platform (DMP). A DMP is a piece of software that enables you to recognize and collect data from many sources,

then manipulate, segment, and so on. DMPs are widely available. The data integration platform Xplenty, SAS, and Salesforce DMP are a few well-known enterprise DMPs. Try several open-source platforms like Pimcore or D:Swarm if you wish to experiment.

3. **Cleaning the Data:**

Once the Data collection is done the next step is to collect the data to provide the data for the better analysis. The Key steps will include Removing major errors, Duplicates or the null values and outliers, removing unwanted data, Filling in the major gaps.

It can be difficult to manually clean up large databases. Fortunately, we have a wide range of tools at our disposal to speed up the procedure. Like OpenRefine, open-source technologies are great for both in-depth analysis and simple data purification. Free tools, however, only provide a limited amount of capability for really huge datasets. Effective techniques for complete data purification include some R tools and Python packages (such as Pandas). One of the best data matching tools available is called Data Ladder.

4. **Analyzing the Data:**

Once the Data cleaning is done the next step is Analyzing the data. Your goal will have a significant impact on the type of data analysis you conduct. But there are numerous methods accessible. You may have heard of time-series analysis, regression analysis, univariate or bivariate analysis, and others.

Determine what has previously occurred using descriptive analysis. Descriptive analytics may be used by TopNotch Learning to examine student course completion rates. Or they might say how many people utilize their products at a certain time. The goal of diagnostic analytics is to determine why something occurred. On the basis of past data, predictive analysis enables you to pinpoint potential trends. Predictive analysis is frequently used in business to make predictions about the future, like future growth. With prescriptive analysis, you can suggest changes for the future. The process' final step in the analytics section is this

5. **Sharing the Results**:

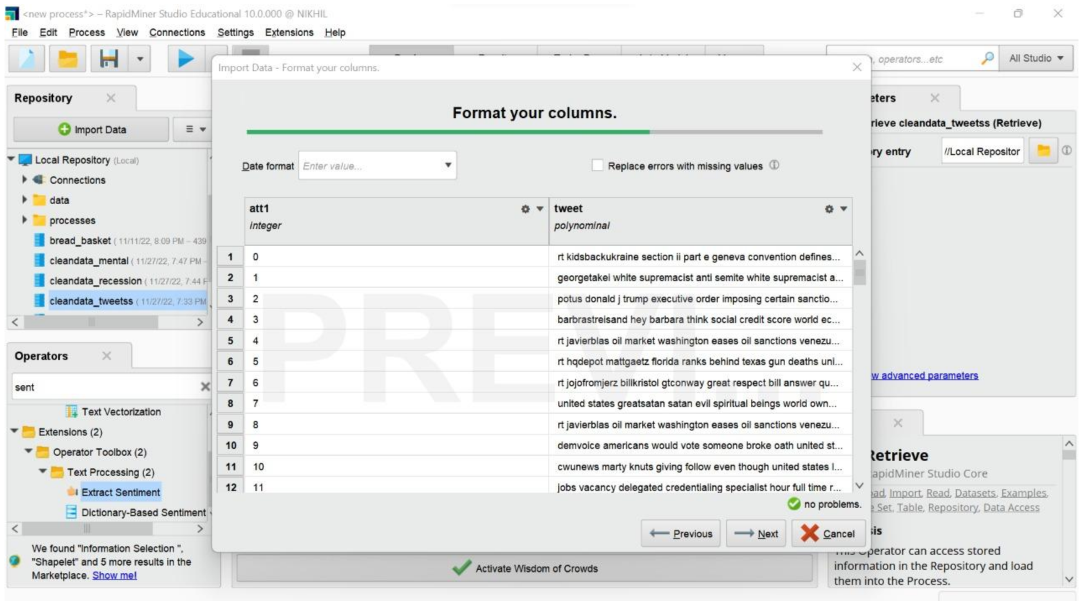
After the analysis the insights will be collected and the last step is to process these with the organizations or clients which can be complex since it will provide the outcomes. Sometimes we try to present these with the help of decision makers. How we present the results will influence the flow of the business. So that the organization can analyze and predict the risk factors or the things that are effecting the business. And it is crucial to find out if there are any gaps or flags that are needed to be considered in the business. There are various tools available to help to share the results and visualize them like Google Charts, Tableau, Datawrapper and Infogram. And in python we can use libraries like Plotly, Seaborn and Matplotlib.

6. **Embracing the Failures:**

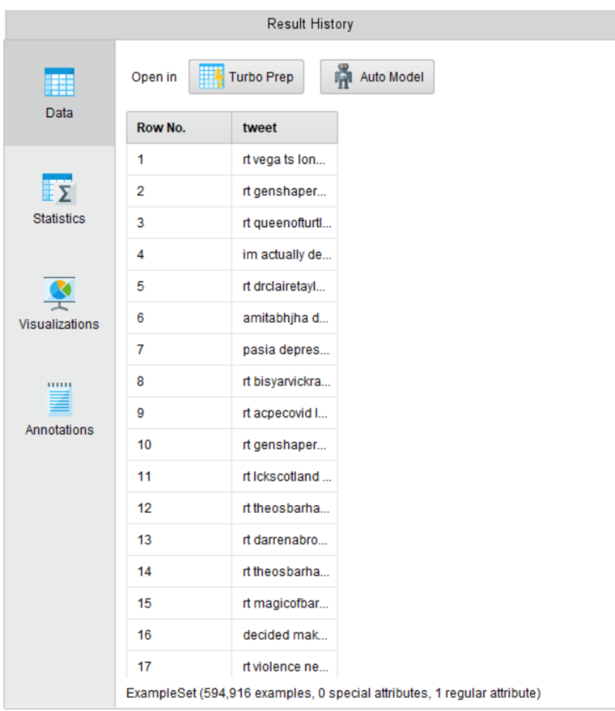
The final step is Finding the failures and understanding.Because data analytics is inherently chaotic, each project will require a distinct approach. For example, while cleaning data, you can see patterns that prompt an entirely new set of inquiries. You might then need to go to step one.

In addition, an exploratory analysis may bring to light a group of data points you hadn't previously considered employing. Or perhaps you discover that your key analyses' conclusions are inaccurate or misleading. This could be the result of data errors or human errors made earlier on in the process.

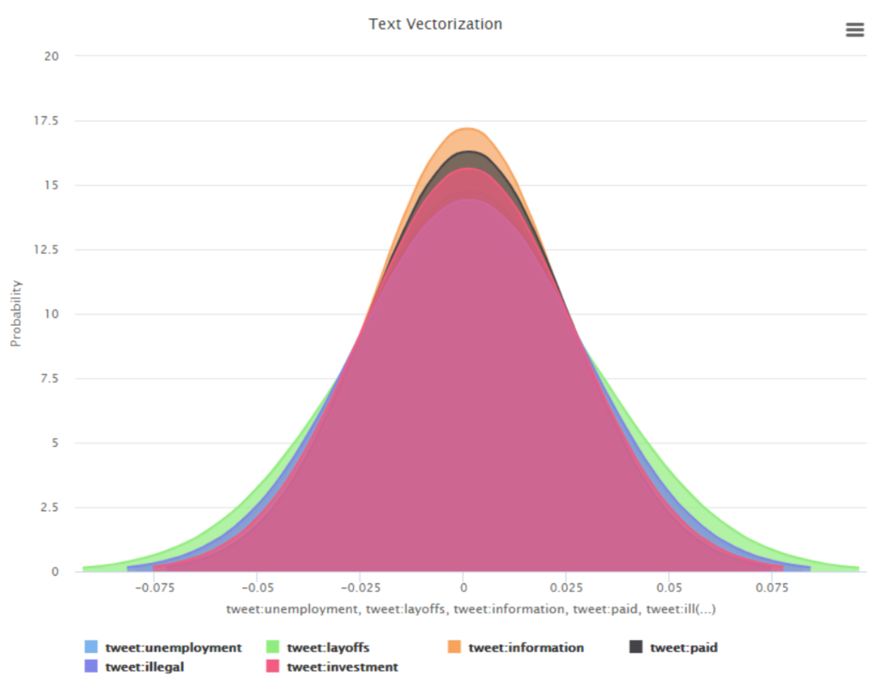
Importing dataset(recession) to rapidminer:



The datasets after cleaning was uploaded in rapidminer and is shown below:

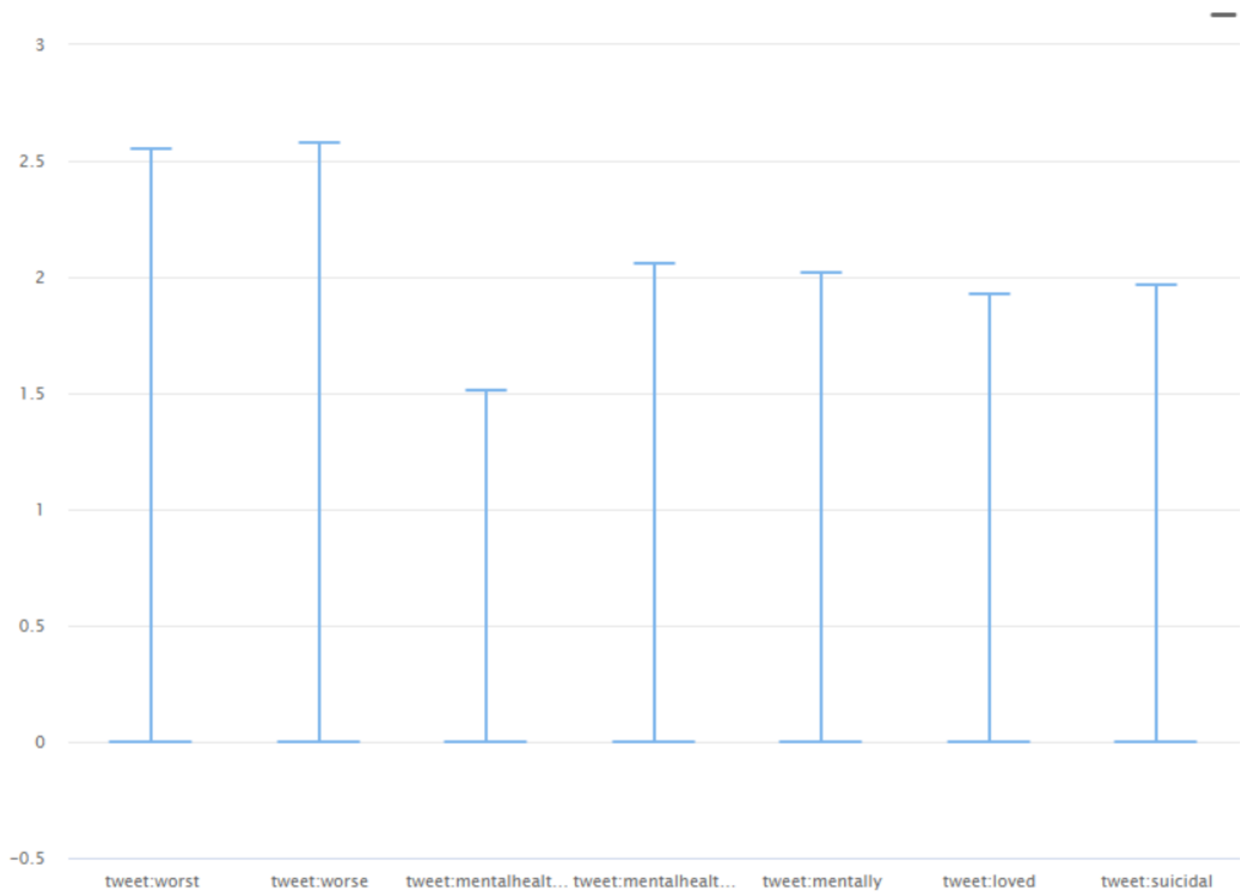


Below is the screenshot after text vectorization of rescission data.



From the graph we can say that the information in tweets is more and the investments in the tweet is pretty low and there are continuous layoff during rescission.

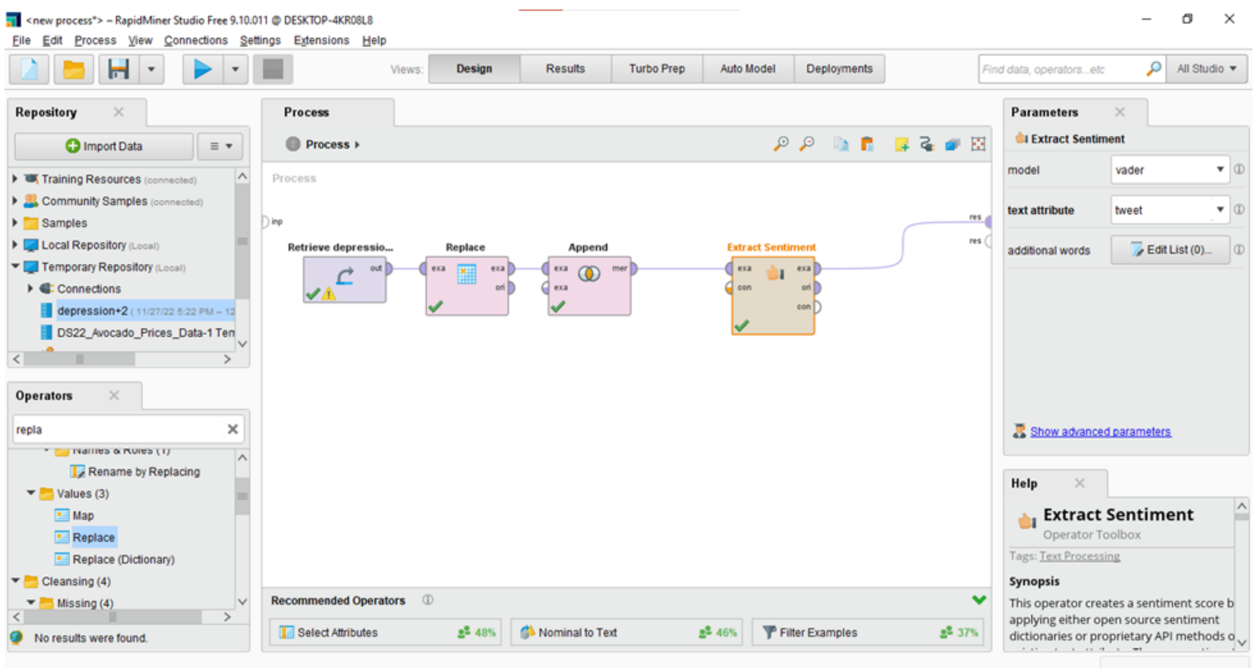
Below is the screenshot for visualization of mental health data with some key words(Box plot):

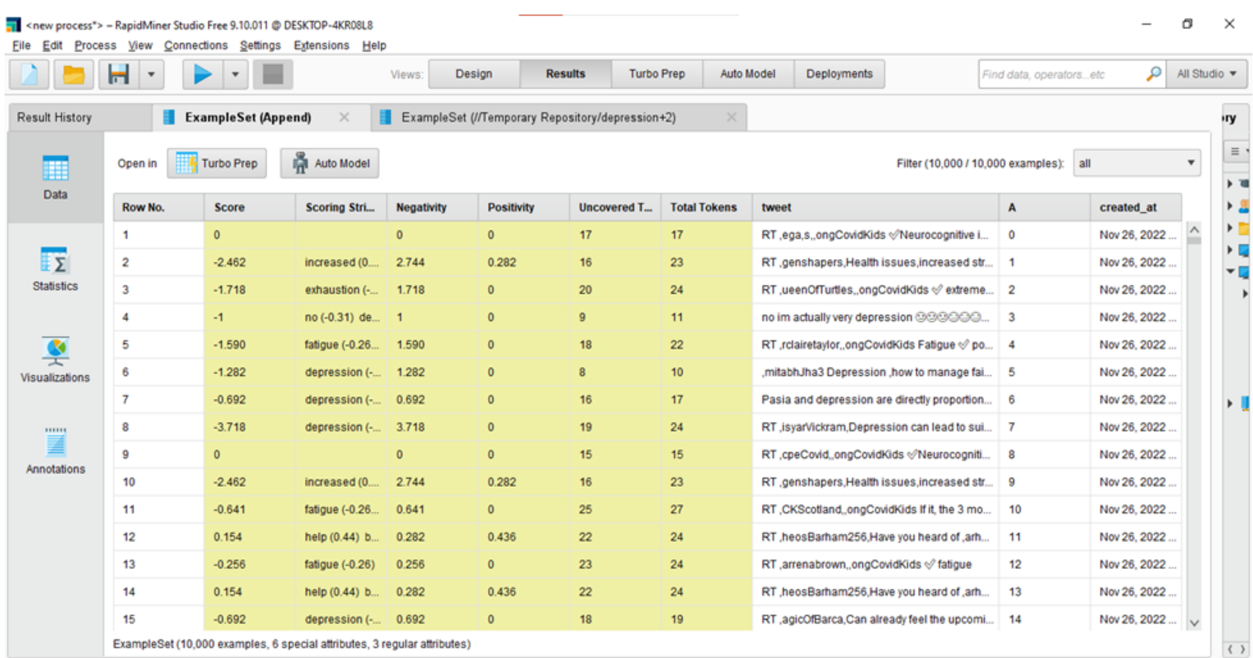


From the tweets we can see that worst and worse mental stability is more compared to rest.

Sentimental analysis

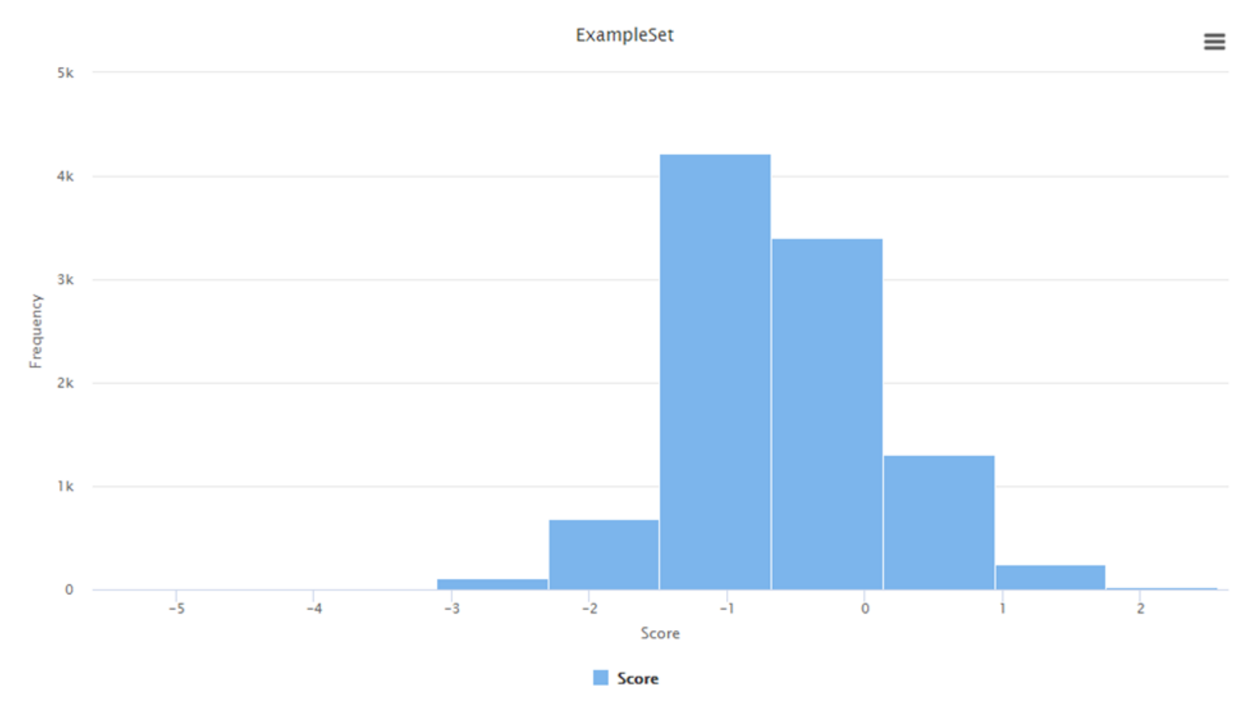
Applying the "extract sentimental" operator to the data is required to complete sentimental analysis. After completing this procedure, the total number of sentiment word scores found in the text will be computed, and the results will then be displayed. Because our data do not contain either the positives or the negatives, this operator must figure out how many positives and negatives there are using numerical values.



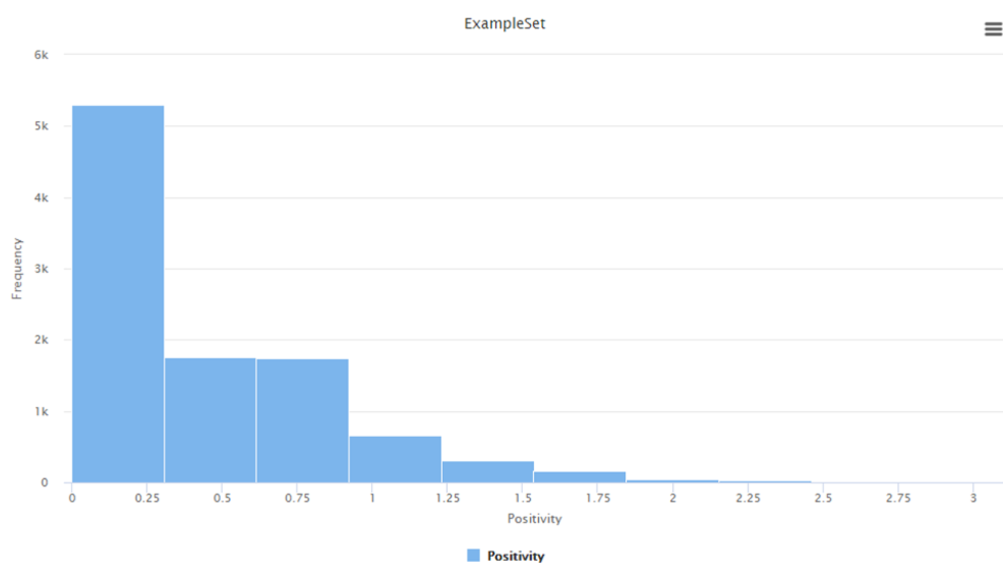


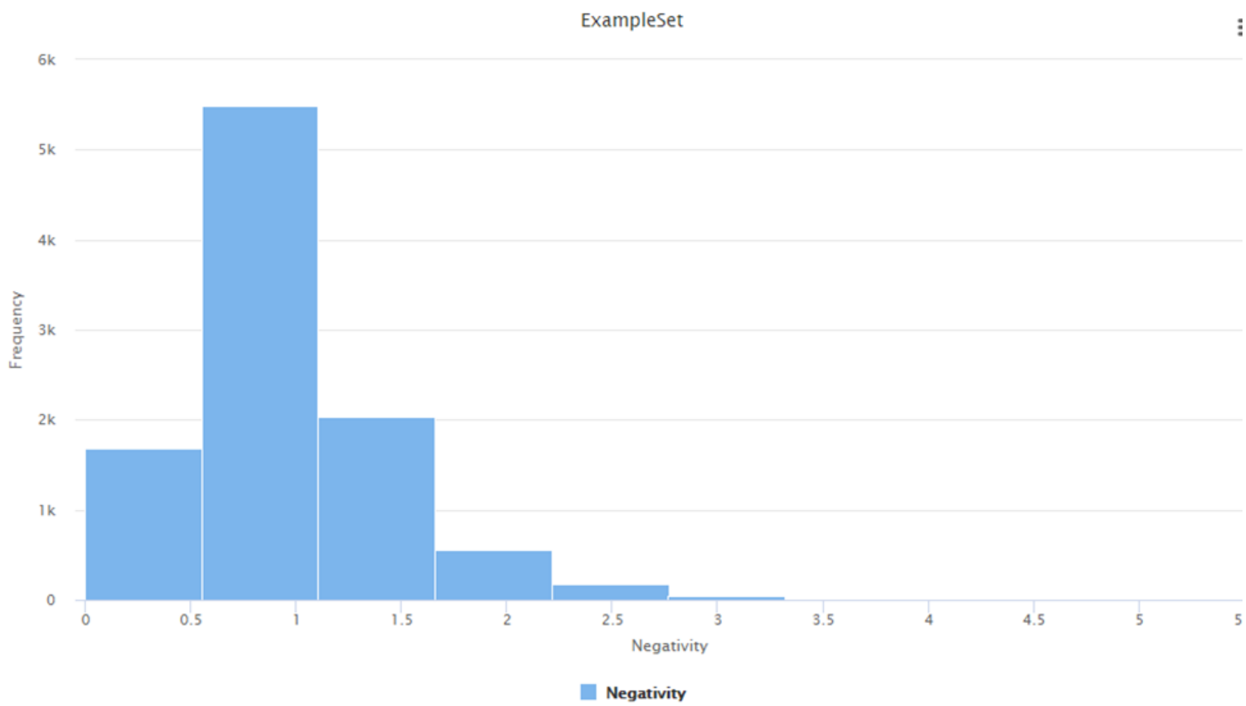
Visualization

The results provided by the application demonstrate that most of the words used in the text were unfavorable, with very few positive terms present among them. This indicates that the negative aspects of depression are more frequently discussed when people talk about it.

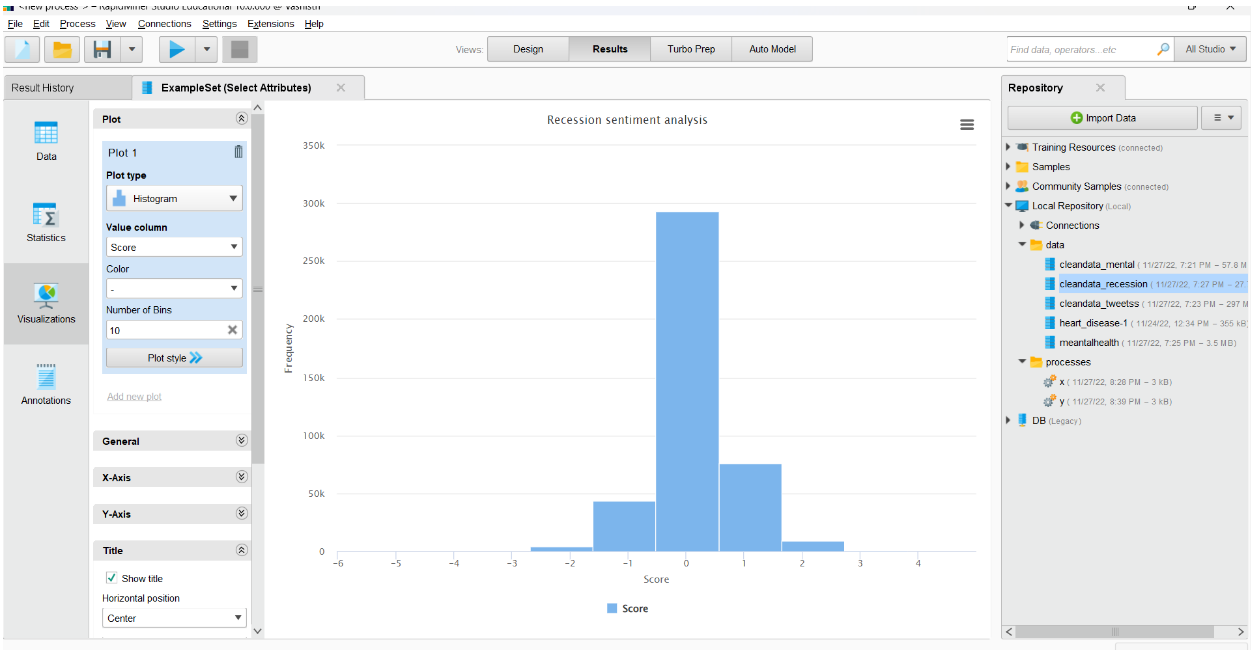


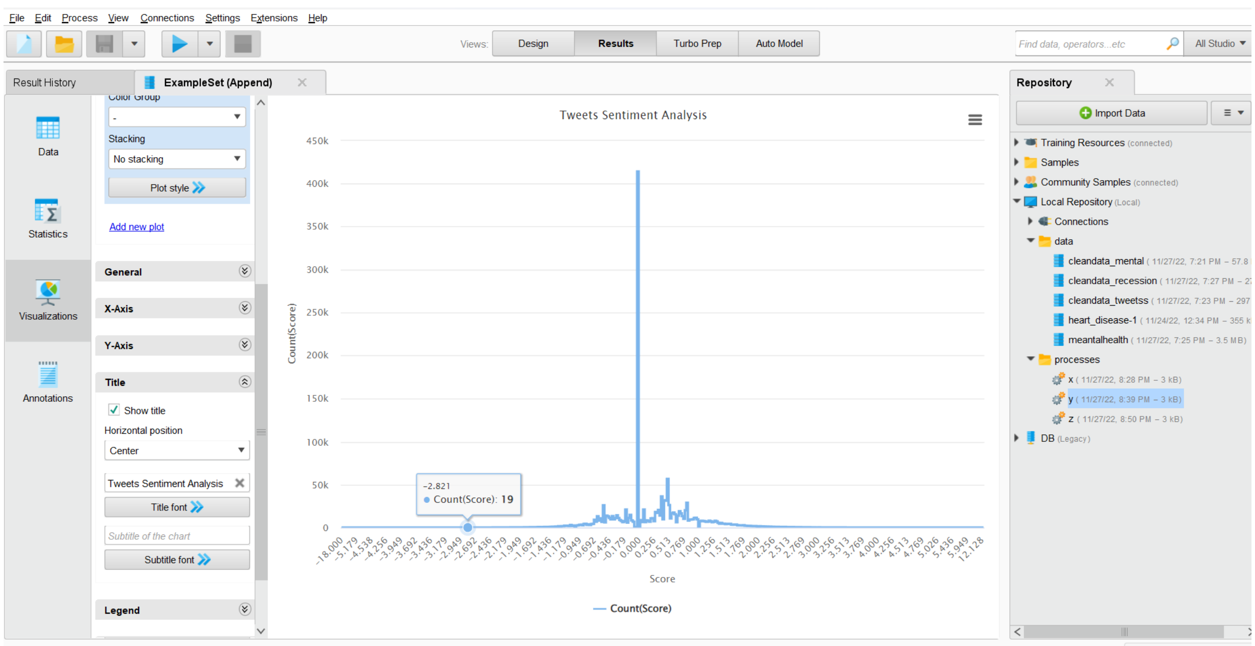
Most messages relating to depression were sent at 2:00. This indicates that, at this point, most respondents felt safe discussing their feelings about depression.



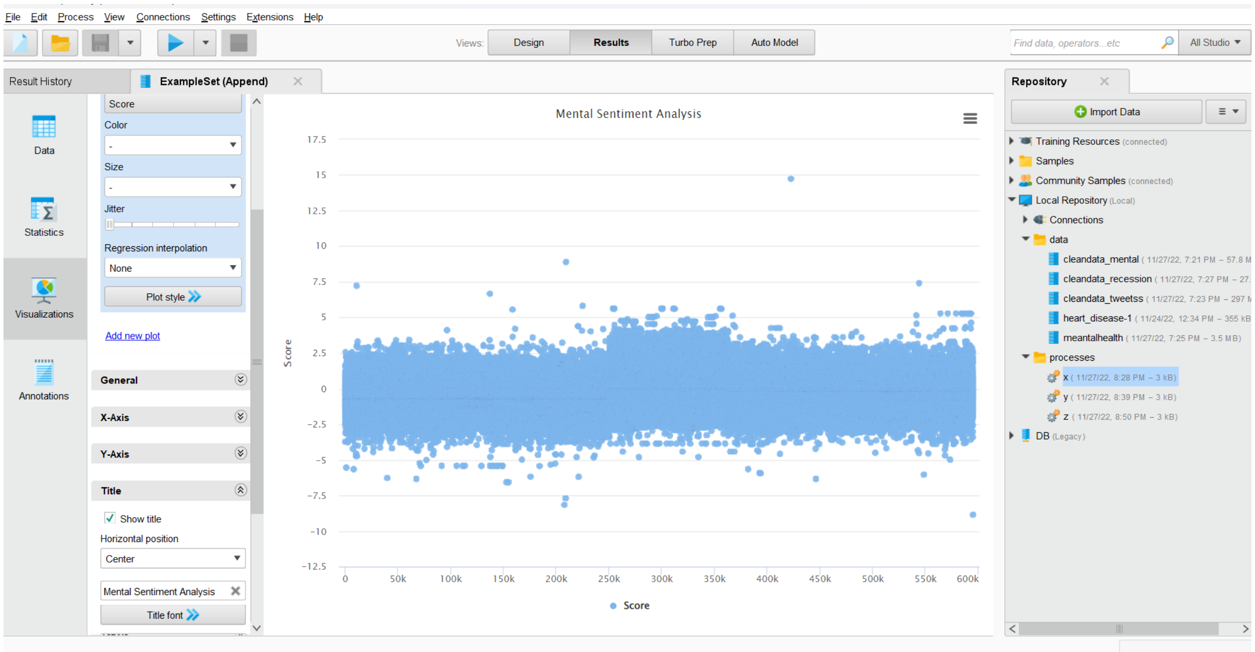


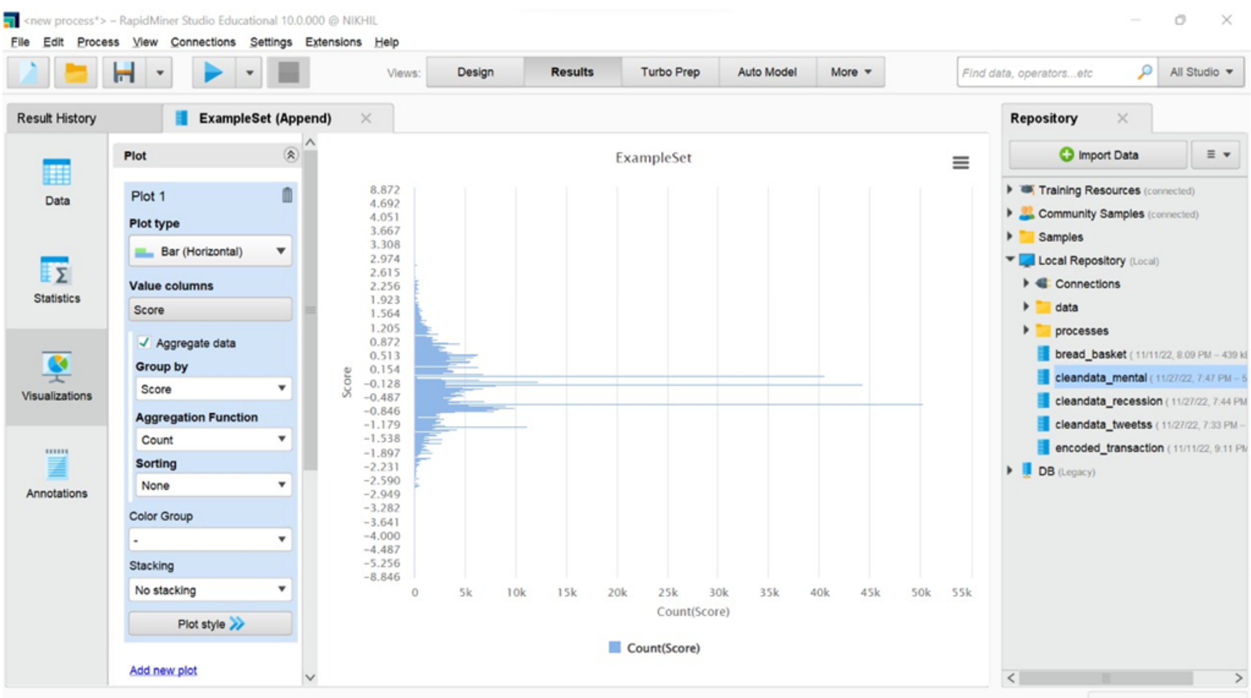
The two entities can be distinguished from one another in a meaningful way. This indicates that the dataset has a greater amount of negative information than positive information. This is due to the fact that the average level of negativity is 0.893, whilst the average level of optimism is only 0.370. The average of these scores is -0.522, which indicates that the negatives in the dataset are more prevalent than the positives. Based on the discourse that took place inside the tweets included in the dataset, it is clear that other individuals are attempting to offer solutions to the problem of depression. Among the various preventative steps that can be taken, the most obvious ones include talking about it, getting counseling, and avoiding stress.


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Select Attributes)
Data
Statistics
Visualizations
Annotations
Plot
Plot 1
Plot type
Histogram
Value column
Score
Color
Number of Bins
10
Plot style
Add new plot
General
X-Axis
Y-Axis
Title
Show title
Horizontal position
Center
Recession sentiment analysis
Frequency
Score
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
cleandata_mental
cleandata_recession
cleandata_tweetss
heart_disease-1
meantalhealth
processes
DB (Legacy)

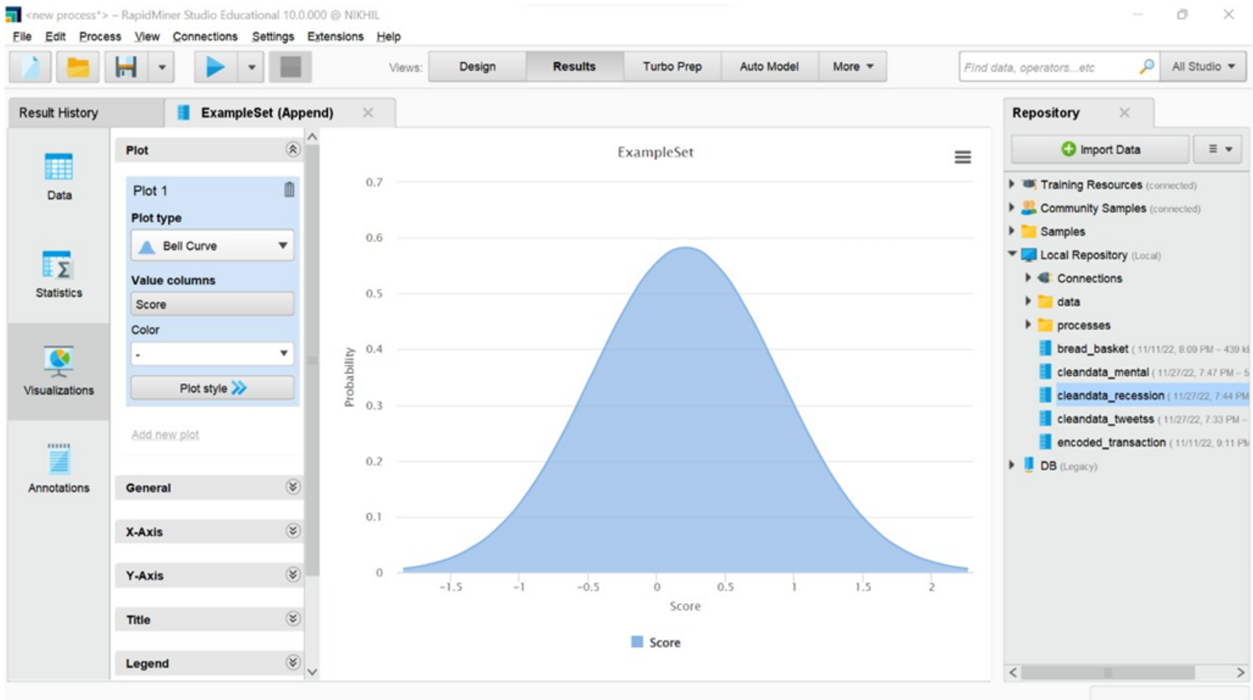

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Append)
Data
Statistics
Visualizations
Annotations
Stacking
No stacking
Plot style
Add new plot
General
X-Axis
Y-Axis
Title
Show title
Horizontal position
Center
Tweets Sentiment Analysis
Title font
Subtitle of the chart
Subtitle font
Legend
Tweets Sentiment Analysis
Count(Score)
-2.821
Count(Score): 19
Score
Count(Score)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
cleandata_mental
cleandata_recession
cleandata_tweetss
heart_disease-1
meantalhealth
processes
DB (Legacy)

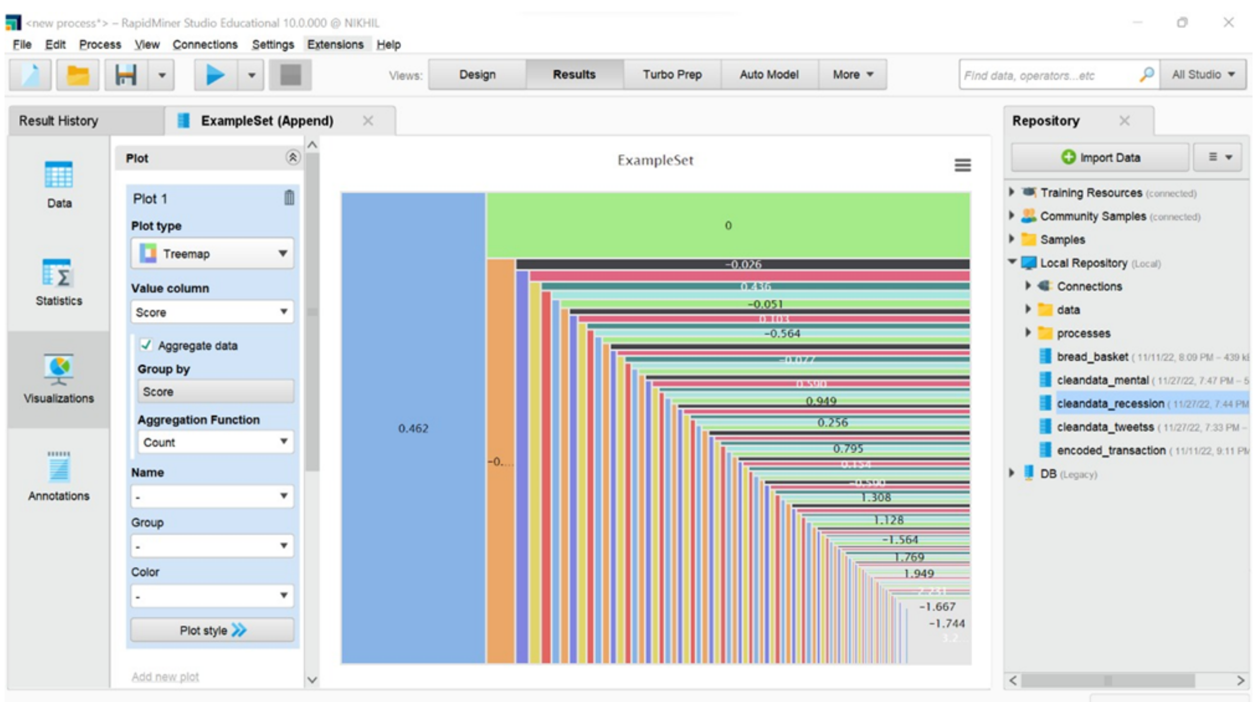
Scatter plot and bar plot of sentimental score of mental health data of nation.

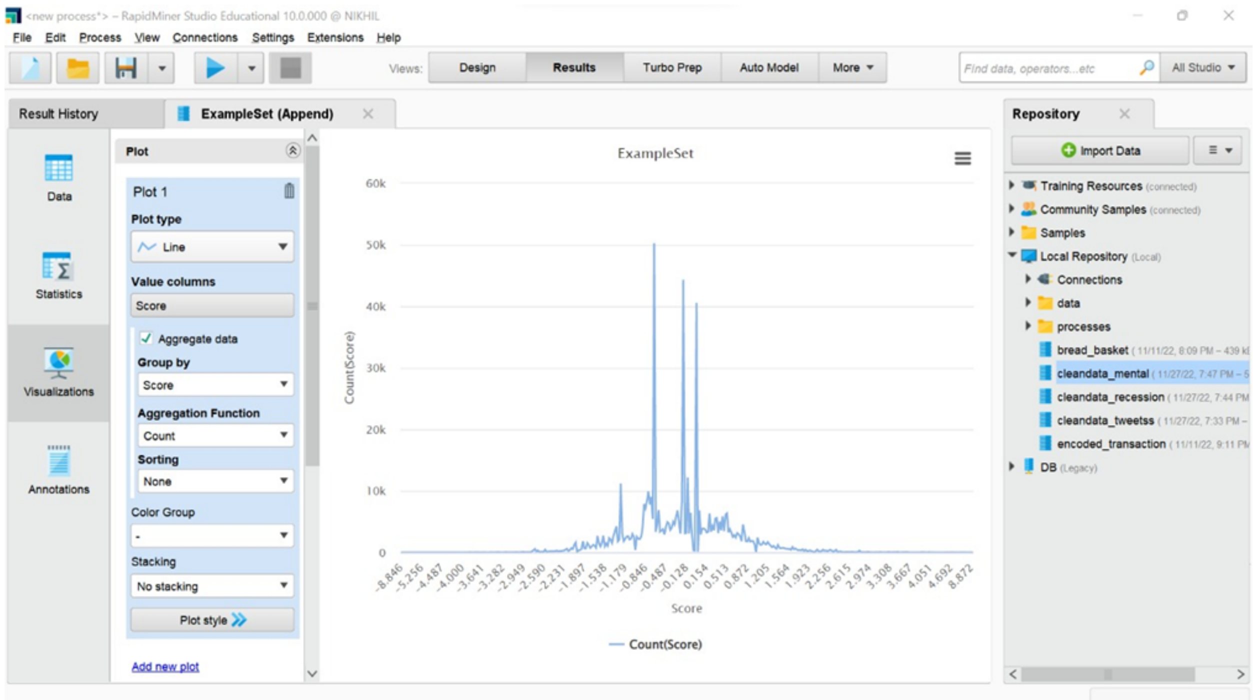




Belt curve for highest score for keywords in recession dataset







CONTRIBUTIONS

As a whole group, everyone participated in all the discussions and meetings which happened over zoom. Below are candidates contribution to the project until now:

HARSHAVARDHAN GYARALA - Extracted tweets for data collection using twitter api account in python, cleaned data inPython and imported in github ,contributed towards the report.

SNEHA SAHITHI PAMARTHI - Contributed in Data collection and gaining developer access, storing Data helped in writing report

VAISHNAVI SIMHACHALAM - Exploratory Data Analysis of cleaned data by writing essential keynotes Which was helpful towards visualization.

NIKHIL POLISETTI - Creating Data visualizations and explaining them, helped towards writing Report.

VASHISTH SUKHADIYA - Contributed in Data visualizations and writing report.

VAMSI ANDE - Contributed in Exploratory Data Analysis and helped towards research.

GITHUB LINK **- https://github.com/harshavardhan2204/INFO-5810-Section-001-Project-Group-8**

References:

1.Stanford Sentiment Treebank Dataset. Available online: https://nlp.stanford.edu/sentiment/code.html/ (accessed on 11 November 2012).

2. Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. LREc 2010, 10, 1320–1326.

3. Alm, C.O.; Roth, D.; Sproat, R. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 579–586.

4. Bartlett, M.S.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Recognizing facial expression: Machine learning and application to spontaneous behavior. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 568–573.

5. Heraz, A.; Razaki, R.; Frasson, C. Using machine learning to predict learner emotional state from brainwaves. In Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), Niigata, Japan, 18–20 July 2007; pp. 853–857
https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/

6..Twitter. (n.d.). Engineering. Twitter. Retrieved November 1, 2022, from https://blog.twitter.com/engineering/en_us

7.Staszkiewicz, P., Chomiak-Orsa, I., & Staszkiewicz, I. (1970, January 1). Table 3 from dynamics of the covid-19 contagion and mortality: Country factors, social media, and market response evidence from a global panel analysis: Semantic scholar. undefined. Retrieved November 1, 2022, from https://www.semanticscholar.org/paper/Dynamics-of-the-COVID-19-Contagion-and-Mortality%3A-a-Staszkiewicz-Chomiak-Orsa/2e173cc2311f6f05d8ac3d44a984809c8b74ec1d/figure/5

8.Ahmad, A. R., & murad, H. R. (2020). the impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan Online Questionnaire study. Journal of Medical Internet Research, 22, E19556. - references - scientific research publishing. (n.d.). Retrieved November 1, 2022, from https://scirp.org/reference/referencespapers.aspx?referenceid=3326343

9.SR;, K. A. B. A. (n.d.). Negative impact of social media panic during the COVID-19 outbreak in India. Journal of travel medicine. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32307545/

10.Yang, K.-C., Torres-Lugo, C., & Menczer, F. (2020, April 29). Prevalence of low-credibility information on Twitter during the COVID-19 Outbreak. arXiv.org. Retrieved November 1, 2022, from https://arxiv.org/abs/2004.14484v1

11.CS;, P. B. N. (n.d.). Mental health and the COVID-19 pandemic. The New England journal of medicine. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32283003/
12.Brooks SK;Webster RK;Smith LE;Woodland L;Wessely S;Greenberg N;Rubin GJ; (n.d.). The psychological impact of quarantine and how to reduce it: Rapid review of the evidence. Lancet (London, England). Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32112714/

13.AR;, K. (n.d.). Covid-19 in people with mental illness: Challenges and vulnerabilities. Asian journal of psychiatry. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32298968/

14.Li G;Miao J;Wang H;Xu S;Sun W;Fan Y;Zhang C;Zhu S;Zhu Z;Wang W; (n.d.). Psychological impact on women health workers involved in covid-19 outbreak in Wuhan: A cross-sectional study. Journal of neurology, neurosurgery, and psychiatry. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32366684/