Mining of Social Data Lab: Warmup

Datasheet for: Twitter/Network Dataset[^1]

Authors: Anand Mathew M S, Finn Vaughankraska

Related Files:

./data/accounts.tsv ./data/tweets.dat

Version 1.0 2024-11-05

Motivation

For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* => The network dataset is created as part of the course 'Mining of Social Data' at Uppsala University. The main purpose of this dataset is to further the learning of the authors as part of the course. It also serves as a source of information to analyze visual persuasion on Twitter.

The main learning outcomes from creating the dataset are as follows:

- Generate datasets to be used for other learning activities during the course.
- Familiarise with typical data formats.
- Identify and reflect on data quality and validity issues.
- Practice the documentation of social data.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? => The dataset was created by Anand Mathew M S and Finn Vaughankraska as part of the course 1DL465 'Mining of Social Data' at Uppsala University under the instruction set presumably created by Prof. Matteo Magnani.

Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.* => Since the dataset was created as part of an academic course, there is no funding involved to the authors. However, since both the authors are fee-paying students enrolled in this programme, an argument could be made that the dataset is funded by the authors.

The original data was funded as part of the *PolarVis* project. The supporters of PolarVis are listed as:

> Project PolarVis is supported by FORTE, the Swedish Research Council for Health, Working Life and Welfare; Uddannelses - og Forskningsstyrelsen, the Danish Agency for Higher Education and Science; the National Research, Development and Innovation Office Hungary; and FWF, the Austrian Science Fund under CHANSE ERA-NET Co-fund programme, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement

Any other comments? The data for the dataset was initially collected as part of the *PolarVis* project which studies networked visual persuasion in and around climate movements in Europe. The link to the project can be found here: [[https://polarvis.github.io/]].

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.* => The dataset comprises of Tweets data as well as Account data which are represented as two different elements.

The Tweets data comprise of the following attributes:

1. **text**: The main content of the tweet, including any retweet mentions, hashtags, URLs and other text content.
2. **entities**: Contains various subfields to identify special elements within the tweet text.
   - **mentions**: Lists users mentioned in the tweet.
   - **hashtags**: Lists the hashtags used within the tweet.
   - **annotations**: Gives contextual metadata on keywords used within the tweet including *probability*, *type* (Place, Other etc.), and *normalized_text*
   - **urls**: Lists the URLs used within the tweet. It includes the *url*, *expanded_url* and *display_url*.
3. **possibly_sensitive**: Boolean flag indicating whether the content of the tweet might be sensitive.
4. **edit_history_tweet_ids**: List of IDs representing the historical edits of the tweet, with the current tweet's ID.
5. **lang**: Language code of the tweet. (eg. "en" for English, "fr" for French etc.)
6. **created_at**: Timestamp of when the tweet was posted in ISO 8601 format.
7. **referenced_tweets**: Contains data on other tweets referenced within the tweet. This includes *type* (eg. "retweeted") and *id* of the referenced tweet.
8. **author_id**: ID of the user who posted the tweet. The key matching the accounts.tsv.
9. **conversation_id**: ID of the conversation thread this tweet belongs to, usually the same as **id** for the root tweet.
10. **id**: Unique identifier for the tweet.
11. **attachments**: Stores information about any media attached to the tweet, such as images or videos, referenced by *media_keys*.
    - **media_keys**: The key matching the media_lists.txt entries with the full image extension (file name pointer).
12. **public_metrics**: Contains engagement metrics for the tweet:
    - **retweet_count**: Number of retweets.
    - **reply_count**: Number of replies.
    - **like_count**: Number of likes.
    - **quote_count**: Number of times the tweet was quoted.
13. **context_annotations**: Metadata that categorizes the tweet within broader domains and entities (e.g., "Extreme Weather + Climate Change"), including:
    - **domain**: Category of context (e.g., "Events") and an *id*.
    - **entity**: Specific entity name and *id*.

The Account data comprises of:

1. **author_id**: ID of the user who tweeted.
2. **Type**: Lists the role of the author with regards to their involvement with the United Nations COP. (eg. "Private individuals", "Advocacy actors", "Journalistic

actors" etc.)

3. **Lang**: Language code used by the author. (eg. "en" for English, "fr" for French etc.)
4. **Stance**: Account's position on the resolutions passed by COP21. (eg. "For", "Unclear", "Against")

How many instances are there in total (of each type, if appropriate)? => There are 2,260,916 entries of tweets and 1936 accounts of authors with their type, language and stances recorded.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).* => The dataset is technically a subsample of the data available on Twitter during that time period. It contains tweets from the time period 30/11 to 12/12 2015 with the hashtag "cop21". This is not representative of the larger set of all tweets on Twitter during the same period as the collection was focused only on a single hashtag.

The authors were not involved in the data collection process, but merely downloaded the data in November 2024.

What data does each instance consist of? *Raw data (e.g., unprocessed text or images) or features? In either case, please provide a description.* => Both, *raw data* (text) and semi-structured features representing the interactions and meta-data surrounding the tweet.

Is there a label or target associated with each instance? *If so, please provide a description.* => No, there is no target feature assigned (yet).

Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.* => No, we have not omitted any features from the original data.

Are relationships between individual instances made explicit (e.g., users, movie ratings, social network links)? *If so, please describe how these relationships are made explicit.* => Yes, each tweet is referenced by an ID and users are referenced by both their twitter handles and *author_id*. Links are made with the author_id to the users data.

Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.* => No

Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.* => Not that the authors are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? => No, the dataset is not self-contained since there are URLs listed in the tweet entries. *If it links to or relies on external resources*: a) *Are there guarantees that they will exist, and remain constant, over time?* => There are no guarantees that the external tweets will still exist as users could delete or edit the source and the entry could cease to exist or be changed on the internet. b) *Are there official archival versions of the complete dataset? (i.e., including the external resources as they existed at the time the dataset was created)* => NA c) *Are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.* => The dataset download was protected by a password access granted by the course instructor with the link expiring after 5th Novemeber 2024.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.* => The dataset contains Tweets that could be classified as personal identifiable information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.* => Yes, social media and Twitter are known for having offensive, insulting, threatening, and anxiety-causing content since they allow a vast range of unfiltered opinions and reactions from users across the world.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.* => Yes

Does the dataset identify any subpopulations (e.g., by age, gender)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* => If we consider people on social media who post and interact with content under the "cop21" hashtag as a subpopulation, then yes.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.* => Yes, the data represents Tweets made by real people on Twitter. You can identify users by their username, name, author id, and by the content they post.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.* => Yes, images and urls in the dataset show individuals' faces and therefore their ethnic origins.

Any other comments?

Collection Process

How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.* => The original data was collected via the Twitter Academic API but the authors collected it via https://uppsala.box.com

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?* => The original collection mechanism (Academic API) cannot be validated but the download link was verfied via seeing the same data downloaded by two seperate devices.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? => By choosing the "cop21" hashtag and collecting tweets under that search key. The authors of this data card have not subsampled anything more.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? => The original data collected by the PolarVis project consists of people involved in "FORTE, the Swedish Research Council for Health, Working Life and Welfare; Uddannelses - og Forskningsstyrelsen, the Danish Agency for Higher Education and Science; the National Research, Development and Innovation Office Hungary; and FWF, the Austrian Science Fund under CHANSE ERA-NET Co-fund programme". They have received funding from the European Union’s Horizon 2020 Research and Innovation Programme.

Prof. Matteo Magnani who gave the authors access to the data is assumed to be paid by Uppsala Univesity. The authors Anand and Finn who are students are unpaid.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? *If not, please describe the time-frame in which the data associated with the instances was created.* => Tweets were collected in 2023 of the period 30/11 to 12/12 2015. The authors created/downloaded this dataset in November 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.* => No, as of now

Does the dataset relate to people? *If not, you may skip the remainder of the questions in this section.* => Yes

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? => Third parties: Uppsala Box

Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.* => No

Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.* => No but the students as researchers have legitimate interest and are operating under instructions of their professor.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).* => NA

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.* => No

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.* => No

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? *If so, please provide a link or other access point to the "raw" data.* => No

Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point.* => NA

Any other comments?

Uses

Has the dataset been used for any tasks already? *If so, please provide a description.* => Not yet by the authors.

Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point.* => No

What (other) tasks could the dataset be used for? => Studying language, social interactions, climate change persuasion, training LLMs and identifying popular Twitter users within the subset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?* => No

Are there tasks for which the dataset should not be used? *If so, please provide a description.* => Training LLM to identify individuals or anything outside the Mining of Social Data course.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description.* => No

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? *Does the dataset have a digital object identifier (DOI)?* => NA

When will the dataset be distributed? => NA

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.* => NA

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.* => No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.* => No

Any other comments?

Maintenance

Who is supporting/hosting/maintaining the dataset? => There is no support for this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? => Anand Mathew M S: anandmathewms@gmail.com Finn Vaughankraska: vaughankraska@gmail.com

Is there an erratum? *If so, please provide a link or other access point.* => No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?* => Possibly. The users of this dataset should only be the authors so they will notify each other of updates via Github. The authors may add the data to a database in order to more easily and performantly process it.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.* => No, there are no limits on how long the data will be retained for. However, the authors do not intend to keep it longer than the 2024/2025 Uppsala University Fall Term (HT24).

Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.* => No, the older version of this dataset was provided to the authors as hosted on upppsala.box.com by Prof. Matteo Magnani as part of the labs for "Mining of Social Data" course at Uppsala University. However, the instructions for the lab stated that the link will expire after 5th November 2024. More information on obtaining the previous versions need to be addressed to Prof. Matteo Magnani.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.* => No

Any other comments?

Datasheet for: Twitter/Text & Image dataset[^1]

Authors: Finn Vaughankraska, Anand Mathew M S

Related Files:

./data/media_lists.txt ./data/media/* ./data/tweets.dat

Version 1.0 2024-11-05

Motivation

For what purpose was the dataset created? This Dataset was created for the social data mining course. While its main purpose is to serve as a practice dataset for the sake of the authors' learning within the scope of the course, it also serves as a source of information to analyze visual persuasion on Twitter.

Was there a specific task in mind? More explicitly, the listed directives of this dataset is as follows:

- Refresh basic skills for text and data file processing and summarization
- Familiarize the authors with a typical data format.
- Reflect on and/or identify possible data quality and validity issues.
- Generate datasets to be used during the course.
- Practice documenting social data.

Was there a specific gap that needed to be filled? Please provide a description. N/A

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by Anand Mathew and Finn Vaughankraska as students in course *1DL465 11012* (Mining of Social Data) at Uppsala University. More specifically the instruction set was presumed to be created by Matteo Magnani.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. The funding for the original data came from PolarVis. The supporters for the project are listed as:

> Project PolarVis is supported by FORTE, the Swedish Research Council for Health, Working Life and Welfare; Uddannelses - og Forskningsstyrelsen, the Danish Agency for Higher Education and Science; the National Research, Development and Innovation Office Hungary; and FWF, the Austrian Science Fund under CHANSE ERA-NET Co-fund programme, which has received funding from the European Union’s Horizon 2020 Research and Innovation Programme, under Grant Agreement

However both authors of this dataset are fee paying students and an argument could be made that they are funding the creation of this dataset on some level.

Any other comments? Link to the source for the data project

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The JSON data provided represents tweet objects with various fields:

1. **text**: The main content of the tweet, including any retweeted message, hashtags, URLs, and other text content.

2. **entities**: Contains various subfields to identify special elements within the tweet text:

   - **mentions**: Lists users mentioned in the tweet
   - **hashtags**: Hashtags within the tweet
   - **annotations**: Contextual metadata on keywords, including a **probability** score, **type** (e.g., "Place", "Other"), and **normalized_text**.
   - **urls**: Includes URLs within the tweet, with **start** and **end** positions, **url**, **expanded_url** (the full URL), and **display_url**.

3. **possibly_sensitive**: Boolean flag indicating whether the tweet content might be sensitive.

4. **edit_history_tweet_ids**: List of IDs representing historical edits of the tweet, with the current tweet's ID.

5. **lang**: The language code of the tweet (e.g., "en" for English, "es" for Spanish).

6. **created_at**: Timestamp of when the tweet was posted in ISO 8601 format.

7. **referenced_tweets**: Contains data on other tweets referenced within the tweet, including **type** (e.g., "retweeted") and **id** of the referenced tweet.

8. **author_id**: ID of the user who posted the tweet.

9. **conversation_id**: ID of the conversation thread this tweet belongs to, usually the same as **id** for the root tweet.

10. **id**: Unique identifier for the tweet.

11. **attachments**: Stores information about any media attached to the tweet, such as images or videos, referenced by **media_keys**.

    - **media_keys**: The key matching the media_lists.txt entries with the full image extension (file name pointer).

12. **public_metrics**: Contains engagement metrics for the tweet:

    - **retweet_count**: Number of retweets.
    - **reply_count**: Number of replies.
    - **like_count**: Number of likes.
    - **quote_count**: Number of times the tweet was quoted.

13. **context_annotations**: Metadata that categorizes the tweet within broader domains and entities (e.g., "Extreme Weather + Climate Change"), including:

  - **domain**: Category of context (e.g., "Events") and an **id**.
  - **entity**: Specific entity name and **id**.

How many instances are there in total (of each type, if appropriate)? 2,260,916 entries of tweets as outlined above with 44,482 images in the media_list.txt file.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

- If the dataset is a sample, then what is the larger set? It is technically a subsample of the data that is available on Twitter during that period.
- Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). The dataset contains tweets and images from the time period 30/11 to 12/12 2015 with the hashtag "cop21". The subsample is not representative of the population ("population" being all the tweets on twitter) since collection was focused around a single hashtag which is not a random sample of the possible Tweets on the platform. The authors did not collect the original data, but downloaded it in 2024.

What data does each instance consist of? Raw data (e.g., unprocessed text or images) or features? In either case, please provide a description. Both *raw data* (text and images) and semi-structured features representing the interactions and meta-data surrounding the tweet.

Is there a label or target associated with each instance? If so, please provide a description. No, there is not a target feature defined (yet).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. No, we have not omitted any features from the original data.

Are relationships between individual instances made explicit (e.g., users, movie ratings, social network links)? If so, please describe how these relationships are made explicit. Yes, users are referenced by both their Twitter handles (usernames) and by **author_id**. Additionally links are made between tweets and media via the attachments object.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Not that the authors are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? No, the dataset is not self-contained since there are urls listed in the tweets entries.

- If it links to or relies on external resources a) are there guarantees that they will exist, and remain constant, over time; No the users who made tweets could delete or edit the source and the entry could cease to exist on the internet or be changed. b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); NA c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. The dataset download was protected by a password access granted by the course instructor and expiring on the 5th of November 2024.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description. The data contains images and Tweets that could be classified as personal identifiable information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. Yes, social media and Twitter are known for having offensive, insulting, threatening, and anxiety-causing content since they allow a vast range of unfiltered opinions and reactions from users across the world.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. If you consider people on social media who post and interact with content under the "cop21" hashtag, then yes the dataset identifies subpopulations who interact with the hashtag.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. Yes, the data are Tweets made by real people on Twitter. You can identify users by their username, name, author id, images, and by the content they post.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. Yes, images and urls in the dataset show individuals' faces and therefore their ethnic origins (and if neural networks can identify peoples' sexual orientations by only their face, then the dataset also reveals that).

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The original data was collected via the Twitter Academic API but the authors collected it via https://uppsala.box.com

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? The original collection mechanism (Academic API) cannot be validated but the download link was verfied via seeing the same data downloaded by two seperate devices.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? By choosing the "cop21" hashtag and collecting tweets under that search key. The authors of this data card have not subsampled anything more.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Students Anand Mathew and Finn Vaughankraska who are both unpaid. PolarVis (who studies visual persuasion) is "FORTE, the Swedish Research Council for Health, Working Life and Welfare; Uddannelses - og Forskningsstyrelsen, the Danish Agency for Higher Education and Science; the National Research, Development and Innovation Office Hungary; and FWF, the Austrian Science Fund under CHANSE ERA-NET Co-fund programme, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme". Matteo Magnani (who granted the students access) is assumed to be paid by Uppsala Univesity.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created. Tweets were collected in 2023 of the period 30/11 to 12/12 2015. The authors created/downloaded this dataset in November 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section. Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Third parties: Uppsala Box

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. No but the students as researchers have legitimate interest and are operating under instructions of their professor.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). NA.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. No.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. No.

Was the *raw data* saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the *raw data*. No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. NA

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description. Not by the authors.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. No.

What (other) tasks could the dataset be used for? Studying language, social interactions, climate change persuasion, training LLMs and identifying popular Twitter users within the subset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms? No.

Are there tasks for which the dataset should not be used? If so, please provide a description. Training LLMs and identifying individuals or anything outside the Mining of Social Data course.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? NA.

When will the dataset be distributed? NA.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. NA.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

Any other comments?

Maintenance

Who is supporting/hosting/maintaining the dataset? There is no support for this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Finn Vaughankraska: vaughankraska@gmail.com Anand Mathew: anandmathewms@gmail.com

Is there an erratum? If so, please provide a link or other access point. No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)? Possibly. The users of this dataset should only be the authors so they will notify each other of updates via Github. The authors may add the data to a database in order to more easily and performantly process it.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. No, there are no limits on how long the data will be retained for. However, the authors do not intend to keep it longer than the 2024/2025 Uppsala University Fall Term.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users. No. The only backup version will be maintained and hosted as it was originally downloaded on upppsala.box.com. The authors have no control of the availability or contents of that version.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description. No.

Any other comments?

[^1]: From: Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. Commun. ACM, 64(12), 86�92. https://doi.org/10.1145/3458723