SENTIMENT ANALYSIS OF VACCINATION TWEETS

NAME: Devesh Chhimwal
UID: 21BCS9711
Class: 21IST-A

Dataset:

The vaccination-related tweets that make up the sample were intentionally chosen for sentiment analysis. It includes a selection of tweets about vaccinations, immunization, and similar issues that have been made on various social media sites, such Twitter. The dataset attempts to record a wide range of user moods, opinions, and attitudes around immunizations. Key features of dataset are:

1. Text: Each entry in the collection corresponds to a single tweet and contains the pertinent textual information. There may be hashtags, mentions, URLs, and other components that are typical of tweets in the text.
2. Metadata: The dataset may include additional metadata associated with each tweet, such as the timestamp of the tweet, user information (e.g., username, follower count), and location. These attributes can provide contextual information that may be useful for analyzing sentiments across different user groups or geographic regions.
3. Size and diversity: In order to provide reliable sentiment analysis, the dataset is anticipated to have a significant amount of tweets. The precise number of tweets can range from a few thousand to tens of thousands or even hundreds of thousands. The dataset attempts to include tweets from a diverse group of users, including people, businesses, celebrities, and media organizations, expressing a wide range of viewpoints and attitudes.

Problem Statement Description:

The problem statement centers on the analysis of opinions shared in tweets about immunizations. The main goal is to create a sentiment analysis model that can correctly categorize tweets into several sentiment categories, such as positive, negative, or neutral. This analysis tries to comprehend public beliefs, attitudes, and worries about immunizations, offering insightful information for public health initiatives and communication plans.

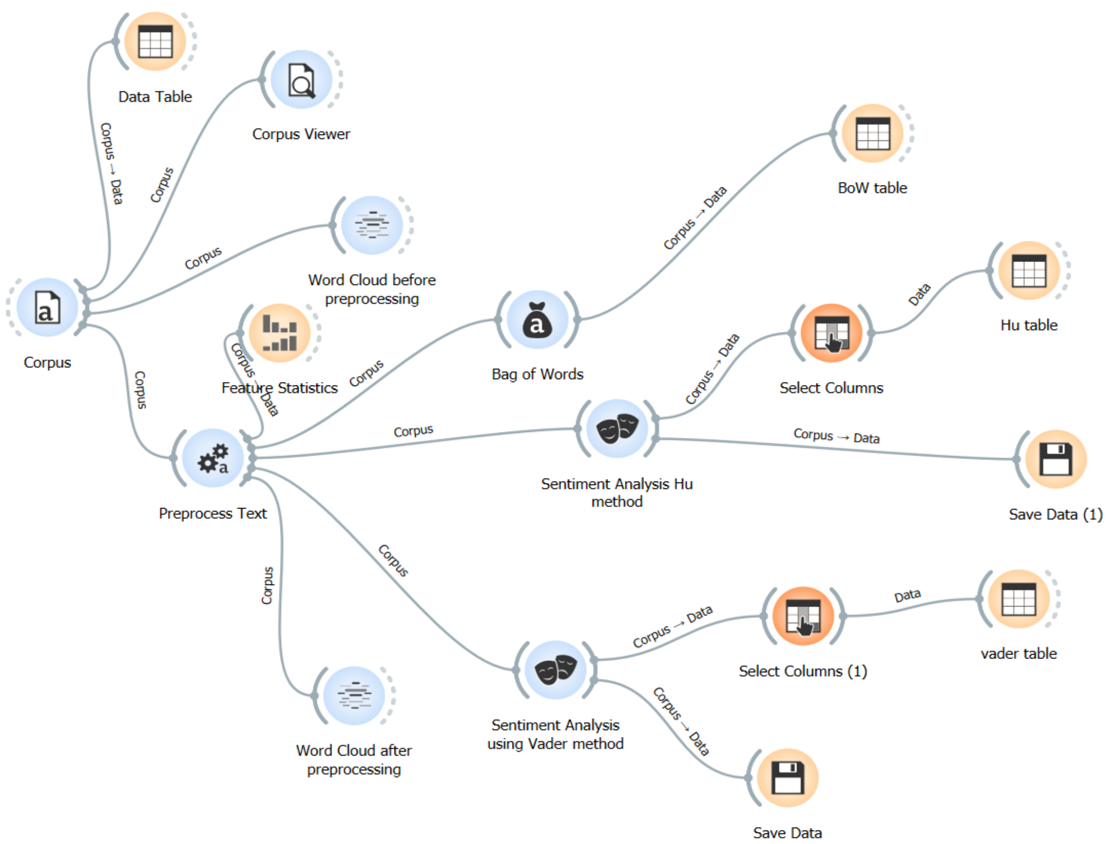
The problem entails the following key aspects:

1. Data Collection: Creating a broad and representative dataset of tweets relating to vaccinations is the first difficulty. This entails gathering a sizable quantity of tweets from numerous sources, including people, businesses, and media outlets. To achieve a thorough analysis, the dataset must include a variety of attitudes expressed against immunizations.
2. Data Pre-Processing: The preprocessing of the data is necessary before training the sentiment analysis model. In order to do this, the text must be cleaned of distracting elements like URLs, hashtags, and mentions. To standardize the text representation, tokenization, stemming, or lemmatization techniques may also be used.
3. Feature Statistics: Features must be extracted in order to represent the tweet text in a way that is appropriate for sentiment analysis. This may involve methods for text

encoding like bag-of-words, n-grams, word embeddings, or other techniques. The difficulty of the sentiment analysis task and the availability of computer resources determine which features should be used.

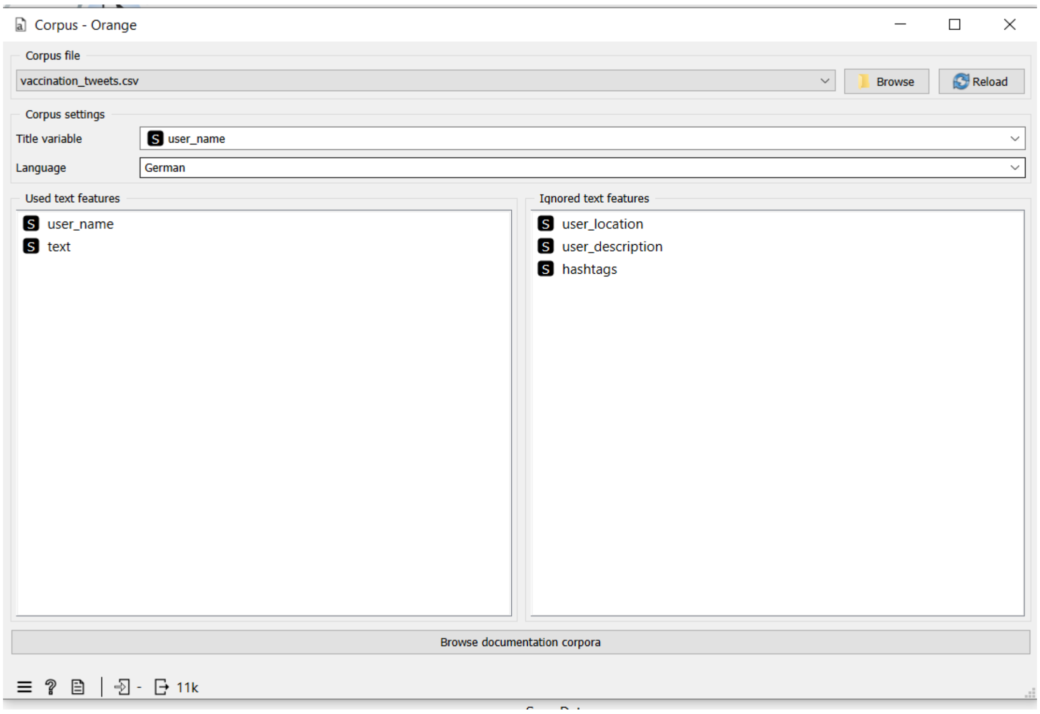
4. Model Selection: The sentiment analysis model can be created using a variety of machine learning or deep learning algorithms. Recurrent neural networks (RNNs), decision trees, random forests, logistic regression, and support vector machines are examples of common methodologies. The model should be educated and assessed using the proper evaluation metrics such as accuracy, precision, recall, or F1 score to assess its performance.
5. Getting Sentiment Insights: Once the sentiment analysis model is created, it can be used to categorize the sentiments of unlabeled tweets. Insights into the sentiment distribution across tweets about vaccinations are provided by this, enabling a better understanding of the public's perceptions, worries, and attitudes. These data can be used to improve public health communication tactics, identify prevalent sentiment trends, address vaccine reluctance, and focus vaccination programs to certain target populations.

Solution with used widget description:

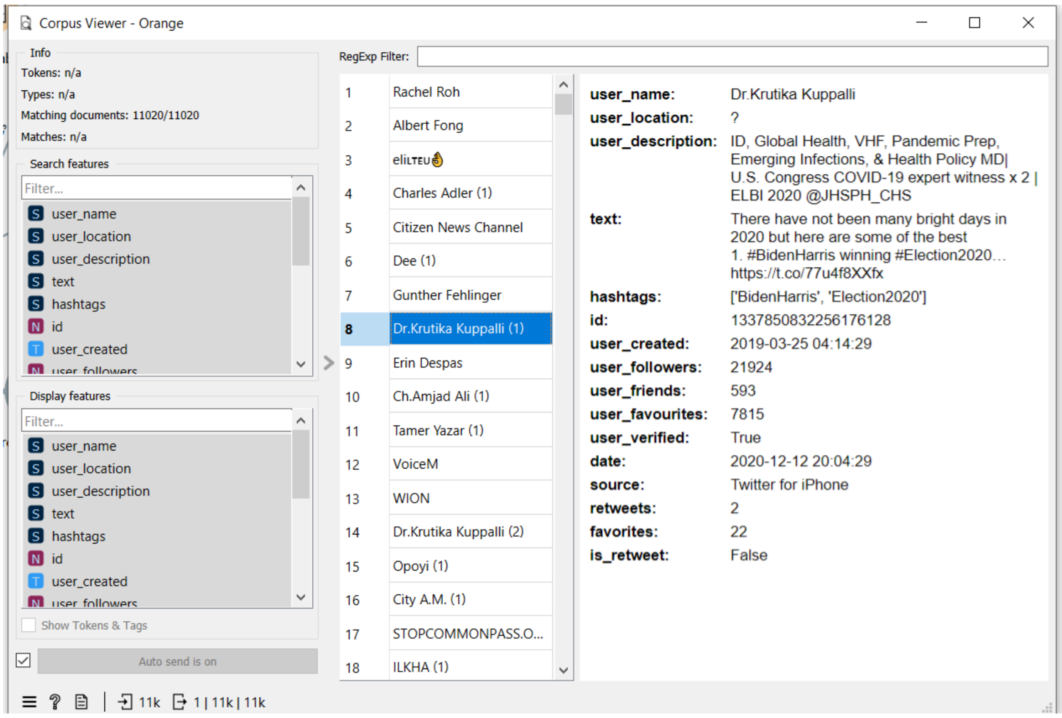


To solve the problem of sentiment analysis of vaccination tweets, the following steps can be taken:

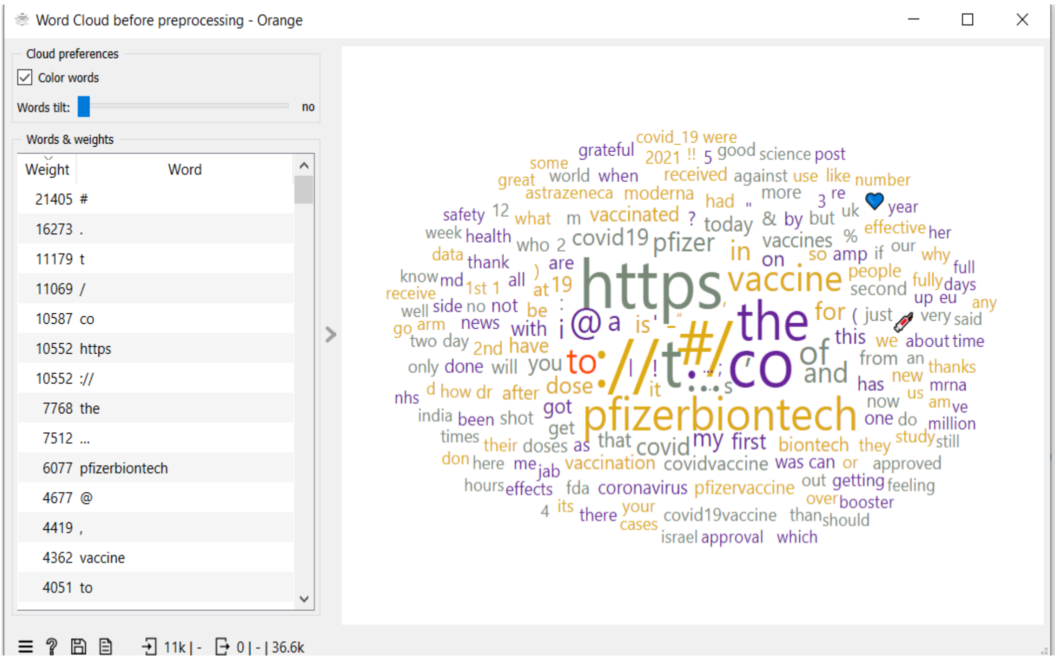
1. Import Dataset: Importing dataset using Corpus which allows to import text data from a file, we’re using vaccination tweets dataset from twitter which has tweets regarding how people feel about vaccination in covid times.



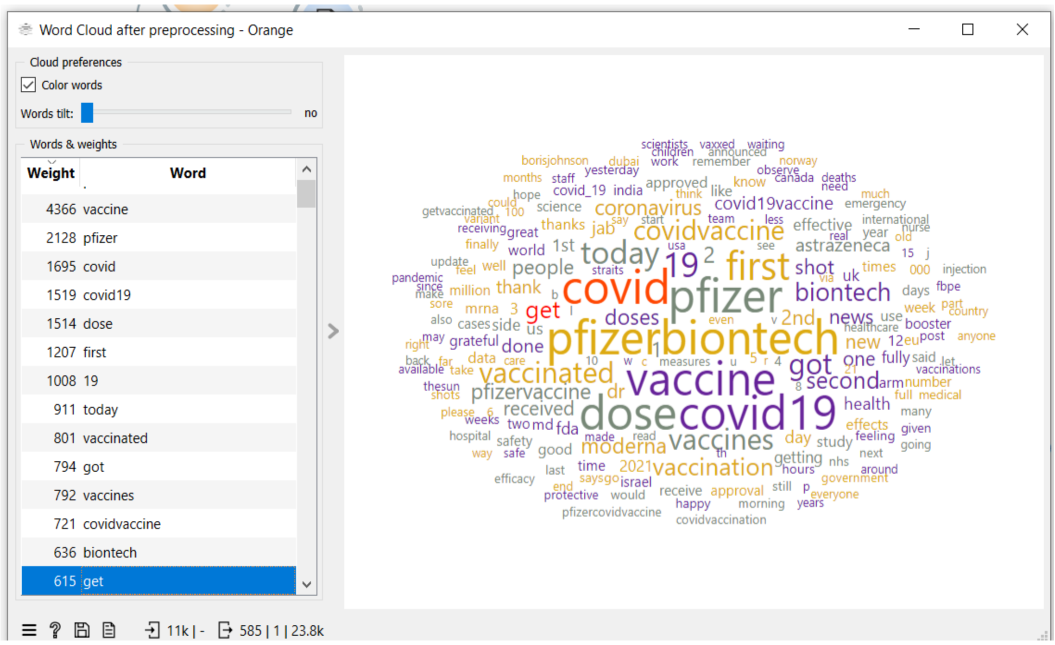
2. Corpus Viewer: The loaded corpus is represented visually by this widget. We can browse the corpus's documents, view their content, and look through any associated metadata. It enables us to choose particular documents for additional analysis and get a general understanding of the text data.



3. Word Cloud: This allows you to create word clouds from text data. Word clouds are graphical representations of text where the size of each word corresponds to its frequency or importance within the given text corpus.

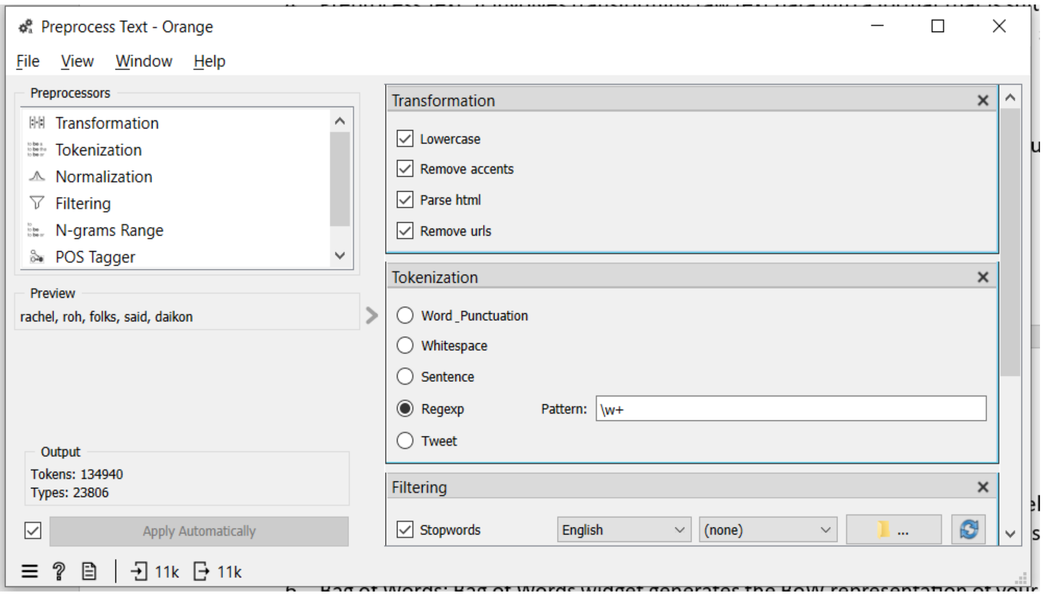


Before Preprocessing



After Preprocessing

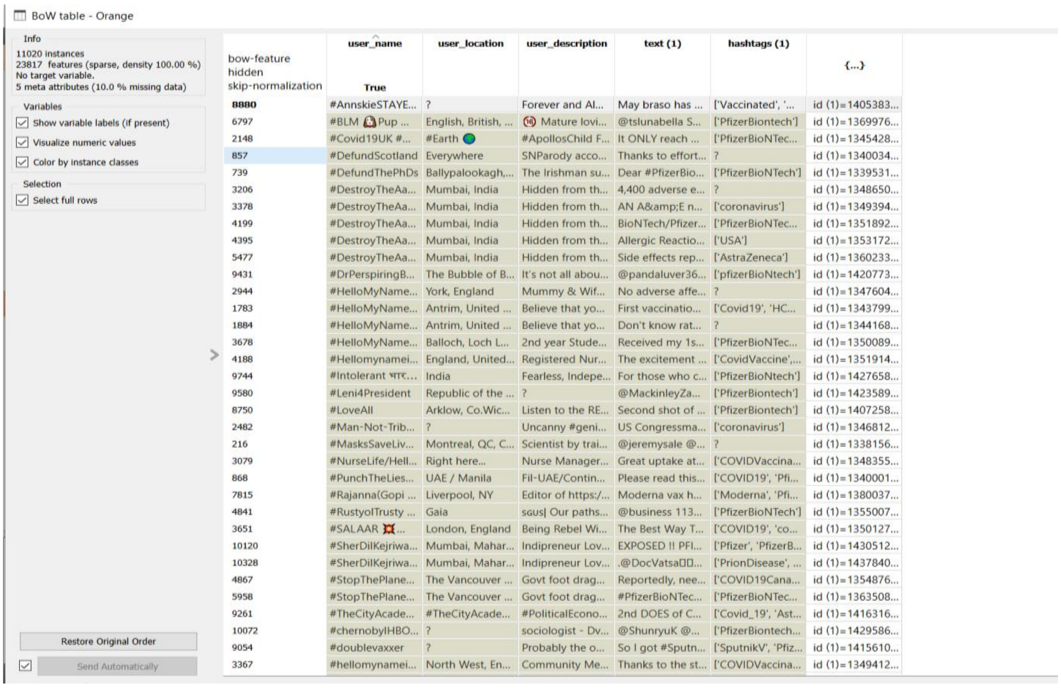
4. Preprocess Text: It involves transforming raw text data into a format that is suitable for analysis and modeling. Text preprocessing techniques are tokenization, stop word removal, lowercasing, etc.



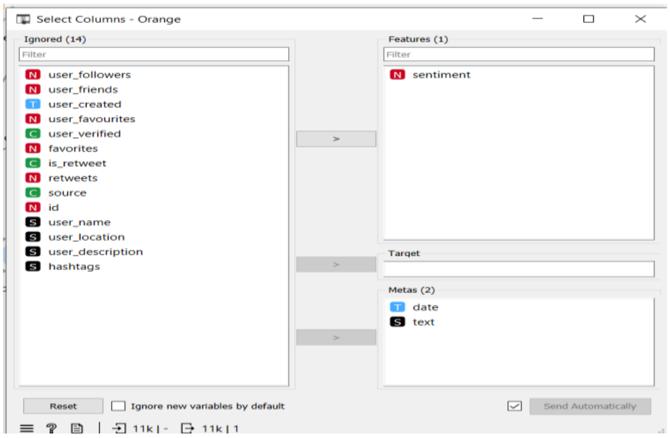
5. Feature Statistics: Feature statistics provide valuable insights into the distribution and characteristics of individual features or variables in a dataset. They help in understanding the data, identifying patterns, and making informed decisions during the data analysis process.

6. Bag of Words: Bag of Words widget generates the BoW representation of your text

data. Each document is transformed into a vector where the elements correspond to the frequency or presence of words or N-grams. The output can be connected to subsequent widgets for further analysis, such as clustering, classification, or topic modeling.

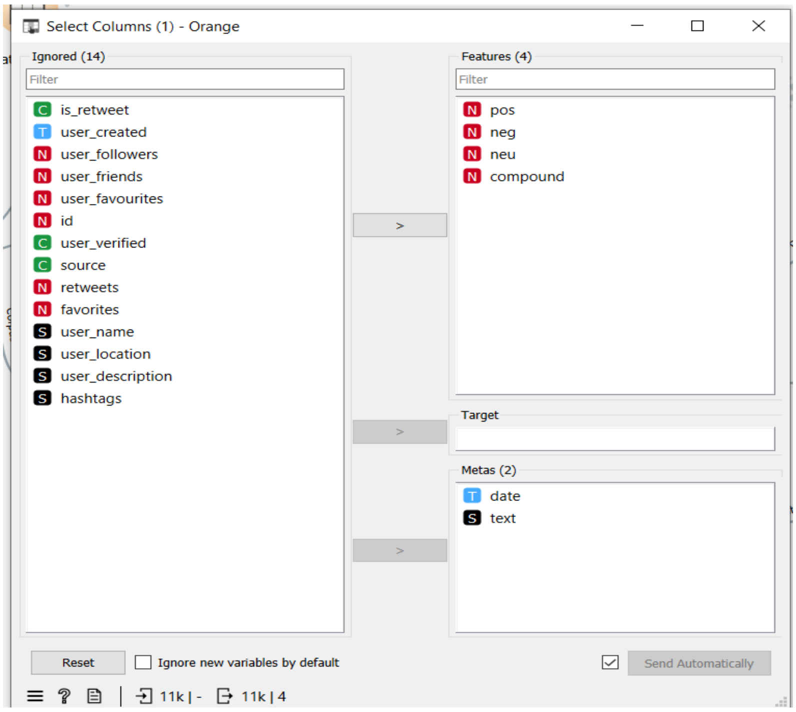


7. Liu Hu Sentiment Analysis: For each document or text snippet, the Liu Hu method computes the sentiment score by counting the occurrence of positive and negative sentiment words. It assigns a sentiment polarity based on the overall sentiment score:
   - If the count of positive sentiment words exceeds the count of negative sentiment words, the sentiment polarity is classified as positive.
   - If the count of negative sentiment words exceeds the count of positive sentiment words, the sentiment polarity is classified as negative.
   - If the counts are equal or there is no sentiment word detected, the sentiment polarity is classified as neutral.
8. Selecting Columns for Liu Hu analysis:



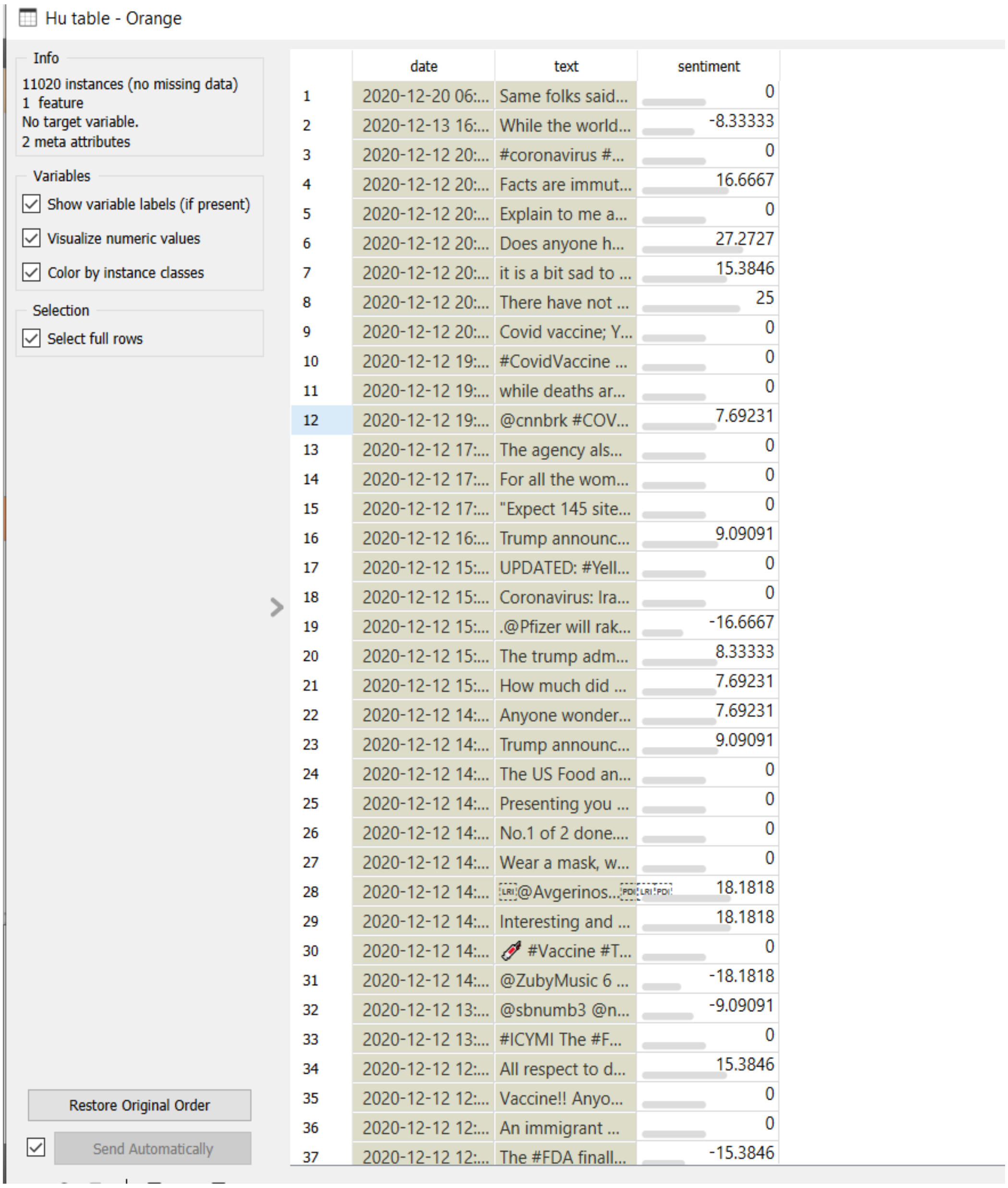
9. Vader Sentiment Analysis: Valence Aware Dictionary and Sentiment Reasoner method is a rule-based sentiment analysis technique specifically designed for social media text. VADER assigns sentiment scores to each word or phrase in the text based on their presence in the sentiment lexicon. It considers the sentiment intensity modifiers, negations, and contextual valence shifts to calculate a sentiment compound score. The compound score represents the overall sentiment polarity of the text, ranging from -1 (extremely negative) to +1 (extremely positive).

10. Selecting Columns for Vader analysis:



Result and Validation:

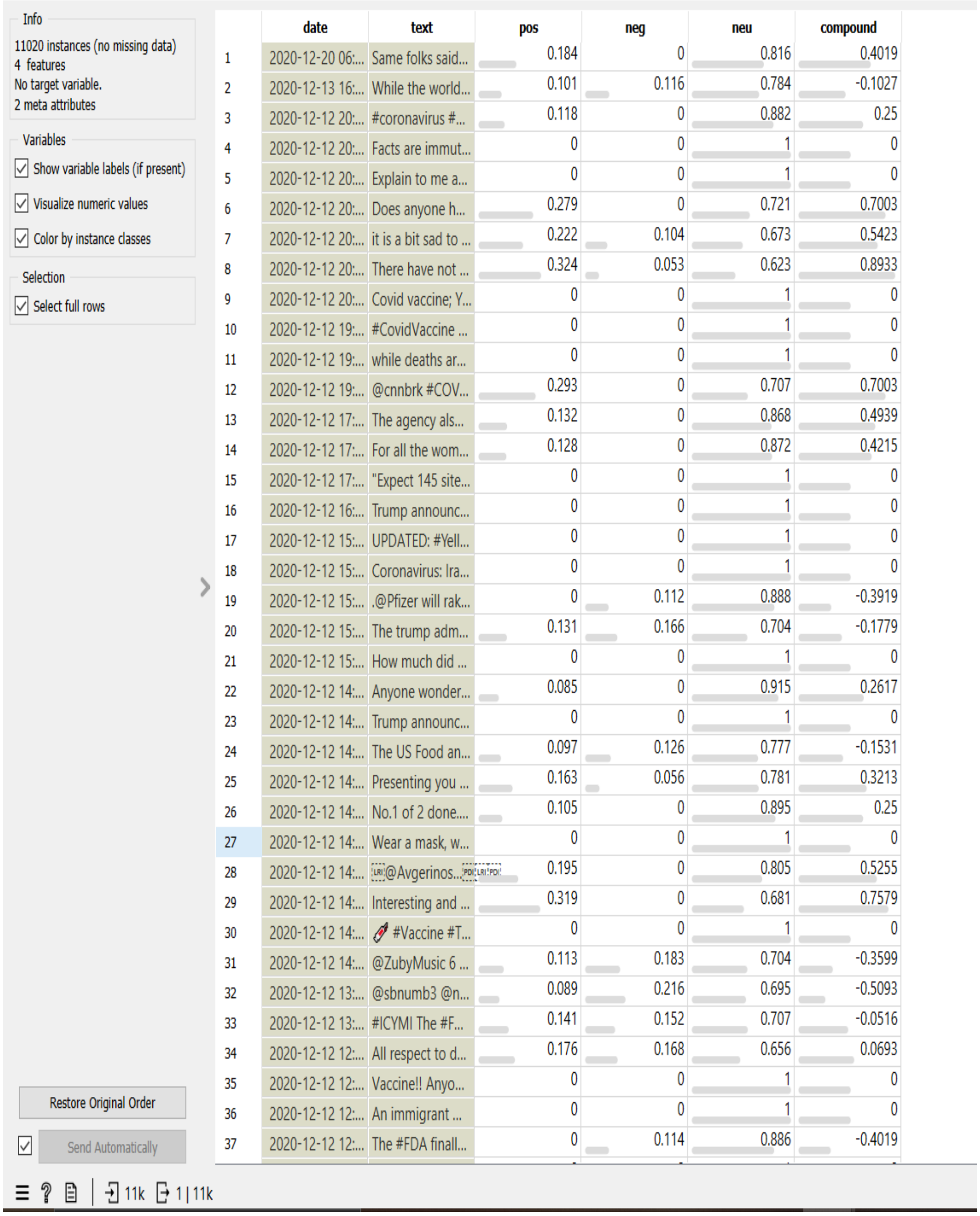
Liu Hiu Analysis Result:



	date	text	sentiment
1	2020-12-20 06:...	Same folks said...	0
2	2020-12-13 16:...	While the world...	-8.33333
3	2020-12-12 20:...	#coronavirus #...	0
4	2020-12-12 20:...	Facts are immut...	16.6667
5	2020-12-12 20:...	Explain to me a...	0
6	2020-12-12 20:...	Does anyone h...	27.2727
7	2020-12-12 20:...	it is a bit sad to ...	15.3846
8	2020-12-12 20:...	There have not ...	25
9	2020-12-12 20:...	Covid vaccine; Y...	0
10	2020-12-12 19:...	#CovidVaccine ...	0
11	2020-12-12 19:...	while deaths ar...	0
12	2020-12-12 19:...	@cnnbrk #COV...	7.69231
13	2020-12-12 17:...	The agency als...	0
14	2020-12-12 17:...	For all the wom...	0
15	2020-12-12 17:...	"Expect 145 site...	0
16	2020-12-12 16:...	Trump announc...	9.09091
17	2020-12-12 15:...	UPDATED: #Yell...	0
18	2020-12-12 15:...	Coronavirus: Ira...	0
19	2020-12-12 15:...	.@Pfizer will rak...	-16.6667
20	2020-12-12 15:...	The trump adm...	8.33333
21	2020-12-12 15:...	How much did ...	7.69231
22	2020-12-12 14:...	Anyone wonder...	7.69231
23	2020-12-12 14:...	Trump announc...	9.09091
24	2020-12-12 14:...	The US Food an...	0
25	2020-12-12 14:...	Presenting you ...	0
26	2020-12-12 14:...	No.1 of 2 done....	0
27	2020-12-12 14:...	Wear a mask, w...	0
28	2020-12-12 14:...	@Avgerinos...	18.1818
29	2020-12-12 14:...	Interesting and ...	18.1818
30	2020-12-12 14:...	💉 #Vaccine #T...	0
31	2020-12-12 14:...	@ZubyMusic 6 ...	-18.1818
32	2020-12-12 13:...	@sbnumb3 @n...	-9.09091
33	2020-12-12 13:...	#ICYMI The #F...	0
34	2020-12-12 12:...	All respect to d...	15.3846
35	2020-12-12 12:...	Vaccine!! Anyo...	0
36	2020-12-12 12:...	An immigrant ...	0
37	2020-12-12 12:...	The #FDA finall...	-15.3846

The final score of Liu Hu sentiment analysis shows the difference between the sum of positive and sum of negative words, normalized by the length of the document and multiplied by a 100. The final score reflects the percentage of sentiment difference in the document.

Vader Sentiment Analysis:



	date	text	pos	neg	neu	compound
1	2020-12-20 06:...	Same folks said...	0.184	0	0.816	0.4019
2	2020-12-13 16:...	While the world...	0.101	0.116	0.784	-0.1027
3	2020-12-12 20:...	#coronavirus #...	0.118	0	0.882	0.25
4	2020-12-12 20:...	Facts are immut...	0	0	1	0
5	2020-12-12 20:...	Explain to me a...	0	0	1	0
6	2020-12-12 20:...	Does anyone h...	0.279	0	0.721	0.7003
7	2020-12-12 20:...	it is a bit sad to ...	0.222	0.104	0.673	0.5423
8	2020-12-12 20:...	There have not ...	0.324	0.053	0.623	0.8933
9	2020-12-12 20:...	Covid vaccine; Y...	0	0	1	0
10	2020-12-12 19:...	#CovidVaccine ...	0	0	1	0
11	2020-12-12 19:...	while deaths ar...	0	0	1	0
12	2020-12-12 19:...	@cnnbrk #COV...	0.293	0	0.707	0.7003
13	2020-12-12 17:...	The agency als...	0.132	0	0.868	0.4939
14	2020-12-12 17:...	For all the wom...	0.128	0	0.872	0.4215
15	2020-12-12 17:...	"Expect 145 site...	0	0	1	0
16	2020-12-12 16:...	Trump announc...	0	0	1	0
17	2020-12-12 15:...	UPDATED: #Yell...	0	0	1	0
18	2020-12-12 15:...	Coronavirus: Ira...	0	0	1	0
19	2020-12-12 15:...	.@Pfizer will rak...	0	0.112	0.888	-0.3919
20	2020-12-12 15:...	The trump adm...	0.131	0.166	0.704	-0.1779
21	2020-12-12 15:...	How much did ...	0	0	1	0
22	2020-12-12 14:...	Anyone wonder...	0.085	0	0.915	0.2617
23	2020-12-12 14:...	Trump announc...	0	0	1	0
24	2020-12-12 14:...	The US Food an...	0.097	0.126	0.777	-0.1531
25	2020-12-12 14:...	Presenting you ...	0.163	0.056	0.781	0.3213
26	2020-12-12 14:...	No.1 of 2 done....	0.105	0	0.895	0.25
27	2020-12-12 14:...	Wear a mask, w...	0	0	1	0
28	2020-12-12 14:...	@Avgerinos...	0.195	0	0.805	0.5255
29	2020-12-12 14:...	Interesting and ...	0.319	0	0.681	0.7579
30	2020-12-12 14:...	💉 #Vaccine #T...	0	0	1	0
31	2020-12-12 14:...	@ZubyMusic 6 ...	0.113	0.183	0.704	-0.3599
32	2020-12-12 13:...	@sbnumb3 @n...	0.089	0.216	0.695	-0.5093
33	2020-12-12 13:...	#ICYMI The #F...	0.141	0.152	0.707	-0.0516
34	2020-12-12 12:...	All respect to d...	0.176	0.168	0.656	0.0693
35	2020-12-12 12:...	Vaccine!! Anyo...	0	0	1	0
36	2020-12-12 12:...	An immigrant ...	0	0	1	0
37	2020-12-12 12:...	The #FDA finall...	0	0.114	0.886	-0.4019

The final output of VADER sentiment analysis shows the positive, negative and neutral score in a particular statement or tweet.