Capstone Project - CIND820

Literature Review

Identifying Trends and Sentiment in Twitter Data

Name: Pallavi Thirunavukarasu

Student Number: 501114919

Supervisor: Dr Ceni Babaoglu

Date: February 14, 2022

Table of Contents

[Introduction 3](#_Toc95771342)

[Related Work 3](#_Toc95771343)

[Sentiment Analysis: 3](#_Toc95771344)

[Twitter Sentiment Analysis (TSA) 4](#_Toc95771345)

[Challenges: 5](#_Toc95771346)

[Methods for Twitter Sentiment Analysis: 5](#_Toc95771347)

[Feature Selection for Twitter Sentiment Analysis: 6](#_Toc95771348)

[Topic modelling and Sentiment Analysis on Twitter Data 7](#_Toc95771349)

[Project Methodology 7](#_Toc95771350)

[Data description 8](#_Toc95771351)

[Overview of the Dataset 9](#_Toc95771352)

[Missing Values 10](#_Toc95771353)

[Variables 10](#_Toc95771354)

[References 14](#_Toc95771355)

# Introduction

Social media has become a rich corpus of text and a communication channel that reflects the events happening in the community and opinions/reviews of products. It has become important to identify the trends in social media. Trends may be broadly described as the top noteworthy topics that are discussed.

Sentiment detection and classification is also of great importance in the analytics of social media data. There are several practical applications for identifying trends and sentiments in social media data such as in recognising consumer sentiments towards a product, understanding public opinion on government policies, financial predicting, etc.

Leveraging social media content presents multiple challenges in interpreting the data due to the presence non- standard words(slang), use of symbols, abbreviations and emoticons.

Substantial amount of research work has been done in this area and there is ongoing research to address the unique challenges of mining social media data.

# Related Work

## Sentiment Analysis:

The basic idea of sentiment analysis is to take a text and classify that text as positive, negative or neutral. Sentiment mining can be done at different levels **[1]**.

* Document level sentiment classification: This level classifies the entire document as positive, negative or neutral
* Sentence level sentiment classification: This level classifies each sentence as positive, negative or neutral
* Aspect or Entity level sentiment classification: This level identifies the aspect in each sentence and then classifies the aspect as positive, negative or neutral.

Liu and Zhang **[2]** defined an opinion as a quintuple, consisting of an entity, aspect of the entity, orientation(sentiment) of the opinion of the user about that aspect, the user and the time. For example, if the following review was posted by user user1, on 14.02.2022,

The camera quality of my new iPhone13 is great.

The entity for which the opinion is expressed is the iPhone 13, the aspect is the camera, the sentiment is positive, the user is user1 and the time is 14.02.2022. The quintuple would be (iPhone13, camera, positive, user1, 14.02.2022). All the five components are considered as essential to sentiment analysis.

Given a collection of opinionated documents, there are 5 steps to find all the quintuples(opinions) in the document.

* Extracting all the entities in the document
* Extracting all the aspects of the entities
* Extracting the opinion holder and the time
* Determining if the opinion of the aspect is positive, negative or neutral
* Producing all the possible quintuples for this document

Immense quantities of user created web based and social networking content has several applications. Some examples include:

* Stock prediction – in this study **[3]** conducted in 2007, stock discussions on message boards were collected using web scraping programs, various classifier algorithms were used to classify the sentiments for the messages, a voting mechanism was used on the classifier outcomes to determine the sentiment index. Relationship of the sentiment index to the stock values was analysed.
* Government Policy making – in this study **[4],** conducted in 2015, sentiment analysis was used to determine how opinion of local government social media posts influences citizen involvement on Twitter. Relationship of positive sentiment expressed on social media to citizen’s digital participation was studied
* Political results – in this study **[5],** conducted in 2010, text analysis software was used to conduct a sentiment analysis of 100,000 messages which mentioned a politician or political party. An analysis of the tweets political sentiments showed that public sentiment is close to the political positions of the parties and politicians.

## Twitter Sentiment Analysis (TSA)

Twitter is considered as one of the most popular microblogging sites. As of Q4 of 2020, twitter has 192M monetizable daily active usage\*. Each tweet is a single message posted on twitter that can be 280 characters long. A user can register with the platform to post the tweets. Each user can follow other users. Retweets are tweets that are re-distributed by other users.

Twitter is considered as one of the largest datasets of user generated content due to the large number of users, easy access to data/downloading of published posts.

Twitter is a dynamic forum with continuous updates and also has unique challenges, hence a novel approach is required for Twitter sentiment analysis.

*\* https://twitter.com/about*

### Challenges:

Some of the challenges associated with Twitter data sentiment analysis are discussed in detail in this survey **[6]**.

1. Text Length – The short length of the twitter messages makes the sentiment analysis very different from the sentiment analysis of blogs, movie reviews or product reviews. Bermingham and Smeaton conducted a study **[7]**, to check if the brevity of text was an advantage by comparing Support Vector Machine (SVM) and Multinomial Naïve Bayes. They concluded that SVM performed better in Twitter Sentiment Analysis.
2. Incorrect English – Tweets have peculiarities of use of slang, emphatic lengthening, emphatic uppercasing, abbreviations. Preprocessing and cleanup of these inaccuracies is required
3. Topic Relevance – To classify tweets and to take into account the topic relevance of tweets, the keywords and hash tags in the tweets are considered
4. Data Sparsity - Tweets contain a lot of noise due to misspellings, slang and this results in data sparsity.
5. Stop Words – Stop words that are normally filtered out in text processing such as ‘like’ may usually have discriminatory powers when it comes to tweets. Saif et al. [2014] **[8]** work focussed on comparing the different methods of stop word removal on 4 different datasets and analysing the impact on the classifiers and the data sparsity.
6. Tokenization – Tokenization by white spaces may not work well for Twitter. Some studies have developed twitter specific tokenization.
7. Multilingual content – Tweets can contain mixed languages even in a single tweet, tweets can be done in several languages. Multi lingual classifiers have been developed for this.
8. Multimodal – Tweets can also videos, images. Extracting features from multimodal content is ongoing research.

### Methods for Twitter Sentiment Analysis:

There are several approaches to Twitter Sentiment Analysis – Supervised Machine Learning methods, Ensemble methods, Lexicon Based, Hybrid and Deep Learning based.

Supervised Learning – Twitter data is easily available and also in large quantities, yet there is a challenge in getting labelled data for sentiment analysis which makes the use of Supervised Learning methods a challenge. In one of the early Twitter Sentiment Analysis done by Go et al., 2019**, [9]** the approach was to collect the tweets with emoticons, then use the emoticons as a kind of noisy labels. The labelled tweets where then stripped of the emoticons and different features like unigrams, unigrams and bigrams, unigrams with part of speech tagging were built. The labelled data was now used to built classifier models using Support Vector Machine, Naïve Bayes and Max Entropy algorithms.

Ensemble Methods – In this study [**10**] , the mined tweets were subjected to an ensemble classification method. Multiple classifiers were trained to obtain better predictive power. The results of the different classifiers were averaged to obtain the sentiment predictions

Lexicon based – In this study **[11],** tweets were collected using keywords, lexicon or dictionary-based approach was used to classify the sentiments. The words from the tweet were matched with the words from the dictionary. If the dictionary word was positive the tweet word was tagged positive. If the dictionary word was negative the tweet word was tagged negative. If the dictionary word was not positive or negative, the tweet word was tagged as neutral. A sentiment score was calculated to classify the tweet. Sentiment mining was also done at the aspect level.

Hybrid methods – Balage et al [**12]**, built a hybrid system for tweet classification. The system was built with 4 main components. A normalization component that was used for correcting and normalizing the collected tweets. The three other components were created as a pipeline - a rule-based classifier, a lexicon-based classifier and a machine learning based classifier. The tweet was passed through each component in the pipeline sequentially until the tweet was classified with a certain confidence. This hybrid pipeline of classifiers was found to give more accuracy than the Support Vector Machine classifier.

Deep learning based – Several deep learning-based approaches to sentiment analysis are underway now.

### Feature Selection for Twitter Sentiment Analysis:

Features such as automatic part-of-speech tags and resources such as sentiment lexicons are useful for sentiment analysis in other domains. However, feature selection for microblogging sites like twitter is not a trivial task.

Kouloumpis et al, 2011 **[13],** conducted a study to check theutility of linguistic features for determining sentiments in microblogging sites such as twitter. They used three datasets in this study. The Hash Tagged dataset HASH, is a subset of the Edinburgh twitter corpus. The top 15 hashtags were first identified from the Edinburg twitter corpus, then the hashtags with at least 1000 tweets were taken. Emoticon dataset EMOT, was created by Go, Bhayani, and Huang for a project at Stanford University by collecting tweets with positive and negative emoticons. The iSieve dataset contains a set of 4000 tweets which were collected and annotated by the iSieve corporation.

The data was preprocessed – tokenization, normalization, part of speech tagging was done. Abbreviations and emoticons were converted to individual tokens. Intensifiers such as upper case were identified and converted to lowercase. Emphatic lengthening was identified and repeated characters were replaced with single characters.

Variety of features like unigrams, bigrams, part of speech features, lexicon-based features were used in the classification. Microblogging domain specific features were also included. Accuracy of the classifiers for different feature combinations was analyzed. Part of Speech features were found to have a negative impact on the twitter sentiment analysis.

Agarwal et al [14**]** [2011], conducted a study in which they compared different combinations of feature sets. The dataset consisted of manually annotated twitter data that was collected. Two models were built. One was a binary class classifier that classified the tweets as positive or negative. The second was a multi class classifier that classified the tweets as positive, negative or neutral. Three types of models were built, a unigram model, a feature-based model and a tree-based model. Using the unigram model as baseline, the study compared combination of models with different combinations of features. The combinations were found to outperform the baseline unigram model. Contrary to the previous study’s finding, this study found that part of speech tagging increased accuracy. Twitter specific features did not impact the outcome.

## Topic modelling and Sentiment Analysis on Twitter Data

In the past tweets have been analyzed at the time of disasters and spread of diseases. For example, this study on Zika outbreak**[17]**  used twitter data, analyzed how timely topics like the Zika virus are addressed on social media. The study examined the emerging themes during a Center for Disease Control (CDC) hosted live Twitter chat and used text mining to evaluate the public’s concerns about the Zika virus and the CDC’s response to the public’s questions.

Similar studies have been conducted at the time of Ebola outbreak **[18],** Hurricane Irma **[19]** and Japanese earthquake of 2011**[20]**.

In the present pandemic situation, many studies have been conducted using twitter data to gauge public opinions and identify trends.

In this study **[15]** , Using the Twitter API and tools, tweets which contained a predefined set of keywords related to covid19 were downloaded from twitter. The top topics in these tweets was identified using the Latent Dirichlet Allocation (LDA). Topics were grouped together as themes manually. The sentiment analysis was done. The mean sentiment was found to be positive for ten topics and negative for two topics.

In this study **[16] ,** Data from Twitter was collected using a streaming application over a period of time. Tweets were collected with keywords related to covid. The data was cleaned and preprocessed (removal of stop words, removal of URLs and hyperlinks, etc.). The word cloud of frequently used words was created. Sentiment analysis was done using the National Research Council (NRC) lexicon, Total of 8 emotions were evaluated based on the lexicon. Latent Dirichlet Allocation (LDA) was applied to fit a topic model. Coherence was used as an evaluative measure to choose the best number of topics for the topic model. The tweet trends and results of the sentiment analysis were visualized.

## Project Methodology

The purpose of this project is to apply Text Mining and Sentiment Analysis using Natural Language Processing techniques on tweets collected from Twitter with the covid19 hashtag. The purpose of this project is to propose a framework for intelligent analysis of the covid related Twitter data by extracting the semantic topics and the sentiments along those topics and presenting the results in an easy-to-follow visual format.

The following is the methodology for the project.

1. The Twitter API will be used to download the English language tweets with the covid19 and related hashtags. [At present for the obtaining initial results the publicly available dataset of tweets collected with the covid 19 hashtag - <https://www.kaggle.com/gpreda/covid19-tweets> , will be used].
2. The dataset will be cleaned and preprocessing techniques will be applied.
3. Using the different topic modelling approaches like Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorizations (NMF) and Latent Semantic Analysis (LSA) the top topics from the tweet dataset will be identified
4. The tweets will be grouped into clusters based on the topics identified in step 3
5. The feature selection for sentiment analysis will be done
6. The sentiments expressed in the tweets within each topic cluster will be identified
7. The opinion impact of the sentiments expressed by topic, sentiment will be visualized effectively

Figure : Project Methodology

## Data description

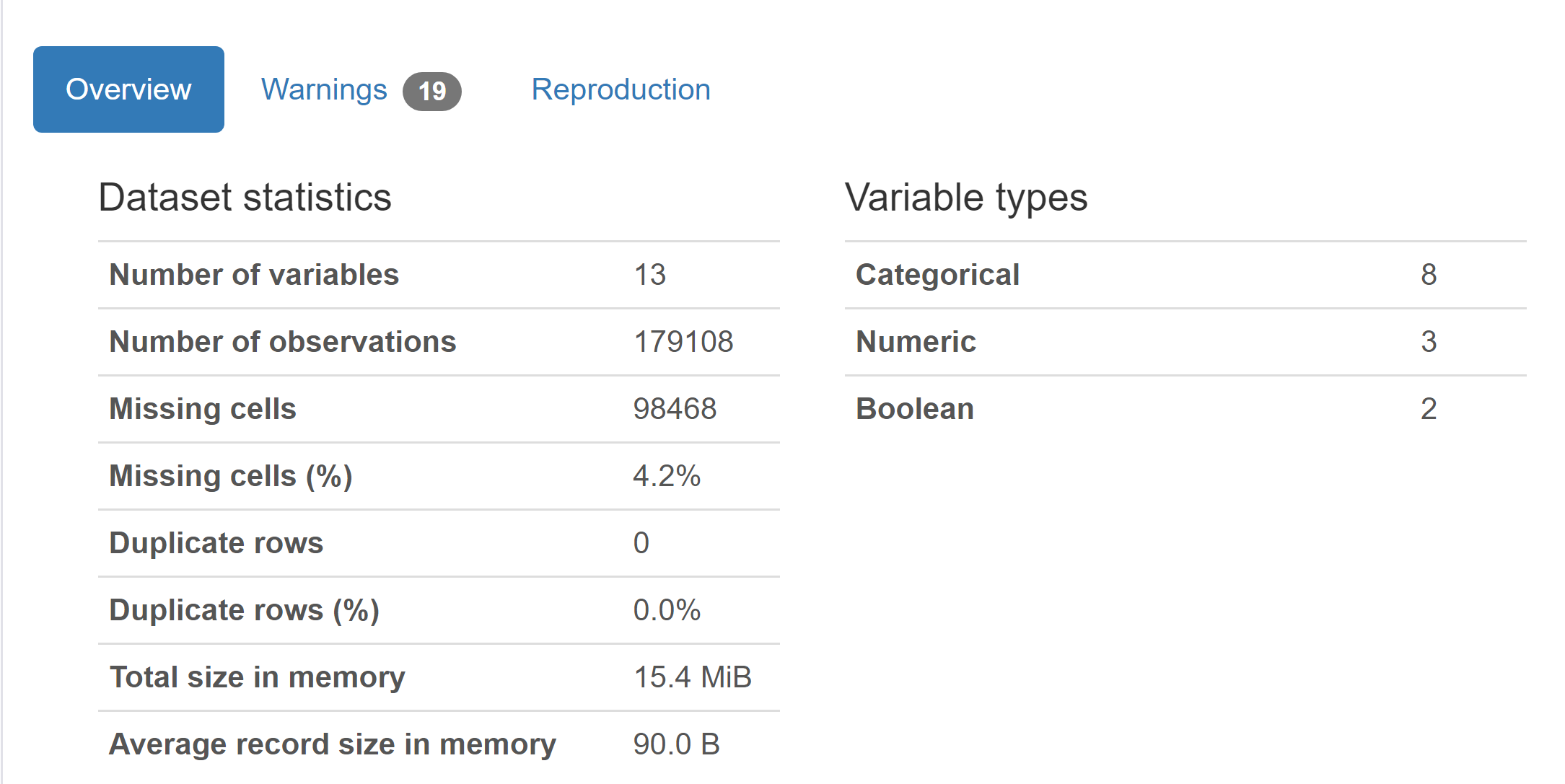
The dataset to be used for obtaining the initial results is the dataset available at - <https://www.kaggle.com/gpreda/covid19-tweets>. These tweets were collected using Twitter API and a Python script. A query for this high-frequency hashtag (#covid19) was run on a daily basis for a certain time period, to collect a larger number of tweet samples.

This dataset consists of tweets in English covering dates from 2020-02-29 to 2020-07-24. The tweets were downloaded from no specific region and are global. Retweets are not included in this dataset.

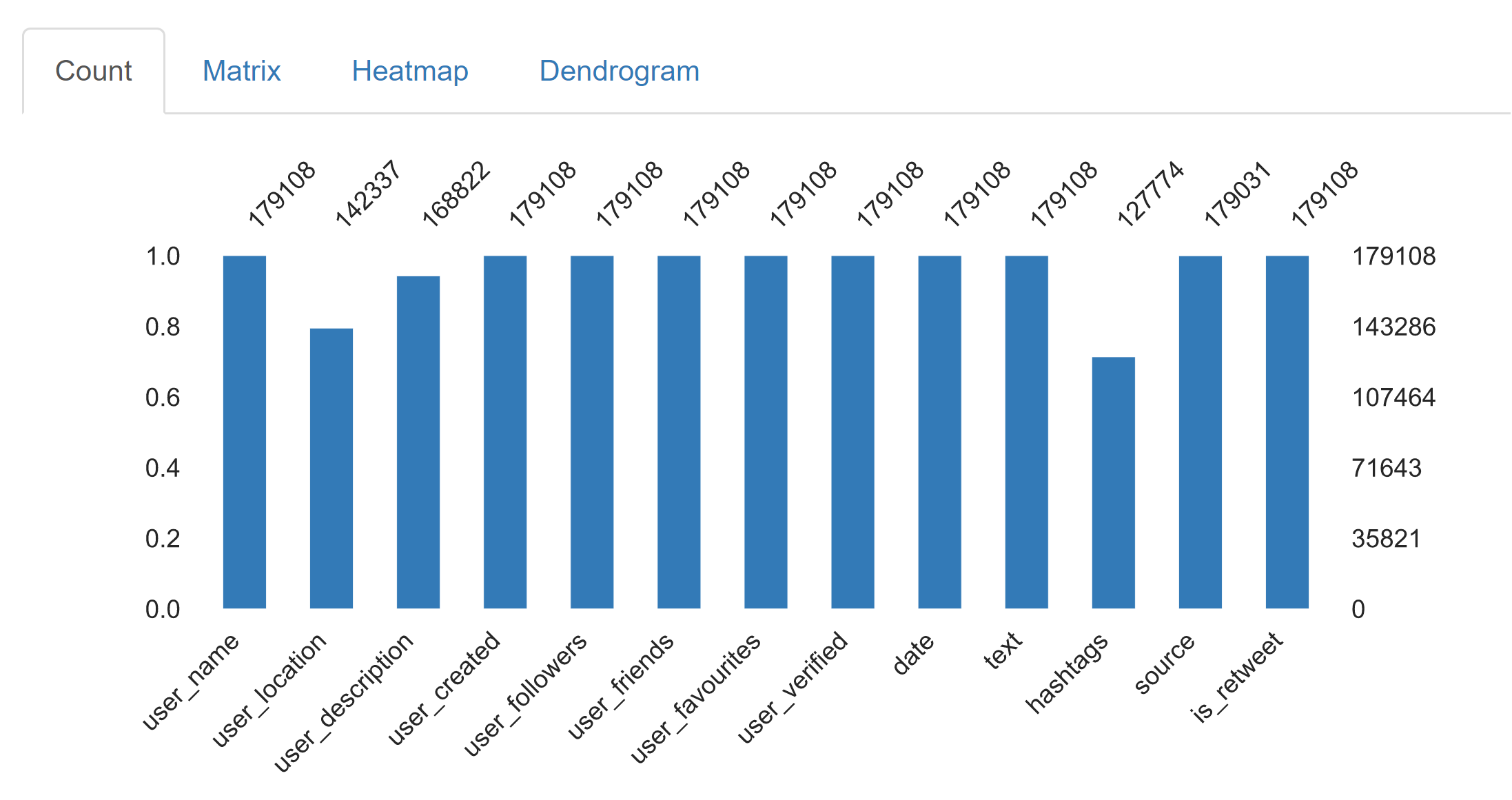
There are 13 columns in this dataset:

1. user\_name – Name of the user on twitter
2. user\_location - Location of the user
3. user\_description – Description of the user on twitter
4. user\_created – When the user was created
5. user\_followers – Number of followers of this user
6. user\_friends – Number of friends of this user
7. user\_favourites – Number of favourites of this user
8. user\_verified – Is the user verified
9. date – Tweet date
10. text -Text of the tweet
11. hastags – List of hash tags
12. source – Source of the tweet
13. is\_retweet – Tweet is a retweet

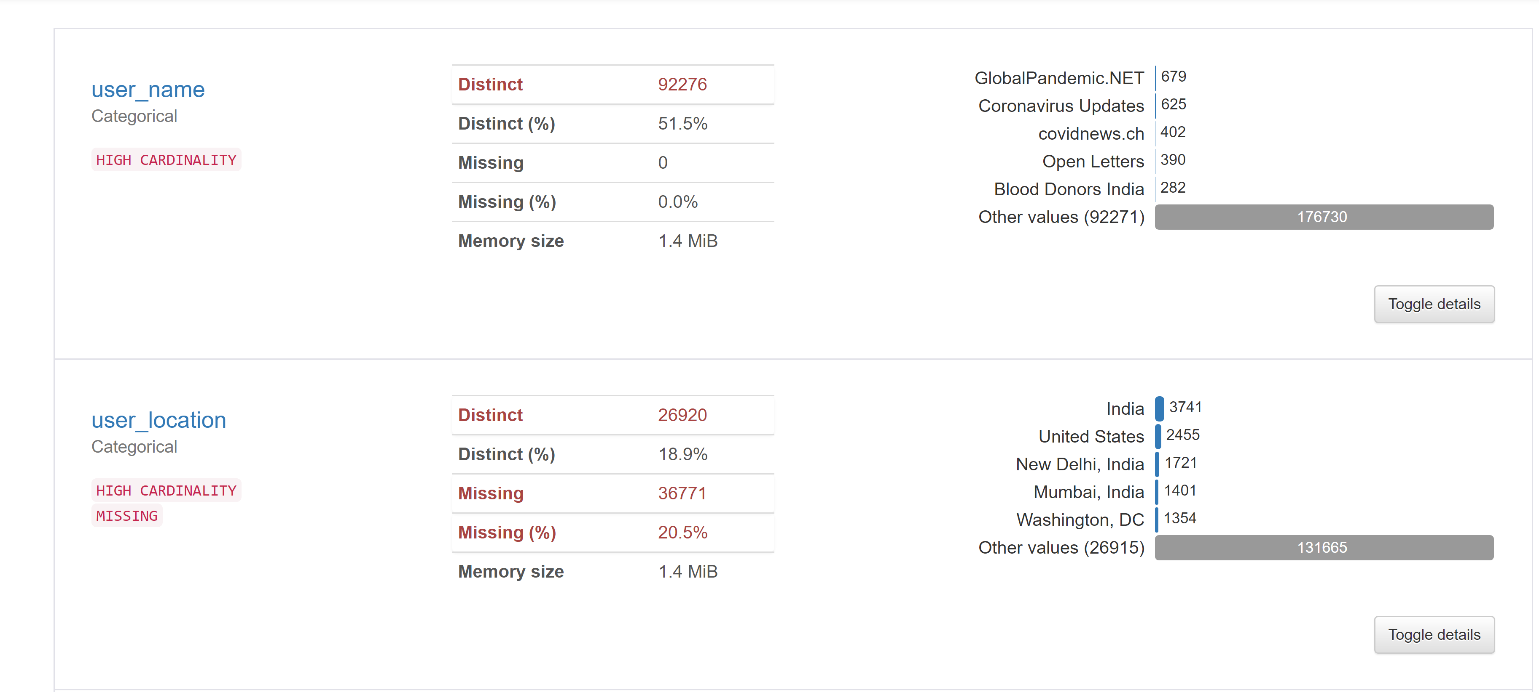
### Overview of the Dataset

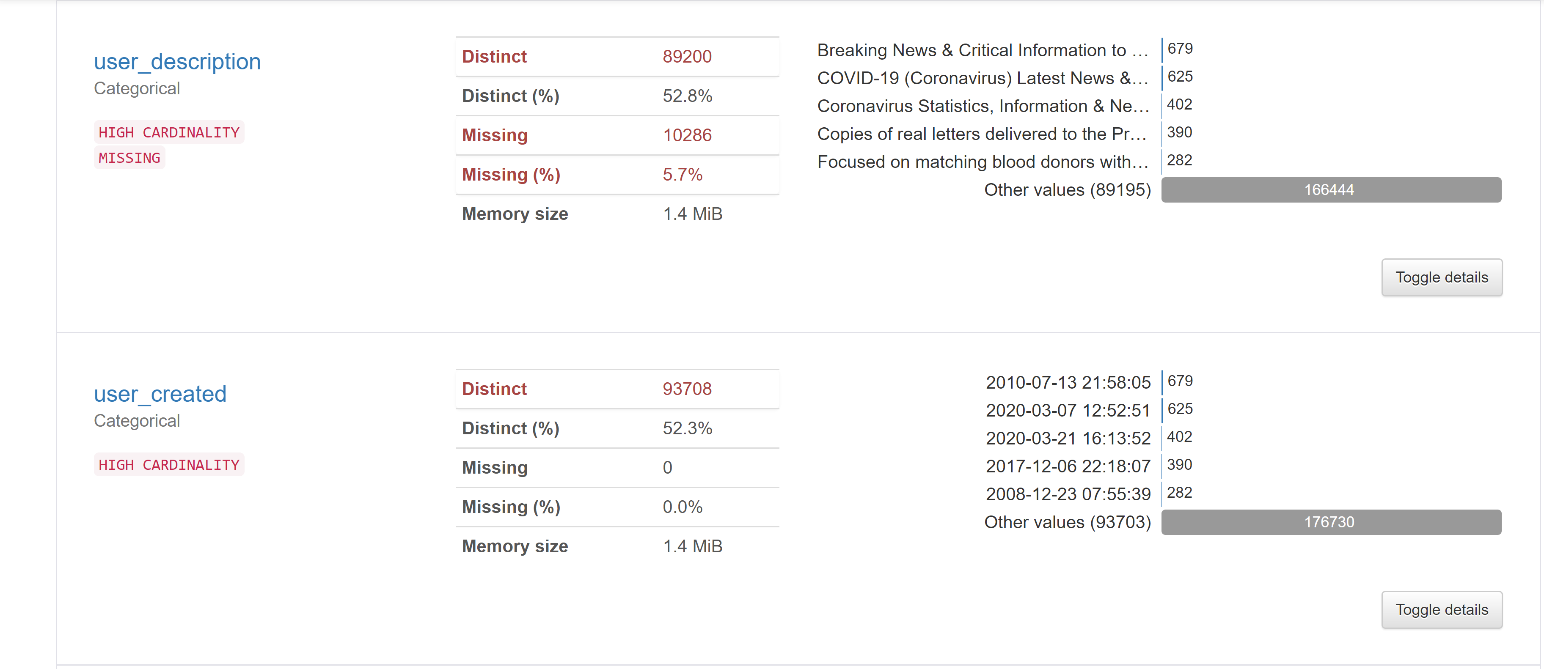


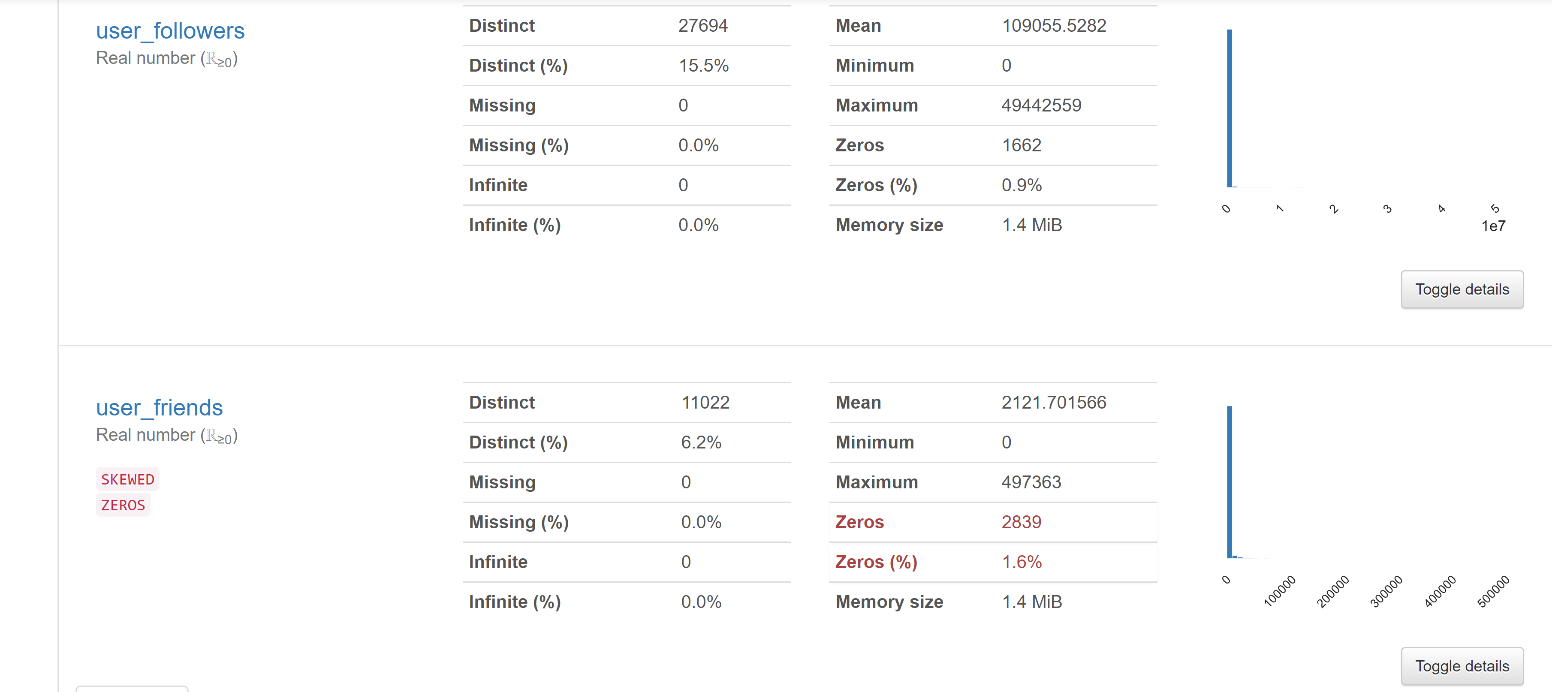
### Missing Values

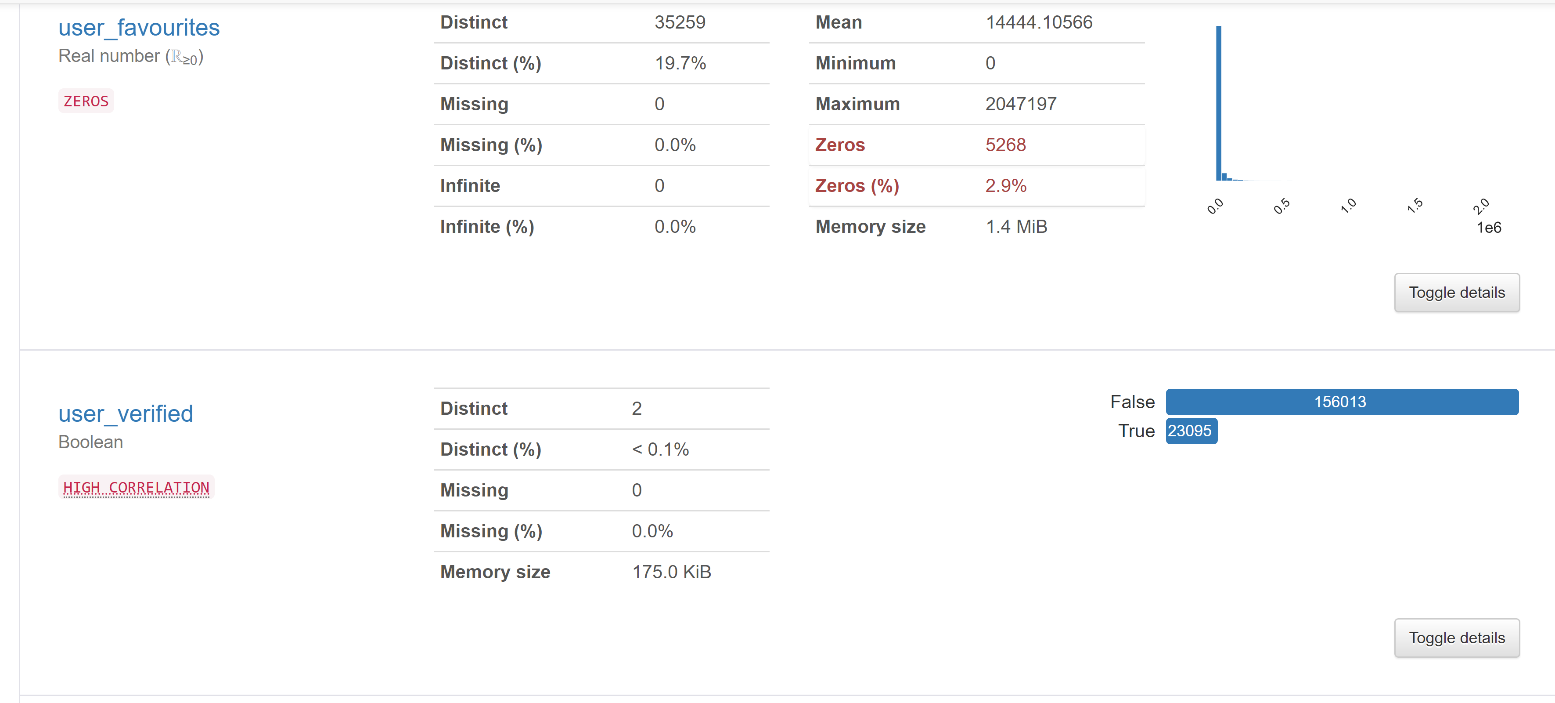


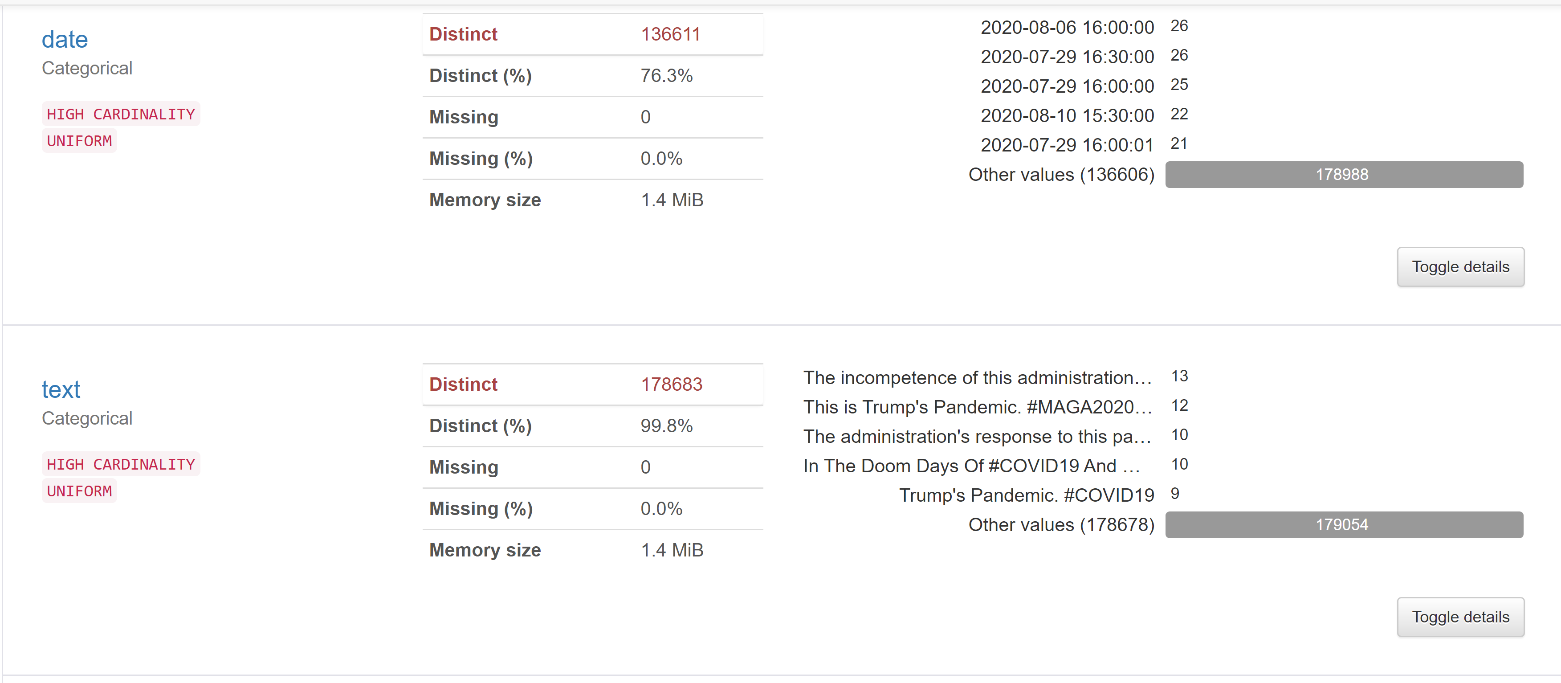
### Variables

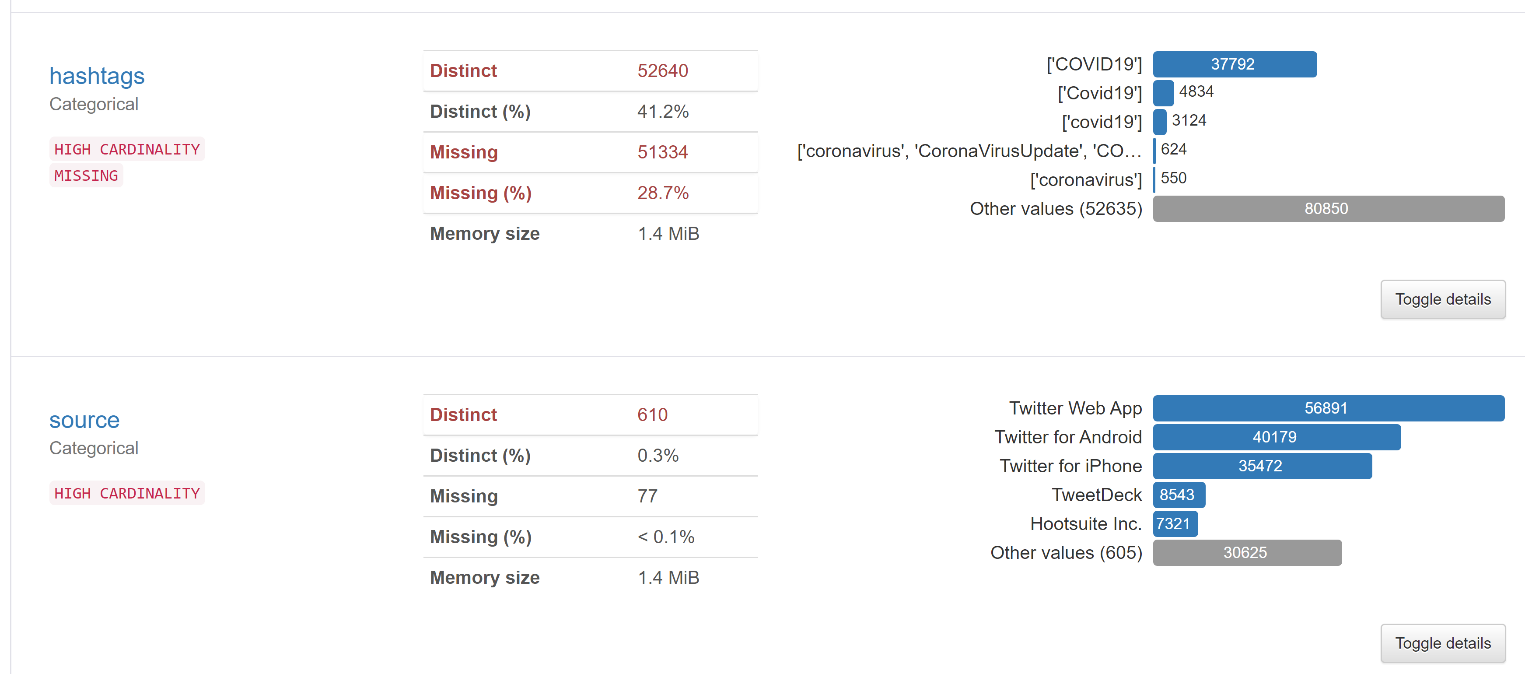




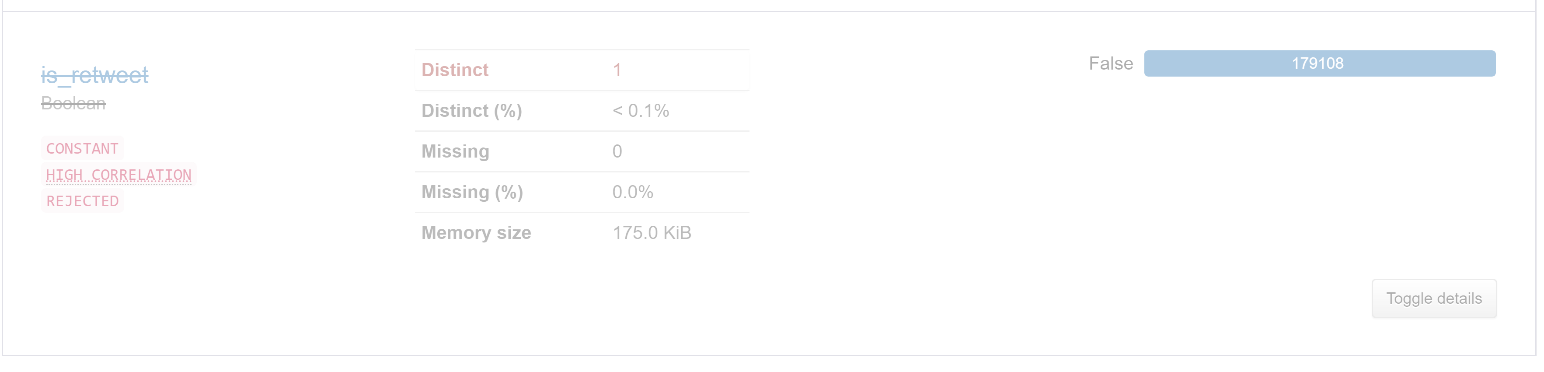








Retweets are removed from the dataset, so there is only one distinct value in this column.



# References

[1] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[2] Liu B., Zhang L. (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. <https://doi.org/10.1007/978-1-4614-3223-4_13>

[3] [**Sanjiv R. Das**](https://pubsonline.informs.org/action/doSearch?text1=Das%2C+Sanjiv+R&field1=Contrib), [**Mike Y. Chen**](https://pubsonline.informs.org/action/doSearch?text1=Chen%2C+Mike+Y&field1=Contrib), (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53(9):1375-1388. <https://doi.org/10.1287/mnsc.1070.0704>

[4] [Staci M.Zavattaro](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)[a](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)[P. EdwardFrench](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)[b](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)[Somya D.Mohanty](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)[c](https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000337" \l "!)A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement. <https://doi.org/10.1016/j.giq.2015.03.003>

[5] Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and social media*, *4*(1), 178-185. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14009>

[6] Anastasia Giachanou, Fabio Crestani, Like It or Not: A Survey of Twitter Sentiment Analysis Methods, [ACM Computing SurveysVolume 49Issue 2](https://dl.acm.org/toc/csur/2017/49/2)June 2017 Article No.: 28pp 1–41 <https://doi.org/10.1145/2938640>

[7] Adam Bermingham, Alan. F Smeaton, Classifying sentiment in microblogs: is brevity an advantage? [CIKM '10: Proceedings of the 19th ACM international conference on Information and knowledge management](https://dl.acm.org/doi/proceedings/10.1145/1871437)October 2010 Pages 1833–1836 <https://doi.org/10.1145/1871437.1871741>

[8] [Saif, Hassan](http://oro.open.ac.uk/view/person/hs4589.html); [Fernández, Miriam](http://oro.open.ac.uk/view/person/mff39.html); [He, Yulan](http://oro.open.ac.uk/view/person/yh375.html) and [Alani, Harith](http://oro.open.ac.uk/view/person/ha2294.html) (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.*, pp. 810–817. [**http://lrec2014.lrec-conf.org/en/**](http://lrec2014.lrec-conf.org/en/)

[9] Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification Using Distant Supervision*. Technical Report. Standford. <https://www-cs-faculty.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>

[10] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques," *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 2015, pp. 169-170, [Doi: 10.1109/ICOSC.2015.7050801](DOI).

[11] P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 2017, pp. 211-216, Doi: [10.1109/ICDMAI.2017.8073512](DOI).

[12] P. Balage Filho and T. Pardo, "NILC\_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (\* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568-572. <https://aclanthology.org/S13-2095.pdf>

[13] Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 538-541. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14185

[14] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM’11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 30–38. <https://aclanthology.org/W11-0705.pdf>

[15] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z, Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study, J Med Internet Res 2020;22(4): e19016, Doi: [10.2196/19016](https://doi.org/10.2196/19016)

[16] Boon-Itt S, Skunkan Y Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study, JMIR Public Health Surveillance 2020;6(4): e21978; Doi: [10.2196/21978](https://doi.org/10.2196/21978" \t "_blank)

[17] Glowacki E.M., Lazard A.J., Wilcox G.B., Mackert M., Bernhardt J.M. Identifying the public’s concerns and the Centers for Disease Control and Prevention’s reactions during a health crisis: An analysis of a Zika live Twitter chat. Am. J. Infect. Control, 44 (12) (2016), pp. 1709-1711, [10.1016/j.ajic.2016.05.025](https://doi.org/10.1016/j.ajic.2016.05.025)

[18] Morin C., Bost I., Mercier A., Dozon J.-P., Atlani-Duault In**formation circulation in times of Ebola: Twitter and the sexual transmission of Ebola by survivors;** PLoS Curr., 1 (10) (2018), [10.1371/currents.outbreaks.4e35a9446b89c1b46f8308099840d48f](https://doi.org/10.1371/currents.outbreaks.4e35a9446b89c1b46f8308099840d48f)

[19] Miyabe M., Miura A., Aramaki E. Use trend analysis of Twitter after the great east Japan earthquake. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, Association for Computing Machinery (2012), pp. 175-178.

[20] Lachlan K., Xu Z., Hutter E., Adam R., Spence P. A little goes a long way: Serial transmission of Twitter content associated with hurricane Irma and implications for crisis communication. J. Strategy. Innov. Sustain., 14 (1) (2019), [10.33423/jsis. v14i1.984](https://doi.org/10.33423/jsis.v14i1.984)