effrosynidis_2022_the_climate_change_twitter_dataset

Year

2022

Author(s)

Dimitrios Effrosynidis and Alexandros I. Karasakalidis and Georgios Sylaios and Avi Arampatzis

Title

The climate change Twitter dataset

Venue

Expert Systems with Applications

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

Nr of inspected topic-words: 15

Label generation

For each topic, there are present: the title of the topic, which was manually given by us, the word cloud of the 1000 most common words in the topic, and the top 15 most unique words that determine the topic. These 15 words are the ones we used to determine the topic's title.

They were extracted using the value 0.2 of the relevance method from Sievert and Shirley (2014).

Fig. 5. Ten topics discovered by the LDA algorithm. For each topic, there are available: its title, the word cloud with the most common 1000 words and the 15 most unique words that were used to determine the topic's title.

Motivation

"One sign of a healthy outcome of a topic modeling procedure is when the top words that define a topic can easily guide a human to provide this topic's title. With more than 10 topics, it was not clear how to name the topics, while with 10 topics we could successfully name 9 topics."

Topic labeling used as a proxy to evaluate the quality of the topic model

---

Topic modeling

LDA

Topic modeling parameters

Nr of topics (k): {10, 15, 20}

Nr. of topics

10

Label

Sentence or single word (one label) manually assigned by the authors

Label selection

/

Label quality evaluation

/

Assessors

/

Domain

Paper: Climate change
Dataset: Social media (Twitter)

Problem statement

This work creates and makes publicly available the most comprehensive dataset to date regarding climate change and human opinions via Twitter.
It has the heftiest temporal coverage, spanning over 13 years, includes over 15 million tweets spatially distributed across the world, and provides the geolocation of most tweets. Seven dimensions of information are tied to each tweet, namely geolocation, user gender, climate change stance and sentiment, aggressiveness, deviations from historic temperature, and topic modeling, while accompanied by environmental disaster events information.

Corpus

Origin: Twitter
Nr. of documents: 15,789,411
Details:

- The Twitter Climate Change Dataset of this work was constructed by merging three publicly available datasets, namely, Credibility of Climate Change Denial in Social Media Data, Climate Change Tweets IDs Data, and Twitter Archive Data.
- The Credibility of Climate Change Denial in Social Media dataset contains 14,353,859 unique tweet IDs that were collected between June 6, 2006, and April 12, 2018, based on a search filter that each tweet contains at least one of the following queries: climate change, climatechange, global warming, and globalwarming.
- The Climate Change Tweets IDs dataset was retrieved from the Harvard Dataverse Repository. This collection consists of 39,622,026 tweet IDs related to climate change that were collected between September 21, 2017, and May 17, 2019. Tweets were retrieved using the following queries: climatechange, climatechangeisreal, actonclimate, globalwarming, climatechangehoax, climatedeniers, climatechangeisfalse, globalwarminghoax, climatechangenotreal, climate change, global warming, and climate hoax.
- Extra Tweets are collected through the Internet Archive from January 1, 2019, to October 1, 2019. We retrieved filtered tweets that contained the queries climatechange and globalwarming.

Document

A tweet together with eight additional dimensions: geolocation, gender, stance, sentiment, aggressiveness, temperature, topics and environmental disaster events.

Pre-processing

- Keyword filtering was applied using the previously mentioned keywords to ensure that the tweets contain at least one of the desired keywords.
- Removal of duplicated tweets.
- Removed Unicode characters, URLs, hashtags in front of words, 'www' in front of words, user mentions, email addresses, newline characters, quotes, and single-character words.
- Lower-cased the tweets
- Removed standard English stop-words plus some other words we defined (e.g. say, get, know, may, one, mr, also)
- Words that were rarely present in the dataset (with total collection count up to 10) were also removed.
- Remaining words were lemmatized

```
@article{effrosynidis_2022_the_climate_change_twitter_dataset,
  abstract = {This work creates and makes publicly available the most comprehensive dataset to date regarding climate change and human opinions via Twitter. It has the heftiest temporal coverage, spanning over 13 years, includes over 15 million tweets spatially distributed across the world, and provides the geolocation of most tweets. Seven dimensions of information are tied to each tweet, namely geolocation, user gender, climate change stance and sentiment, aggressiveness, deviations from historic temperature, and topic modeling, while accompanied by environmental disaster events information. These dimensions were produced by testing and evaluating a plethora of state-of-the-art machine learning algorithms and methods, both supervised and unsupervised, including BERT, RNN, LSTM, CNN, SVM, Naive Bayes, VADER, Textblob, Flair, and LDA.},
  author = {Dimitrios Effrosynidis and Alexandros I. Karasakalidis and Georgios Sylaios and Avi Arampatzis},
  date-added = {2023-03-22 18:08:56 +0100},
  date-modified = {2023-03-22 18:08:56 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2022.117541},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Climate change, Machine learning, Sentiment analysis, Topic
```

```
modeling, Twitter},
  pages = {117541},
  title = {The climate change Twitter dataset},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417422008624},
  volume = {204},
  year = {2022}}
```

#Thesis/Papers/Initial