Expert Systems With Applications 204 (2022) 117541

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

The climate change Twitter dataset

Dimitrios Effrosynidis [a,*], Alexandros I. Karasakalidis [a], Georgios Sylaios [b], Avi Arampatzis [a]

[a] Database & Information Retrieval research unit, Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece
[b] Lab of Ecological Engineering & Technology, Department of Environmental Engineering, Democritus University of Thrace, Xanthi 67100, Greece

ARTICLE INFO

Keywords:
Climate change
Machine learning
Sentiment analysis
Topic modeling
Twitter

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.

1. Introduction

It is of high importance to deduce human opinion on the topics of climate change and global warming. Since data is knowledge, a quality dataset can be the source to answer critical questions about these emerging topics in order to successfully tackle them. The large-scale geophysical phenomenon of climate change torments society for several years. It is driven by the gradual increase in atmospheric greenhouse gases causing raised air temperature, sea-level rise, ocean acidification, and more frequent extreme weather events (Masson-Delmotte et al., 2018; Shukla et al., 2019). Although it is often attributed to human activity, societal opinions seem divided (Brulle, Carmichael, & Jenkins, 2012; Hahnel, Mumenthaler, & Brosch, 2019; Sugg, 2021). Climate change and its contemporary manifestation as global warming are essential to be recognized and addressed by the combined efforts of the public and the authorities. In order to effectively tackle the problem, policy and decision-makers need to understand better public perceptions on climate change, and probably explain better to the society the Driver-Pressure-State-Impact-Response (DPSIR) cycle (Jorgenson et al., 2019). As climate change is a global-scale problem, traditional approaches that measure public opinion, like surveys, are difficult to be employed (Cody, Reagan, Mitchell, Dodds, & Danforth, 2015; Philo & Happer, 2013).

The rise of social media in the recent decade and their use as opinion sharing platforms makes them an invaluable source to study human stance on climate change (El Barachi, AlKhatib, Mathew, & Oroumchian, 2021). One of the most popular social media platforms is Twitter, where anyone can share their opinions and thoughts, or 'tweet', using up to 280 characters. Researchers and social scientists use Twitter to extract and analyze opinions and explore from text hidden patterns and trends on public sentiment and perceptions (Zimbra, Abbasi, Zeng, & Chen, 2018). Moreover, it is now easier than ever to find, combine and use data from multiple external sources. We are in the dawn of the data science and machine learning era and there are numerous tools, algorithms, and methods to handle big data and extract valuable information (Al-Jarrah, Yoo, Muhaidat, Karagiannidis, & Taha, 2015).

In this paper, we create and make publicly available the most comprehensive dataset to date regarding climate change and Twitter, namely The Climate Change Twitter Dataset. The dataset is consisted of more than 15 million tweets about climate change and global warming from 2006 to 2019 in the English language. For each tweet, its exact geolocation is given, the temperature deviation at the time and place it was written, along with the gender of the twitter, the stance, sentiment, aggressiveness and topic of the tweet. The dataset is also accompanied by environmental disaster events that took place during this 13-year period.

The rest of this paper is organized as follows. In the next section, we briefly summarize related works. Section 3 describes the data merging, pre-processing, and enrichment process, together with the algorithms and methods used. Section 4 presents and discusses the experimental results, while conclusions are summarized in Section 5.

* Corresponding author.
E-mail addresses: deffrosy@ee.duth.gr (D. Effrosynidis), alexkara23@ee.duth.gr (A.I. Karasakalidis), gsylaios@env.duth.gr (G. Sylaios), avi@ee.duth.gr (A. Arampatzis).

https://doi.org/10.1016/j.eswa.2022.117541
Received 20 October 2021; Received in revised form 15 January 2022; Accepted 6 May 2022
Available online 14 May 2022

2. Related work

In this section, we review relevant research publications, we identify the main gaps in previous works and we highlight our contributions to fill such shortcomings.

The first, pioneering research attempting to deduce human perceptions on climate change based on social media, like Twitter, was conducted by Kirilenko and Stepchenkova (2014). The authors collected tweets containing the queries 'climate change' and 'global change' during 2012–2013. Global tweets in five different languages (English, German, Russian, Portuguese and Spanish) were analyzed, their geolocation was extracted and their patterns were explored, in relation to major climate change related events. Classical sentiment classification algorithms, like the Naive Bayes and the Support Vector Machines (SVMs) were applied to detect and track opinions regarding climate change from Twitter feeds (An et al., 2014). Feature selection was implemented to filter the tweets' dataset and establish the best search strategy. It was found that major climate events could result in sudden changes in sentiment polarity, verifying the importance of social media mining on the large-scale exploration of public perceptions. The authors of Williams, McMurray, Kurz, and Lambert (2015) created a dataset of 590,608 tweets from January 13 to May 30, 2013, retrieved with queries 'global warming', 'climate change', 'agw' (an acronym for 'anthropogenic global warming'), 'climate' and 'climate realists' and manually classified the 1545 most active users to either activist, skeptic, neutral, or unknown.

Gender differences in the climate change communication on Twitter were analyzed by Holmberg and Hellsten (2015). After identifying the gender of each username, the authors found that male and female tweeters use very similar language in their tweets, but clear differences were observed in the use of hashtags and usernames. Results showed that female tweeters illustrate a more convinced attitude towards the anthropogenic impact on climate change, while male tweeters present a skeptical stance. In a sentiment analysis study, Cody et al. (2015) designed a hedonometer, using a lexicon-based approach, capable of assessing the happiness score of tweets. The examined database consisted of 1.5 million tweets containing the keyword 'climate' between 2008 and 2014. The hedonometer produced a happiness time-series, and its performance was successfully tested after extreme climate change events. In a similar manner, a multi-scale analysis of Twitter activity for the extreme event of Hurricane Sandy was done by Kryvasheyeu et al. (2016). Twitter data were geolocated and a relationship was revealed between the proximity to hurricane Sandy's track and social media activity. Twitter activity was highly correlated to the economic damage caused by Sandy, in per capita terms, while a lexicon-based approach was implemented to identify tweet sentiment during the event.

An overview of how Twitter is being used by scholars is reviewed by Giachanou and Crestani (2016). The authors discuss the existing sentiment analysis methods, identifying that most researchers follow four approaches: machine learning, lexicon-based, hybrid and graph-based. Methodologies on opinion retrieval, tracking sentiments over time, irony detection, emotion detection, tweet sentiment quantification, as well as the available datasets to implement them, are also reviewed.

A dataset of 3.7 million tweets geolocated in the USA, related to climate change and global warming, was collected between 2012 and 2014 (Yeo et al., 2017). Based on their context, tweets were split into two groups; the 'climate change' group and the 'global warming' group, and statistical tests were applied to explore the differences among these groups within specific topics of discussion. As shown, more posts on 'global warming' were found in discussions related to weather and energy, while more 'climate change' tweets were related to the environment and politics. Such results underline the importance of scientifically-based communication on climate change issues. In parallel, air temperature anomalies across the US were found related to the volume of climate change tweets.

Sarcasm and incivility were examined by Anderson and Huntington (2017) on 4,094 tweets during an extreme weather event, where it was found that their instances were low, overall, but with a higher density from climate change skeptics. A comprehensive review on the different areas of research about discussions on climate change in Twitter was done by Fownes, Yu, and Margolin (2018). The authors assessed public views on the topic, specified spatio-temporal trends in activity and content, and discussed how the medium could be actively used to promote dialog. Some limitations are mentioned, such as that Twitter discussions are not yet well mapped onto attitudes about climate change, as captured by surveys.

The nature and contradictions in climate change discussions between different countries and over time was analyzed, performing tweet volume analysis on a collection of 366,244 geotagged tweets, dated from July 1, 2016, to February 28, 2018 (Dahal, Kumar, & Li, 2019). The data was collected using the queries 'climate change', 'carbon dioxide', 'fossil fuel', 'carbon footprint', and 'emissions'. The research of Dahal et al. (2019) is focused on specific extreme events, carrying out lexicon-based sentiment analysis and topic modeling with probabilistic word distribution, using the LDA approach, but without evaluating the process. Topic modeling was found to be useful for event detection, as spikes in topic discussion or total sentiment were in direct correspondence with significant real-world events. The overall sentiment was negative and the USA was found at risk of falling behind other countries in climate change mitigation.

Binary classification (believers vs. deniers) was used by Koenecke and Feliu-Fabà (2019) on 16,360 tweets, after labeling the data based on influential users (manually labeling some tweets of a user and assuming that all other tweets posted by the same user have the same label). As positive was labeled any tweet composed by a Twitter account marked as 'believer', while any tweet composed by a 'denier' was labeled as negative. The authors created an event-related dataset containing five natural disasters, and labeled manually 500 tweets. Results showed that the 2018 hurricanes yielded a statistically significant increase in the average tweet sentiment, affirming climate change. On the contrary, the 2018 blizzard and wildfires did not. Deep Neural Network algorithms for binary classification between believers and non-believers were capable of predicting 205,795 climate change tweets in 2016 (Chen, Zou, & Zhao, 2019), achieving 88.1% in two-class classification model accuracy. The model was trained by 2,000 manually-labeled tweets.

Machine learning models were used to identify tweets relevant to hurricane Irma (Sit, Koylu, & Demir, 2019). Evaluating Logistic Regression, Ridge, SVMs, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) algorithms, the last outperformed all the others. Unsupervised topic modeling was also employed with LDA to mainly detect the topic of affected individuals and perform spatio-temporal analysis on it. A lexicon-based sentiment analysis approach, including novel fuzzy-based Dempster–Shafer fusion, was applied to analyze the polarity in the expressed opinions of people of Alaska, in relation to the state energy mix and the introduction of renewable energy. An information-rich geotagged Twitter dataset was used for this study (Abdar et al., 2020). Similarly, a large set of geolocated tweets were analyzed with the lexicon-based sentiment method to reach the daily variation in expressed sentiment across the US and six other countries (Baylis, 2020). This dataset was combined with meteorological observations to estimate the sentiment response to temperature and profanity. Correlation occurred between the expressed sentiment and both hot and cold temperatures. Gender, emotion and sentiment lexicon-analysis was performed on 811,211 UK-tweets and 961,969 Spanish-tweets collected from the first half of 2019 (Loureiro & Alló, 2020). Findings showed that messages in the U.K. are less negative and the most evoked feeling is fear.

Based on the above review, it is evident that Twitter can be effectively used for sentiment and emotion analysis of public perceptions, especially on global environmental issues, like climate change. The

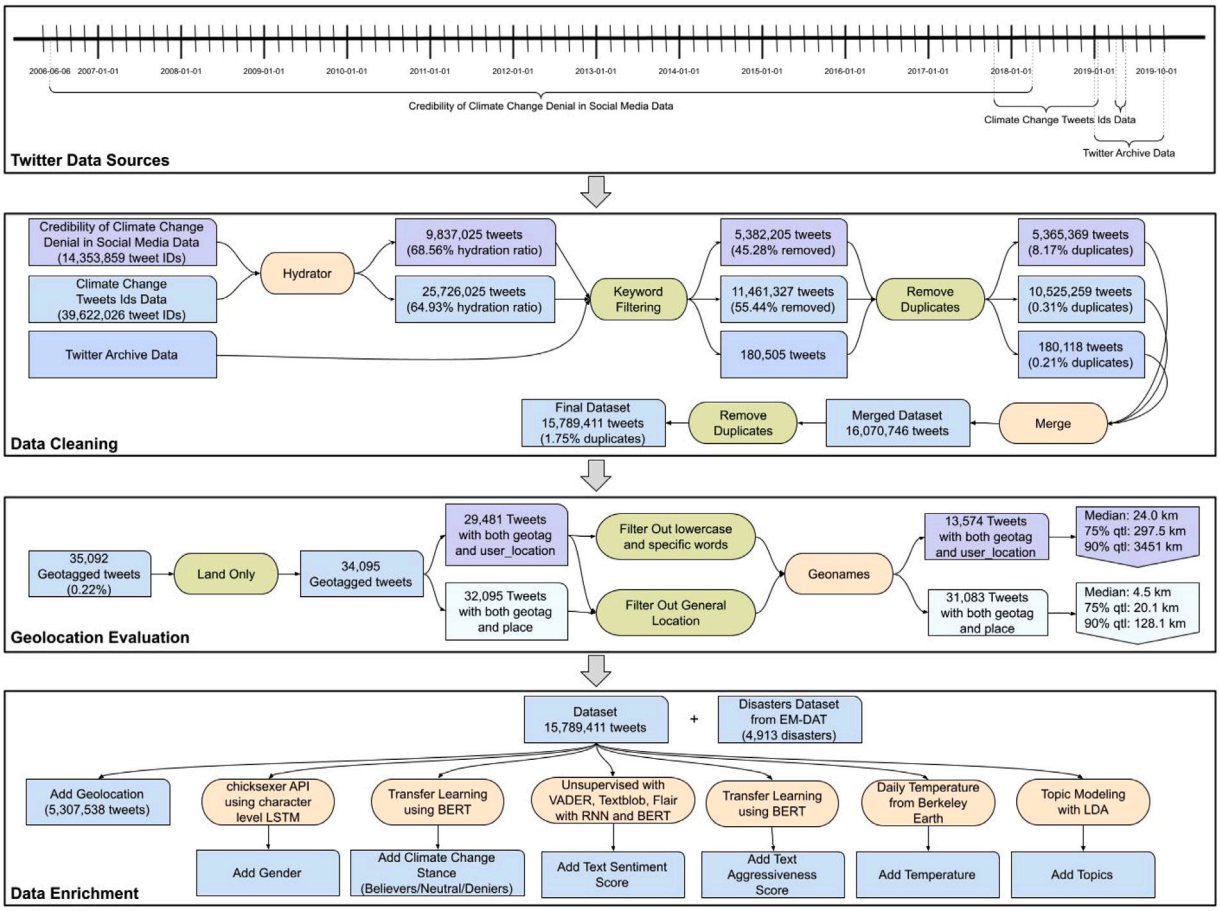

Fig. 1. Data collection, pre-processing, and enrichment.

long-term sentiment analysis of tweets exhibits direct response to the occurrence of extreme events, while spatio-temporal patterns indicate that stronger correlation is shown in CC-prone areas. Although big datasets have been widely used in previous studies, these works focus only on two or at most three aspects of human attitude towards climate change. Moreover, the temporal coverage of the examined tweets is rather limited, varying from a few months to 1–2 years.

In the present work, a comprehensive dataset of eight aspects, namely geolocation, stance, sentiment, aggressiveness, temperature, gender, topics, and disasters, has been developed. The temporal coverage is the heftiest, spanning over 13 years, and the produced geolocations were generated with the highest possible precision.

To our knowledge, this is the most comprehensive climate change Twitter dataset to date. In parallel, some recently published works employ modern algorithms and models, but evaluation in most studies is lacking. Herein, the most competitive, state-of-the-art algorithms are employed, evaluated, and compared, in order to deliver the highest possible accuracy in all studied aspects. More specifically, this is the first work that utilizes the BERT algorithm to classify climate change stance, aggressiveness, and sentiment achieving a 10% increase in performance from the second best algorithm (LSTM). Finally, it is the first work that uses machine learning with evaluation instead of just lexicon-based approaches to predict the sentiment.

3. Materials and methods

In this section, we will describe the datasets and the whole process followed to collect, pre-process, and enrich them, which is graphically shown in Fig. 1. We will also talk about all the algorithms, machine learning models and methods that were employed in this study.

3.1. Dataset merging and pre-processing

The initial step is to collect the data from various sources, merge them, and apply appropriate pre-processing techniques.

3.1.1. Twitter data

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 (Samantray & Pin, 2019) 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 (Littman & Wrubel, 2019) 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, from the Twitter API using Social Feed Manager. There exists a gap in the data between January 7, 2019, and April 17, 2019. Tweets were retrieved using the POST statuses/filter method of the Twitter Streaming API, using the track parameter with the following queries: #climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal, climate change, global warming, and climate hoax.

In order to fill the aforementioned gap in the Climate Change Tweets IDs dataset and further extend the data, we collected extra

tweets through the Internet Archive[1] from January 1, 2019, to October 1, 2019 (the latest available date at the time this task was completed). It is a free online library providing about 1% random Twitter data of the full Twitter database. We retrieved filtered tweets that contained the queries #climatechange and #globalwarming.

Based on the three sources of information that were combined, and the English keywords used to create them, the resulting dataset will be in the English language. We tested that using the spaCy Python library (Honnibal & Montani, 2017), which indicated that the dataset is 98% in English. A bias on the origin locations of the tweets is expected to countries that use the English language as their mother tongue. Nevertheless, if we had to choose only one language, since the English language is the most common second language in the world, it is the most appropriate to construct a global dataset.

3.1.2. Data pre-processing

Since it is against Twitter rules to publish the original tweets, the first two sources of data contain only the IDs of the tweets. In order to obtain the complete tweet text and metadata, a process called 'hydrating' was employed. It basically searches each ID and retrieves its data. The open-access toolkit Hydrator[2] was used for this purpose. A similar process was also followed by Chen et al. (2020) when developing a multilingual COVID-19 Twitter dataset. Hydrating millions of tweets is time-consuming as Twitter has limits on its usage. About two weeks were spent hydrating the data. Since hydrating retrieves tweet ID data at the time of its usage, many tweet IDs and specifically older ones may not be found, as they were deleted by the author or from Twitter itself, or the user deleted his account, or he got banned from Twitter. At the end of the process, we managed to hydrate 9,837,025 tweets (68.56% hydration ratio) from the Credibility of Climate Change Denial in Social Media dataset and 25,726,025 tweets (64.93% hydration ratio) from Climate Change Tweets IDs dataset.

After retrieving the complete tweet data, keyword filtering was applied using the previously mentioned keywords "#climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal" to ensure that the hydrated tweets contain at least one of the desired keywords. We decided to include hashtag-only keywords and not phrases like 'climate change', as hashtags in Twitter are used like a topic specifier. Filtering was followed by a merge of the three datasets and the removal of duplicated IDs. Thus, the final dataset that is used in this work consists of 15,789,411 unique tweet IDs between June 6, 2006, and October 1, 2019. The produced dataset is publicly available.[3]

3.2. Data enrichment

The next step includes the enrichment of the initial dataset with eight dimensions, namely, geolocation, gender, stance, sentiment, aggressiveness, temperature, topics and environmental disaster events.

3.2.1. Geolocation data

Having geolocated data is crucial and can contribute to spatial findings. Tweets could include geolocation information only if the user has enabled this option in his Twitter account. Unfortunately, this option is enabled for a tiny percentage of tweets. Similarly, Graham, Hale, and Gaffney (2014) found that out of the 19.6 billion tweets collected in 2013, only 0.7% contained geolocation information. In the present dataset, this number equals to 0.22% of the total tweets. The discrepancy from the bibliographic source may be attributed to the fact that the current dataset is much newer, Twitter has changed its policy regarding geolocation, and people in recent years are more cautious of publicly sharing their information.

This tiny amount of geolocated tweets is insufficient for a reliable spatial analysis, thus, other tweet metadata were considered to extract spatial information. Previous works in the subject, populated their databases with additional geolocation points using metadata on the 'place', i.e., the location of the tweet in text form. 1.77% of our dataset includes a place. While it is an improvement, it is still insufficient for analysis. On top of that, the sample would be biased because only advanced user groups are aware of such Twitter options and are able to enable them.

Another location information that can be found in Twitter metadata is the user's location. This location is added by the user himself to enrich his profile. It is very easy and a significant amount of users add their location. In our dataset, 75.45% of tweets are accompanied by a user location. It is expected that user location would be sometimes misleading as we are most interested in where the user was when he tweeted and not where he declared that he lives. The user might also have moved to a new location and forgot to change his Twitter information. Nevertheless, using the user's location, our dataset will have more than enough geolocation observations to conduct a reliable and not biased geospatial analysis. We will also prove through empirical evaluation that the uncertainty the user locations may introduce is within small margins.

The GeoNames[4] geographical database is used, which covers all countries and contains over eleven million place names to reverse-geocode either the places or user locations. We are evaluating places and locations separately. We wrote a Python script to use the GeoNames database and apply filters on top of it for better performance. All of the filters were chosen after evaluating them. These filters exclude general locations (e.g. California), skip places with all-lowercase letters, and exclude specific words (e.g. Earth, Heaven). First, for the evaluation of the place metadata, we reverse-geocode the places whose coordinates are already known in the dataset. For these 32,591 computed coordinates, their haversine distance from the real, ground truth coordinates is computed. Its median is only 4.5 km, meaning that this process generated very precise geolocations. For comparison, previous works have found this number to be 13 km (Kirilenko & Stepchenkova, 2014), 11 km (Kirilenko, Molodtsova, & Stepchenkova, 2015), or 27.8 km (Sisco, Bosetti, & Weber, 2017). Next, the same evaluation is done for the user location metadata. The median of the haversine distance is larger as expected, equal to 24 km, but is acceptable. Previous works have not used user location. Finally, we applied this method to the whole dataset and successfully extracted 5,307,538 geolocations from the place and the user location.

3.2.2. Gender

The second extra information that enriches the dataset is the gender of the tweeter. We used a Python package[5] which is a trained gender classifier implemented using character-level multilayer LSTM. As written in its documentation, its architecture is as follows: (1) character embedding layer, (2) first LSTM layer, (3) second LSTM layer, (4) pooling layer and (5) fully connected layer. The classifier was trained on data from the Social Security Administration of USA (2017 version)[6], which includes 32,634 unique names and a collection of datasets for NLP research by Milos Bejda,[7] which includes about 130,000 names. Passing as input the user name of the tweeter, it outputs the probability of the user being male or female (the sum of the two probabilities is one). We kept the prediction only if one of the probabilities was higher than 95%, else we marked it as undefined. The choice of this

[1] https://archive.org/details/twitterstream.
[2] https://github.com/DocNow/hydrator.
[3] https://data.mendeley.com/datasets/mw8yd7z9wc/
[4] https://www.geonames.org.
[5] https://github.com/kensk8er/chicksexer.
[6] https://www.ssa.gov/oact/babynames/limits.html.
[7] https://mbejda.github.io/.

Table 1
Stance training dataset label distribution.

	Deniers	Neutral	Believers
#	3,990	7,715	22,962
%	11.51	22.25	66.24

threshold was applied after manually reviewing 1000 random samples using thresholds of 60%, 75%, 90%, 95%, and 99%, combined by an acceptable number of total tweets that had a gender after each threshold (few undefined). After using the pre-trained model and our 95% rule, we evaluated on the 31,271 names from the Social Security Administration of USA dataset (2020 version) and we got an accuracy of 90%. In total, there were identified 10,307,402 tweets written by a male, 4,895,134 by a female, and 586,875 were undefined, indicating that males use the platform more.

3.2.3. Stance

The third enrichment regards the stance of the tweet towards climate change. That is if the tweet supports the belief of man-made climate change (believer), if the tweet does not believe in man-made climate change (denier), and if the tweet neither supports nor refuses the belief of man-made climate change (neutral). We used Transfer Learning, training a classifier on a third dataset that is in the same domain (climate change tweets), and using it to predict the stance of each tweet in our dataset. The dataset that we used to train our classifier is publicly available[8] and contains 34,667 labeled tweets pertaining to climate change collected between April 27, 2015, and February 21, 2018. The tweets were labeled by 3 independent reviewers. The distribution of 'deniers', 'neutral' and 'believers' on the training data can be seen on Table 1. This dataset will be referred to as '*Stance training dataset*'. We trained and evaluated several models in this dataset (see Section 4.1) and used the best model (BERT) to predict our whole dataset. In total, 11,292,424 tweets were classified as believers, 1,191,386 as deniers, and 3,305,601 as neutral.

3.2.4. Sentiment

The fourth enrichment concerns the sentiment of the tweet. We employed unsupervised machine learning techniques and more specifically two lexicon-based approaches using VADER (Hutto & Gilbert, 2014) and Textblob (Loria, 2018), as well as a pre-trained RNN model and a pre-trained BERT model using the Flair framework (Akbik et al., 2019). All of these approaches return a sentiment score on a continuous scale. This scale ranges from −1 to 1 with values closer to 1 being translated to positive sentiment, values closer to −1 representing a negative sentiment while values close to 0 depicting no sentiment or being neutral. The outcome was five new features, one for each approach and one from the unweighted average of all approaches, measuring the sentiment in the aforementioned scale.

3.2.5. Aggressiveness

The fifth enrichment adds text aggressiveness information to the dataset. Our methodology followed the steps we took in the process of the third enrichment, where we used Transfer Learning. In order to train our classifiers, we combined data from two sources. The first source was a Twitter Sentiment Analysis practice competition[9] on hate speech from Analytics Vidhya containing labeled tweets (0/1 for non-hateful/hateful tweets) in .csv form. The second source was SemEval's 2021 task 7 on detecting and rating humor and offense on short texts, containing small texts labeled as humorous or not (1 or 0) and as offensive or not (1 or 0). The need to merge these datasets emerged from the enormous imbalance of the Analytics Vidhya data (containing 2,242 hatefull tweets out of 31,962). Thus, we created a dataset from these sources, consisted of 2,242 hatefull tweets and 4,500 non hatefull tweets from the Analytics Vidhya data and 2,258 offensive short texts from SemEval's data (9,000 labeled tweets in total), that was balanced in regards to the offensive and non-offensive tweets/texts. As in the second enrichment, after training and evaluating our models (see Section 4.1, we used the best model (BERT) to predict our dataset. In total, 4,527,267 tweets were classified as offensive out of the 15,789,411.

3.2.6. Temperature

The sixth dimension adds temperature to the dataset. We used the Berkeley Earth (Rohde & Hausfather, 2020) daily land-surface temperature datasets. They provide highly detailed data in both space and time. For each 1°×1° lat–long grid cell on earth, the daily average, minimum, and maximum temperature anomaly is reported. This anomaly or deviation is in Celsius and relative to the January 1951–December 1980 average. We wrote a Python script that matches each lat–long pair of the Twitter dataset to the closest available Berkeley Earth point on that particular day the tweet was written.

3.2.7. Topic modeling

The seventh level of information adds topics to the dataset. There are plenty of available topic modeling techniques. A survey by Qiang, Qian, Li, Yuan, and Wu (2020) compares nine of them in Twitter texts (among others), namely Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), four Dirichlet Multinomial Mixture (DMM) (Nigam, McCallum, Thrun, & Mitchell, 2000) approaches named GSDMM (Yin & Wang, 2014), LF-DMM (Nguyen, Billingsley, Du, & Johnson, 2015), GPU-DMM (Li, Wang, Zhang, Sun, & Ma, 2016) and GPU-PDMM (Li et al., 2017), two Global Word Co-Occurrences approaches named BTM (Cheng, Yan, Lan, & Guo, 2014) and WNTM (Zuo, Zhao, & Xu, 2016), and two Self-Aggregation approaches named SATM (Quan, Kit, Ge, & Pan, 2015) and PTM (Zuo, Wu, et al., 2016). The three methods with the highest performance in terms of topic coherence were found to be GPU-PDMM, PTM, and LDA, while LDA had the fastest runtime (×233 faster than GPU-PDMM, and ×20 than PTM). Recently, a flexible framework to combine latent topic information with BERT embeddings on tweet data, called T-BERT was proposed by Palani, Rajagopal, and Pancholi (2021). They compared it with LDA and found both methods to vary on performance in terms of coherence score depending on the number of topics chosen. BERT was superior when topics were less than 10, and LDA when topics where 10 to 17.

Based on the above, we used LDA for our experiments. The goal of LDA is to assign tweets to topics, where a topic consists of probabilistic mixtures of words that represent word co-occurrence trends. Multiple topics are assigned to each tweet with a probability. The topic with the highest probability will be the label of the tweet. Appropriate text preprocessing is essential for successful topic modeling. We used several text pre-processing techniques for Twitter (Effrosynidis, Symeonidis, & Arampatzis, 2017). Specifically, and in this order, we 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. We then lower-cased the tweets and removed standard English stop-words plus some other words we defined (e.g. say, get, know, may, one, mr, also), that were not included in the Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009) we used. Words that were rarely present in the dataset (with total collection count up to 10) were also removed. The remaining words were lemmatized, and the uni-gram and bi-gram terms were fed to the LDA algorithm.

The algorithm was implemented using the gensim library (Rehurek & Sojka, 2010) and was applied on the whole dataset of 15,789,411 tweets. The big volume of data is not expected to affect negatively the model, as Crossley, Dascalu, and McNamara (2017) showed that larger corpora lead to greater accuracy for LDA. The most important

[8] https://www.kaggle.com/edqian/twitter-climate-change-sentiment-dataset.

[9] https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/.

parameter is the number of topics k that the model will discover. It is not known a priory which k is best as too few topics could include well-separated topics inside them that could be further separated, and too many topics would introduce noise and make no particular sense. In that respect, we run LDA with 10, 15, and 20 topics. We did not use a larger number of topics since the results of 20 topics were already inferior and more topics would not be manageable.

We evaluated the produced topics using five different methods. The first method is perplexity and evaluates how good a produced LDA model is. It compares the theoretical word distributions represented by the topics with the actual distributions in the topics discovered. This statistic is relative and has meaning only when different models are compared. Lower values indicate a better model. On top of perplexity, we used three per-topic measures that have been previously used in Sit et al. (2019): (1) Probability of a topic, which represents how common a topic is across all tweets and is calculated by dividing the number of word tokens assigned to the topic by the sum of the token counts for all topics. (2) Entropy of a topic, which represents the distribution characteristics of a topic across tweets. For example, if entropy is high, the topic is distributed evenly over tweets, whereas if entropy is low, the topic is concentrated on a smaller number of tweets. (3) Kullback–Leibler divergence, which represents how far a topic is from the overall distribution of words in the dataset. A greater corpus distance means that the topic is more distinct as compared to the overall distribution of words, whereas a smaller distance means that the topic is more similar to the corpus distribution. The final evaluation criterion was done by visualizing the topics in a two-dimensional space using LDAvis (Sievert & Shirley, 2014). A model where its topics do not overlap is considered superior.

3.2.8. Environmental disaster events

The eighth and last enrichment concerns environmental disaster events. The International Disaster Database (EM-DAT)[10] was accessed to download disaster events. EM-DAT records over 24,000 historic disasters and organizes them by specific categories. In this work, only the natural disasters concerning the time span of our data were retrieved. In total 4,913 events were downloaded with the most important features of the dataset being: disaster type, event name, country, coordinates, start and end date, total deaths, total affected, and total damages in US dollars. There are many null values in some features. The most noticeable ones are coordinates (3,623 nulls), total damages in US dollars (3,158 nulls), total affected (1,700 nulls), and total deaths (1,330 nulls).

3.3. Algorithms and methods

In this section we describe the machine learning methods that were used to create the dimensions of stance, sentiment, and aggressiveness.

For enriching the Twitter dataset by making predictions with Transfer Learning, we employed a number of methods and supervised learning algorithms. Our simpler, conventional classifiers were a linear Support Vector Machine (SVM) and a multinomial Naive Bayes. We pre-processed the tweets using the techniques by Effrosynidis et al. (2017), Symeonidis, Effrosynidis, and Arampatzis (2018) and employed a TF-IDF vectorizer to transform our tweets into vectors of bigrams and trigrams for training these models. We also created word embeddings with vectors of 200 dimensions from our tweets using GloVe in order to train a Convolutional Neural Network and a Recursive Neural Network (LSTM). Some research suggests that, when compared to the most obvious alternative — Word2vec, GloVe performs better on Twitter sentiment analysis tasks (Stojanovski, Strezoski, Madjarov, & Dimitrovski, 2015). GloVe has also been successfully used in our previous, Twitter sentiment analysis based work (Karasakalidis, Effrosynidis, & Arampatzis, 2021). Furthermore, we trained a Transformer model based on Google's renowned BERT. More specifically, we fine-tuned the 'BERT-Base, Uncased' model that was used in the original paper published by Google (Devlin, Chang, Lee, & Toutanova, 2018). That model was pre-trained on a combined corpus of the English Wikipedia and the BooksCorpus. The above-mentioned models were trained and evaluated on the Google Colab. The GPU the platform provides – and hence the GPU used for training – is an Nvidia Tesla K80. On the other hand, the LSTM and CNN models were trained and evaluated using an Nvidia GTX 1070. The target measure for training these models was the F_1 score, while their architecture can be seen in Fig. 2.

Finally, the unsupervised learning models we used for the sentiment analysis of tweets have been mentioned in Section 3.2. We decided to combine the output of these models by taking their average prediction for each tweet. The models are briefly mentioned below:

- SVM: A Support-Vector Machine is a supervised machine learning algorithm used for both regression and classification purposes (Vapnik, 1999). For classification tasks, it finds the hyperlanes that best divide data into two or more classes.
- Naive Bayes: A probabilistic model used for classification, employed on our transfer learning tasks, using a multinomial event model (Kibriya, Frank, Pfahringer, & Holmes, 2004).
- CNN: Deep Learning models where each neuron in a certain layer is connected to all neurons in the next layer.
- RNN (LSTM): LSTM is a type of Recurrent Neural Network that has feedback connections giving it the ability to learn long-term dependencies.
- BERT: Google's Transformer-based pre-trained model for classification, capable of achieving state-of-the-art performance in sentiment analysis (Devlin et al., 2018).
- VADER: A rule-based, lexicon-based model for sentiment analysis of social media text used in numerous academic works (Hutto & Gilbert, 2014).
- Textblob: A lexicon-based model for sentiment analysis (Loria, 2018). Based on Python's 'Pattern' library (De Smedt & Daelemans, 2012), it is designed for sentiment analysis in a more general context than VADER.
- Flair: Packing two pre-trained models, an RNN-based and a BERT-based, both trained on IMDB's movies dataset (Akbik, Blythe, & Vollgraf, 2018; Maas et al., 2011). Regardless of their training corpus, we experimented with them on our sentiment analysis task.

In regards to model training, we focused on the BERT model from early on, as it outperformed the rest of the models both in sentiment and aggressiveness classification. Nevertheless, we tested different preprocessing techniques and vectorizers for training the SVM/NB models and different structures and target measures for our CNN/LSTM models as well. While these models proved inferior to the BERT model, their faster training allowed us to experiment further. We employed a linear SVM classifier and a multinomial Naive Bayes (NB) from Scikit-Learn python library (Pedregosa et al., 2011) with their default hyper-parameters. We pre-processed our tweets by removing URLs, replacing user mentions with the username, we removed hashtags and numbers, we replaced contractions and elongated words, and transformed the processed tweets into TF–IDF vectors using bigrams and trigrams.

As we have already mentioned, we used 200 dimension GloVe to translate the tweets into vectors for training our CNN and LSTM models. Their architecture is depicted in Fig. 2 and they were implemented using the Keras API.[11] The loss function used was categorical or binary cross-entropy – depending on the classification task – while the

[10] https://public.emdat.be/data.

[11] https://keras.io.

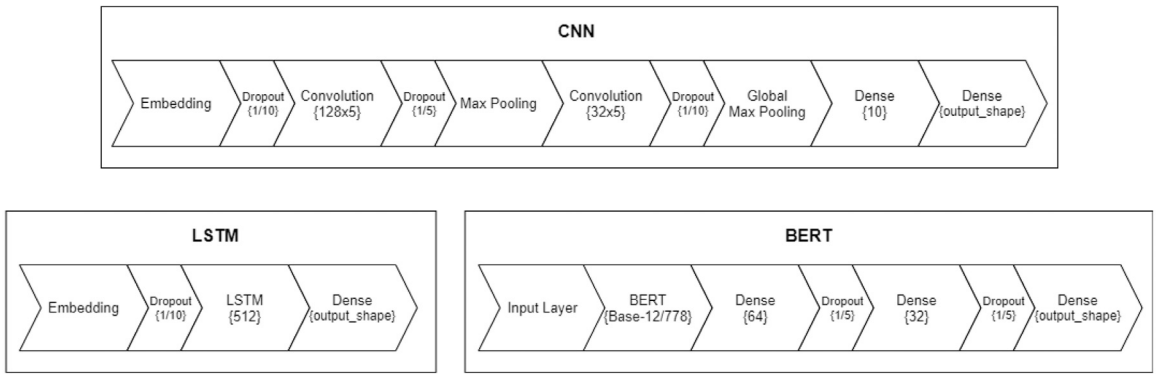

Fig. 2. CNN, LSTM and BERT layer structure.

created_at	id	lng	lat	topic	sentiment	stance	gender	temperature_avg	aggressiveness
2006-12-14 01:39:10+00:00	1092823	-122.41942	37.77493	Ideological Positions on Global Warming	-0.544194675	neutral	male	4.2285404	aggressive
2009-11-06 06:44:52+00:00	5472772435	153.02809	-27.46794	Global stance	0.358623818	neutral	male	0.29895014	aggressive
2009-12-21 19:14:27+00:00	6902468428	-105.2705456	40.0149856	Seriousness of Gas Emissions	-0.342463109	believer	undefined	5.465174	not aggressive
2010-07-20 23:27:57+00:00	19030860012	-1.7907973	53.8385598	Importance of Human Intervantion	-0.277766986	denier	male	2.5027697	not aggressive
2011-03-29 09:21:17+00:00	5.26616E+16	121.0699173	14.5288867	Weather Extremes	0.538100337	neutral	male	-1.2430557	not aggressive

Fig. 3. A sample of five rows of the Climate Change Twitter Dataset.

optimizer we used was the Adam. Their data batch sizes were 1024 and 512 respectively. They were both trained for 100 epochs, saving a checkpoint of their weights each time their validation F_1 score peaked, while the training stopped after 10 epochs of non-improving validation loss. Finally, we reduced the learning rate of 10^{-5} and 10^{-4} respectively by a factor of 0.85, each time validation loss was worsening for two consecutive epochs.

Using dense layers to translate the transformer's outputs, we trained the BERT model back-propagating the error through its entirety, totaling 109,533,636 parameters. All layers but the output layers used relu as an activation function. We used the AdamW optimizer (Loshchilov & Hutter, 2017), an improvement over the Adam while using categorical or binary cross-entropy as a loss function, again, depending on the classification task at hand. We tested different batch sizes and learning rates, and decided to use 32 and 10^{-5} respectively. Finally, we dynamically reduced the learning rate of the model while training by a factor of 0.4 when the loss on the validation set stopped improving and saved the model's weights when the F_1 score was at its peak for the validation set.

In all the deep learning models that we trained, we introduced dropout regularization in order to combat over-fitting. The dropout introduced in the CNN and LSTM models was based on numerous different runs, while the dropout introduced on BERT was less evaluated due to its long training time. We should also note that the structure of the BERT model is identical for both the classification tasks it served. The CNN and LSTM models were more fine-tuned for each task, in regards to their batch size, learning rate, layer activation functions, and dropout regularization.

4. Experiments and results

The main result and contribution of this work, is the produced Climate Change Twitter Dataset. A sample of five tweets is given in Fig. 3.

In the rest of this section, we will talk about how the dimensions of stance, sentiment, aggressiveness and topics were created, evaluated and compared from a plethora of machine learning methods. We will also conduct a brief volume, spatial and temporal analysis of the data, and finally, discuss how this dataset can be used.

4.1. Climate change stance classification

In this section, we will compare the different machine learning models we used for Transfer Learning and justify the choice of the BERT model.

Before pre-processing the Stance training dataset 3.2.3 – that was later used for training our models – we checked the IDs of its tweets to see if they match with tweets from our Twitter data. 6,416 of these tweets were found to be included in our own dataset, so they were separated from the Stance training dataset and became the test set that we used to evaluate our trained models. We split the remaining data to two sets – one for training (22,601 tweets) and one for validation (5,650 tweets) – while we used the above-mentioned 6,416 tweets as a test set for evaluating our models. These models were also evaluated using 5-fold cross-validation on the Stance training dataset, each fold consisting of 5,650 tweets used for validation while training on the other 22,601 tweets. Furthermore, each fold was created with the same label distribution as the label distribution of the corresponding training set. The number of folds was selected based on our limited resources — a larger number of folds would result in an intolerable amount of training time for the BERT models. Accuracy, precision, recall and F_1, were the measures calculated for each model. These measures can be seen in Tables 2 and 3. Based on these evaluation steps, we fine-tuned our models as discussed in Section 3.3.

Naive Bayes was certainly the worst performing model with an F_1 score being behind the SVM by 0.12 in cross validation and 0.08 in evaluation set, at a range of 0.52–0.55. The CNN and LSTM models had a much better performance than NB and SVM both in cross validation and on the evaluation set. Although our BERT model's accuracy on the evaluation set only slightly outperformed the accuracy of CNN and LSTM, it was still much better able to differentiate between the classes (Denier-Neutral-Believer) as it had an F_1 score of more than 0.75, i.e. around 10% higher than the second best model.

Although using dropout layers for the deep learning models while trying to reduct hidden layers and experimenting with early-stopping and dynamically reducing the learning rate based on the performance of the models on the validation data, overfitting persisted as accuracy, precision and recall had a 10% difference at best between the cross-validation (Table 2) and test (Table 3) scores. Although this phenomenon was not observed with the NB and SVM models, their already poor performance on our data was enough to deter us from

Table 2
Stance classification — evaluating the models trained via transfer learning using five-fold cross-validation. The scores shown on the table are the average of accuracy, precision, recall and F_1 scores of all folds. Variance is included in parenthesis. Precision, recall and F_1 scores (for each fold) were calculated using macro averaging.

	Accuracy	Precision	Recall	F_1
NB	0.7319 $(1.36 \cdot 10^{-5})$	0.7533 $(6.23 \cdot 10^{-5})$	0.4964 $(2.00 \cdot 10^{-5})$	0.5223 $(4.16 \cdot 10^{-5})$
SVM	0.7663 $(1.23 \cdot 10^{-5})$	0.7356 $(4.85 \cdot 10^{-5})$	0.6071 $(5.97 \cdot 10^{-5})$	0.6418 $(4.57 \cdot 10^{-5})$
CNN	0.8558 $(4.21 \cdot 10^{-6})$	0.8313 $(4.09 \cdot 10^{-5})$	0.7684 $(3.84 \cdot 10^{-5})$	0.7986 $(2.37 \cdot 10^{-4})$
LSTM	0.8285 $(1.08 \cdot 10^{-5})$	0.7942 $(1.21 \cdot 10^{-5})$	0.7295 $(1.49 \cdot 10^{-5})$	0.7605 $(5.87 \cdot 10^{-5})$
BERT	0.8789 $(4.16 \cdot 10^{-4})$	0.8666 $(5.37 \cdot 10^{-4})$	0.8392 $(9.27 \cdot 10^{-4})$	0.8527 $(7.17 \cdot 10^{-4})$

Table 3
Stance classification — transfer learning models trained using the entirety of the training data (22,601 tweets) and evaluated on the test set (6,416 tweets). Macro average used for calculating precision, recall and F_1 scores.

	Accuracy	Precision	Recall	F_1
NB	0.7180	0.7245	0.5092	0.5438
SVM	0.7414	0.6888	0.5935	0.6236
CNN	0.7804	0.7479	0.5841	0.6559
LSTM	0.7812	0.7636	0.5773	0.6575
BERT	0.7884	0.7802	0.7406	0.7599

Table 4
Stance classification — label counts on training data, evaluation data and the Climate Change Twitter Dataset. Percentage of each label count (out of total labels of each dataset) is included in parenthesis. The label are analogously distributed in all 3 datasets.

Label	Training data	Evaluation data	Climate Change Twitter Dataset
'-1'	3,373 (11,94%)	617 (9,62%)	1,191,386 (7,54%)
'0'	5,972 (21,14%)	1,743 (27,16%)	3,305,601 (20,94%)
'1'	18,906 (66,92%)	4,056 (63,22%)	11,292,424 (71,52%)

any further experimentation. Our conclusion being that enriching our training data further has to be investigated to tackle overfitting.

Our BERT model's performance seems promising, even when compared to other related research works. Not only does it have a better performance than any other machine learning algorithm employed for classifying climate change related tweets, but we are amongst the first researchers to use it in this field. These are the reasons we elected BERT for predicting climate change stance on our data. Nonetheless, the training time required for the BERT model was around 5 h (293 min). If we are to incorporate the training time for every model on cross-validation (for BERT), that would result on adding approximately 7 h and 20 min (440 min) to the previous time. Hence, for evaluating a single BERT model, it would result to about 12 h and 20 min of training time.

While trying to explain the uneven classes in the labels of the training data (see Table 1), we realized that the same imbalance was found on the evaluation data. This realization led us to the belief that this imbalance could be a good representation of real-world data. This belief was further strengthened while we were manually revising both the unknown data and our algorithm's predictions on that data. You can see the count of each class for each dataset on Table 4.

4.2. Climate change sentiment and aggressiveness

In this section, we will compare the different methods we used to extract sentiment aggressiveness from the tweets.

The unsupervised models we have used for sentiment prediction have already been discussed in Sections 3.2 and 3.3. In order to get an insight into the performance of these models, we wanted to evaluate them on labeled data. It is common in many research works where models like VADER are used, that the researchers trust the results of such pre-trained models. We wanted to have a picture of the performance of these models on predicting sentiment for tweets from our work's domain. With that in mind and while we were not able to find tweets with sentiment labeled on a continuous scale, we decided to transform our predictions into a discrete scale, in order to evaluate them based on the popular Sentiment140 dataset (Go, Bhayani, & Huang, 2009).

Table 5
Sentiment regression — evaluating the unsupervised models on Sentiment140 data. The outputs of the models was transformed from continuous (−1 to 1) to categorical (−1, 0 or 1) for evaluation. The average of all models balances precision and recall scores.

	VADER	Textblob	Flair (LSTM)	Flair (BERT)	Average
Accuracy	0.6612	0.6076	0.7011	0.6996	0.7214
Precision	0.6083	0.5672	0.6941	0.6999	0.7069
Recall	0.8998	0.8997	0.7164	0.6959	0.7539
F_1	0.7259	0.6957	0.7051	0.6979	0.7296

The evaluation set (Sentiment140) was balanced in regards to the labels, so we translated the continuous values of the models evenly, as: [−1, −0.35] Negative Sentiment / (−0.35, 0.35) Neutral Sentiment / [0.35, 1] Positive Sentiment. After sampling 100,000 tweets from that set in a balanced manner in regards to the labels, we performed the evaluation and measured the accuracy precision, recall and F_1 score for each of our models. These measures can be seen in Table 5. The individual models gave satisfactory performances, varying in F_1 and accuracy. The highest F_1 was achieved by VADER (72.59%), followed by Flair LSTM (70.51%), while the highest accuracy was observed with Flair LSTM (70.11%), followed by Flair BERT (69.96%). The unsupervised VADER and Textblob models, traded precision for recall, while the Flair models had balanced results. Nonetheless, the generalized nature of the models, as well as the translation of their outputs from continuous values to categorical, are factors that made the regression task difficult. The average of the chosen models seemed to provide the most accurate results out of all individual four, with an F_1 of 72.96% and an accuracy of 72.14%, so this is the one we will choose.

The −+0.35 threshold that was used for transformation of the continuous sentiment to discrete was not randomly chosen. It represents the points where a clear separation between the distributions of 'Negative'-'Neutral' and distributions of 'Neutral'-'Positive' sentiments can be spotted on our data when using the 'Average' of the predictions. While analyzing the evaluation results for sentiment, we found that the accuracy for the 'Negative' sentiment was just over 30%, far less than the 'Neutral' and 'Positive' sentiments. As these models come pre-trained, there is only a small overhead for corrections. The far better method of transfer learning that we used in stance and text aggressiveness classification was unfeasible due to the lack of such, publicly available, Twitter data.

Evaluating the Transfer Learning models for the aggressiveness of a tweet was challenging, as there was a lack of accessible, labeled Twitter data on this subject. Considering that, we validated our models using cross-validation and we built on its results. We split the training data into 10 folds, each consisting of 7,200 tweets for training, 900 for validation and 900 for testing — all of them being balanced in regards to the labels. The choice of splitting the data in 10 folds was made due to the far smaller amount of data (when compared to the data used for stance transfer learning). The evaluation results can be seen in Table 6.

All models – but BERT – had similar performance. Text aggressiveness seems to be a relatively easy classification task as even the simplest

Table 6
Aggressiveness classification — evaluating the transfer learning models using ten-fold cross-validation. The scores are calculated by averaging accuracy, precision, recall and F_1 of all folds accordingly while their corresponding variance is in parenthesis.

	SVM	NB	CNN	LSTM	BERT
Accuracy	0.8626 $(3.26 \cdot 10^{-4})$	0.8676 $(1.57 \cdot 10^{-4})$	0.8600 $(6.52 \cdot 10^{-5})$	0.8709 $(4.25 \cdot 10^{-5})$	0.9740 $(1.19 \cdot 10^{-3})$
Precision	0.8631 $(3.26 \cdot 10^{-4})$	0.8679 $(1.56 \cdot 10^{-4})$	0.8693 $(4.99 \cdot 10^{-4})$	0.8637 $(3.92 \cdot 10^{-4})$	0.9706 $(1.26 \cdot 10^{-3})$
Recall	0.8626 $(3.25 \cdot 10^{-4})$	0.8675 $(1.56 \cdot 10^{-4})$	0.8572 $(4.00 \cdot 10^{-4})$	0.8841 $(2.92 \cdot 10^{-4})$	0.9782 $(1.13 \cdot 10^{-3})$
F_1	0.8650 $(3.29 \cdot 10^{-4})$	0.8687 $(1.83 \cdot 10^{-4})$	0.8628 $(2.92 \cdot 10^{-4})$	0.8697 $(1.18 \cdot 10^{-4})$	0.9742 $(1.19 \cdot 10^{-3})$

Table 7
Text aggressiveness classification — label count on the training data and the Climate Change Twitter Dataset. The percentage of each label count (out of total labels of each dataset) is included in parenthesis.

Label	Train	Climate Change Twitter Dataset
'0'	3,909 (50%)	11,262,144 (71,33%)
'1'	3,909 (50%)	4,527,267 (28,67%)

classifiers can perform on par with more complex ones, while BERT can achieve near-perfect results — it scored more than 10% higher both in accuracy and F_1 than any other model. This is also supported by the performance of the SemEval's competition participants — the competition we collected part of our training data from 3.2.5. The slightly higher variance in scores – when compared to the score variance of 5-fold cross-validation for stance classification – can be explained by the larger, 10-fold cross-validation. This evaluation step shows once more the cutting edge performance of transformer based models. Due to the significantly smaller training dataset than the one used for stance classification, the training time of the BERT model was reduced to 34 min (93 min for evaluating using cross-validation). Nonetheless, the scores shown in Table 6 are based only on cross-validation whereas a public, independent and hand-labeled dataset was not to be found for a more thorough evaluation. Thus, the high scores from cross-validation may warn us of over-fitting. We can get a hint that this may not be a possible issue by looking at the label counts of aggressive/non-aggressive texts in the training and the predicted data (see Table 7), as the former's labels differs from the latter's. This can only be an estimation though, since we have no insights of a real-world data on the subject.

The most valid method that is obvious for us to explore this possible issue would be to create a gold-standard dataset of tweets with appropriate labeling and testing our models on it. This applies not only to the aggressiveness classification but to both stance and sentiment classification. We make the argument though, that creating such datasets would outdo our methodology of transfer learning, as these datasets could be used for training instead of evaluating our models.

4.3. Topic modeling

Using LDA, we discovered 10, 15, and 20 topics, and evaluated them using five methods. In this section we will discuss the evaluation results.

The four measures of evaluation are shown in Table 8, while the graphical LDAvis is shown in Fig. 4. Perplexity is only shown for a sanity check and completeness, as it is the default measure produced by the gensim LDA module. A lower perplexity indicates that the data is more likely. It is expected to decrease as the number of topics increases. Perplexity has real meaning when comparing models that produce the same number of topics.

We emphasize more on the per-topic evaluation measures. Probability shows how common a topic is across all tweets. Too small probabilities indicate very specific topics that might not contain enough information and not be sufficient for analysis, and too big probabilities indicate that the topic could be split further. We set a healthy range of probabilities to be between 0.10 and 0.50. It is also desirable, although not possible, for the topics to have equal probabilities. In that respect, the 10 topics model has probabilities that range from 0.1116 to 0.3575 (3.2 times difference between largest and smallest topic, the 15 topics model from 0.0903 to 0.3091 (3.4 times difference), and the 20 topics model from 0.0549 to 0.2515 (4.58 times difference).

The lower entropy values of the 10 topics model imply that the topics of this model are more distributed across tweets. While it is expected for the KL divergence to increase with more topics, this increase is not marginal in our case. The means of KL of the 3 models are 0.59, 0.62, and 0.66 respectively, suggesting that even when we make more distinct topics, the topic distributions from the overall distribution do not change significantly.

When the topics are seen visually, it is clearer and easier for the human eye to conclude that the 10 topic model produces much more well-separated topics. The other two models have many overlapping topics.

In addition to the four measures and the one graphical, 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.

Thus, according to these evaluation criteria, the 10 topic model will be used, combined with the fact that 10 topics are more manageable.

The ten topics discovered by the algorithm are depicted in Fig. 5. 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).

4.4. Volume analysis

In this section, we will conduct a brief volume, spatial and temporal analysis of the data in order to understand them.

The number of tweets that make up the data, as mentioned in Section 3.1.1, is 15,789,411. They span from June 6, 2006, to October 1, 2019. One third of them (5,307,538) has geolocation information as discussed and evaluated in Section 3.2.1. The most common hashtag is #climatechange followed by #globalwarming. Twitter gives the option to include a website link in a tweet as a reference for the context of the tweet. As tweets have the power to influence others, it is crucial to investigate these external references. Most of the URLs present in tweets are not from formal news domains, but rather from social media domains like Twitter, Google, YouTube, Medium, and Facebook. After excluding them, along with other unrelated sources, such as URL shortening links, we are left with about 30% of the initial websites. The top-40 such domains are presented in Table 9, with relative frequencies normalized by the count of the first of the list. The most referenced domain in Twitter climate change discussions is theguardian.com, followed by washingtonpost.com and nytimes.com.

If we plot each one of the 5,307,538 geolocated tweets on the world map we get Fig. 6. Since the keywords and hashtags used to create the dataset are in the English language, there is a bias on the origin locations of the tweets. Most of the tweets are located in the United States of America, Canada, United Kingdom, Australia, New Zealand, but also in Europe, where English is very common as a

Fig. 4. LDA visualization on the two dimensional space. The three models displayed (from left to right) consist of 10, 15, and 20 topics.

Table 8
Evaluation of topic modeling using LDA with 10, 15 and 20 topics.

	10 topics model			15 topics model			20 topics model		
Perplexity	−7.001			−7.448			−7.88		
Topics	Probability	Entropy	KL	Probability	Entropy	KL	Probability	Entropy	KL
1	0.1116	6.455	0.9489	0.0903	7.149	0.6343	0.0549	6.246	1.8130
2	0.1477	6.944	0.4601	0.1184	7.083	0.6686	0.0803	6.660	0.6666
3	0.1565	7.552	0.7027	0.1218	7.251	1.6760	0.0832	6.656	0.8581
4	0.1604	6.762	0.4112	0.1261	6.723	0.5434	0.0848	6.765	0.6145
5	0.1607	6.998	0.5637	0.1343	6.912	0.6562	0.1011	6.979	0.6893
6	0.1609	6.716	1.0721	0.1377	6.840	0.5337	0.1036	7.241	0.8048
7	0.2457	6.867	0.3170	0.1383	6.906	0.6087	0.1043	6.964	0.6566
8	0.2498	7.240	0.7048	0.1497	6.524	0.4682	0.1097	6.784	0.4968
9	0.2963	6.613	0.5523	0.1605	6.763	0.4143	0.1160	6.683	0.5448
10	0.3575	7.121	0.2076	0.1620	6.845	0.5091	0.1170	7.440	0.8521
11				0.1681	6.584	0.4201	0.1183	7.220	0.5072
12				0.1874	7.057	0.8930	0.1252	6.511	0.5615
13				0.2179	7.178	0.4103	0.1280	6.287	0.4847
14				0.2344	7.094	0.3274	0.1489	6.828	0.4233
15				0.3091	6.727	0.6089	0.1540	6.623	0.4757
16							0.1733	7.131	0.4492
17							0.2088	6.587	0.6232
18							0.2097	7.036	0.3575
19							0.2217	7.383	0.7176
20							0.2518	6.999	0.7459

Table 9
Most referenced climate change websites excluding social media domains (e.g. Twitter, Google, YouTube, Medium); relative percentages, normalized by the count of the first of the list.

Domain	%	Domain	%	Domain	%	Domain	%
theguardian.com	100.00	thinkprogress.org	11.06	forbes.com	6.75	thehill.com	4.63
washingtonpost.com	67.01	ft.com	10.49	nationalgeographic.com	6.68	thestar.com	4.62
nytimes.com	53.48	dailymail.co.uk	10.33	worldbank.org	6.27	wattsupwiththat.com	4.50
bbc.com/co.uk	28.77	abc.net.au	9.73	foxnews.com	5.81	ecowatch.com	4.49
cnn.com	22.99	grist.org	9.28	nbcnews.com	5.49	cbsnews.com	4.47
reuters.com	20.21	nature.com	9.25	telegraph.co.uk	5.37	dailykos.com	4.22
huffingtonpost.com	14.62	latimes.com	8.81	scientificamerican.com	5.36	europa.eu	3.88
cbc.ca	13.06	weforum.org	8.43	nrdc.org	5.11	sciencemag.org	3.31
smh.com.au	11.73	bloomberg.com	7.37	nasa.gov	5.04	theatlantic.com	3.08
independent.co.uk	11.62	npr.org	7.13	businessinsider.com	4.67	theeconomist	2.94

second language. Countries with high populations like India, Japan, the Philippines, Indonesia, and Malaysia have also a high number of English tweets. Tweets in China are almost absent because the country has blocked Twitter. Other remarkable locations with tweet activity are the Gulf of Guinea, Uganda, Kenya, South Africa, Central America, and the East coast of Brazil.

Regarding basic temporal analysis, Fig. 7 illustrates the number of tweets over different sampling periods. At the top left, the number of tweets for each year in the dataset is visible. As the years pass by, Twitter activity is increased, especially after 2014, with a peak of 6 million tweets in 2018. At the top right, the total number of tweets written in each month of the year is presented. The months of December and May are the most tweeted, while the lowest activity happens during the summer months. At the bottom left, the day-of-the-week activity is shown. It is clear that people do not tweet as much during weekends. Finally, at the bottom right, the activity over the hour-of-the-day can be seen. Most tweets are posted between 13:00 and 22:00, with a peak at 16:00 (local times).

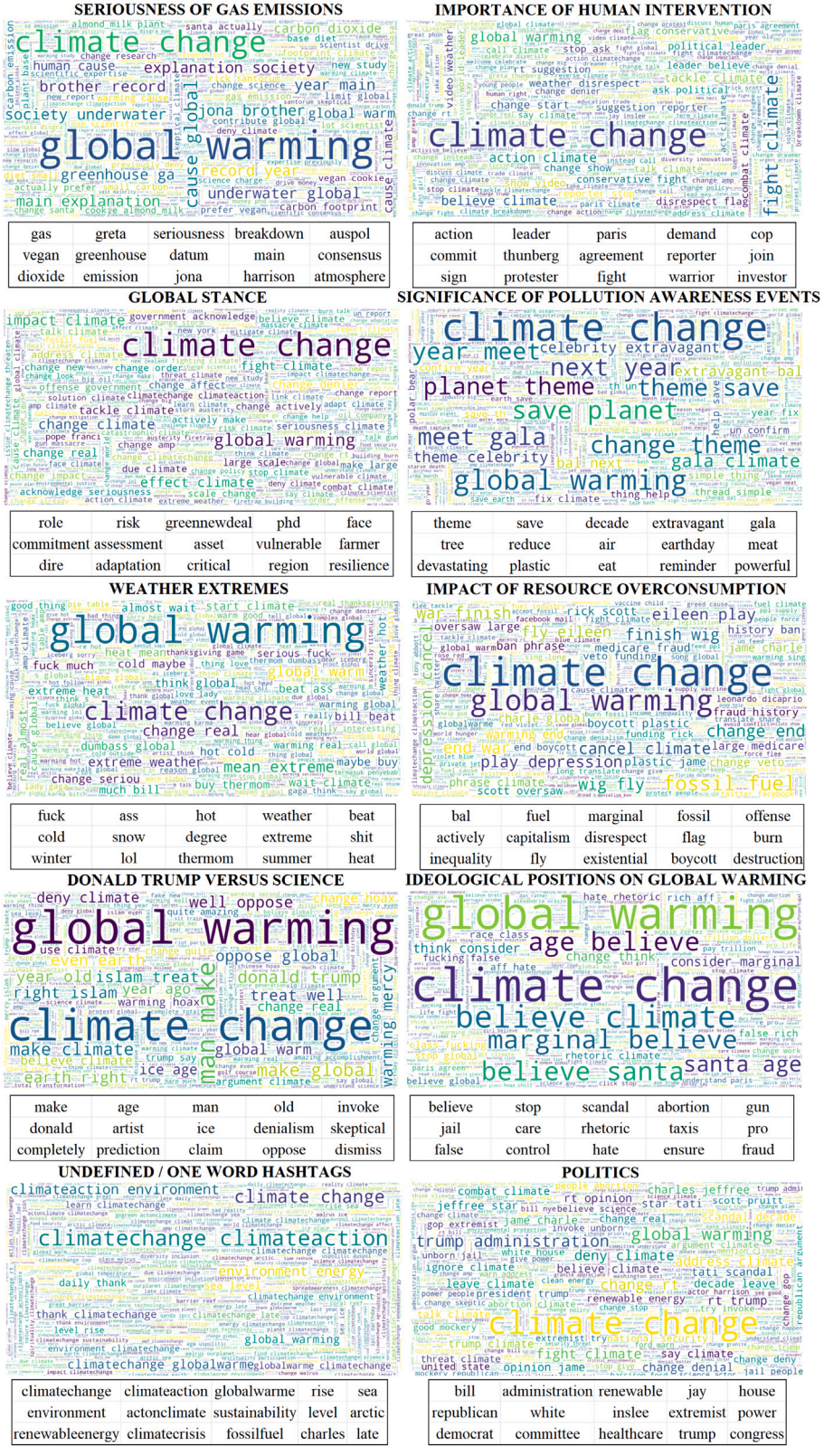

Fig. 5. Ten topics discovered by the LDA algorithm. For each topic, there are available: its title, the word cloud with the most common 1000 words and the 15 most unique words that were used to determine the topic's title.

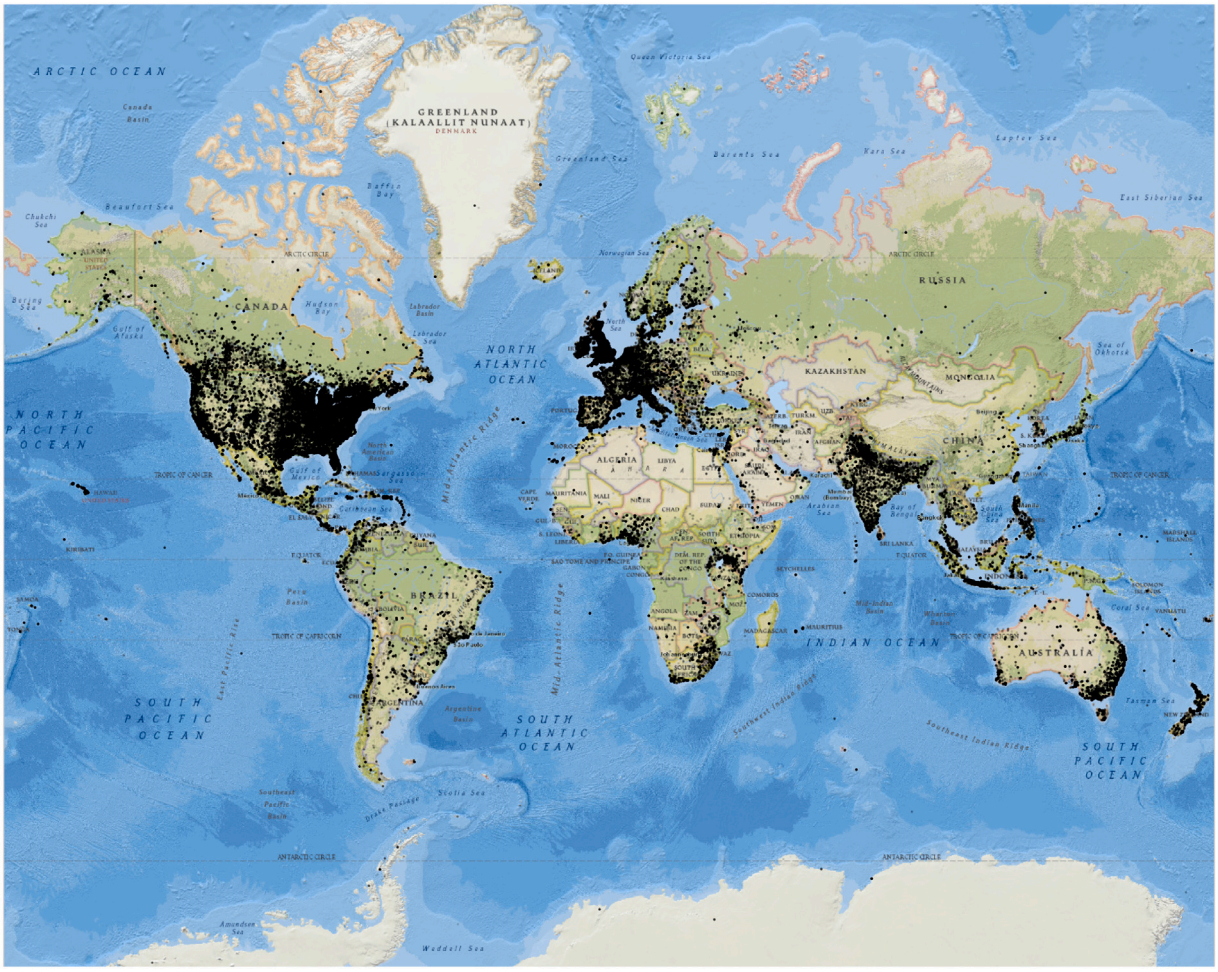

Fig. 6. Climate change tweets (black dots) on the world map.

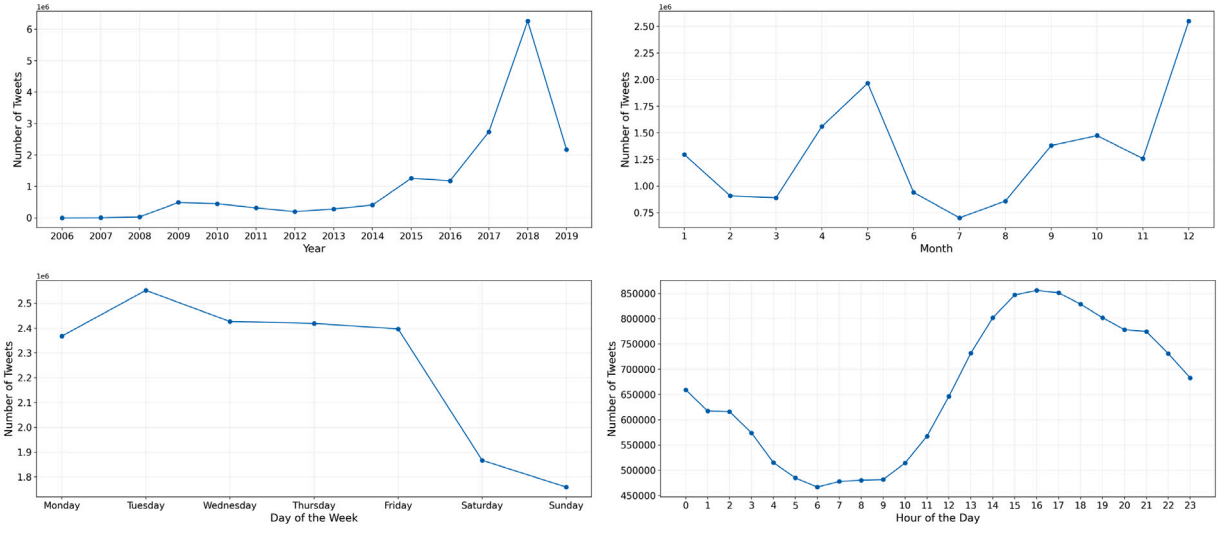

Fig. 7. Temporal analysis of the tweets on different time sampling periods.

Table 10
Ten examples of tweets accompanied with topic, sentiment, stance, aggressiveness, and temperature deviation from historic average as assigned by the proposed models and methodology.

#	Tweet text	Topic	Sentiment	Stance	Aggressiveness	Temperature deviation
1.	*Canada's Prime Minister Stephen Harper announces a new $1.3bn fund to combat climate change.*	Importance of Human Intervention	−0.12	believer	not aggressive	3.32
2.	*Well, no snow. At least I'm not in Alaska where its -78 degrees! That's NEGATIVE SEVENTY-EIGHT flippin degrees! Damn that global warming.*	Weather Extremes	−0.74	believer	aggressive	1.46
3.	*Why does MSM always accept assumption that CO2 emissions cause global warming? Fail. #msmbias #tcot*	Seriousness of Gas Emissions	−0.54	denier	not aggressive	−5.83
4.	*Gov. Rick Scott banned the term 'climate change' in Florida. The state was just hit by its worst storm ever.*	Importance of Human Intervention	−0.94	believer	not aggressive	0.35
5.	*This humidity is oppressive. It's 6:10am and 27C with 88% humidity. That's unnatural. Which moron said climate change doesn't exist?!*	Weather Extremes	−0.93	believer	aggressive	2.00
6.	*Wrapping up a wonderful presentation from IFTF board member, Larry Smarr, on climate change opportunities and dilemmas in the ICT world.*	Global stance	0.93	believer	not aggressive	−0.92
7.	*Congrats to Paul and the environment team! And thank you to all the residents and groups over the years for pushing our city to do our part in fighting climate change. Amazing work to celebrate today and build upon in days to come! #ygk #ClimateAction*	Importance of Human Intervention	0.91	believer	not aggressive	7.46
8.	*30 degrees Celsius outside? So much for global warming...Canada could use warmer winters!* #copenhagen #climate #climatechange	Weather Extremes	0.07	denier	not aggressive	−16.14
9.	*Anti-Trump chef to Trump: If climate change isn't real, why are you building a sea wall at your golf course?*	Donald Trump versus Science	−0.44	believer	aggressive	−13.19
10.	*If at this point you still don't believe human activity has a huge effect on #climatechange you are either really uninformed or an idiot.*	Ideological Positions on Global Warming	−0.55	believer	aggressive	15.81

4.5. How can this dataset be used?

The most important use of this dataset is to use it through exploratory data analysis in order to extract valuable information regarding climate change and human opinion. Some research questions that can be answered by utilizing this dataset are the following: 'How do climate change deniers differ from believers?', 'Where around the globe the climate change denier/believer ratio is high?', 'Is there any correlation between human sentiment and deviations from historic temperature?', 'Is there any correlation between each of the discovered topics and the climate change stance, sentiment, and aggressiveness?'.

Another way to use this dataset is to train new machine learning models. But can this dataset be used to train new models with supervised methods? All produced dimensions except temperature deviations were computed by an algorithm or a model. This means that this dataset has silver standard labels and not gold. Generating large training data with manual gold labels is very expensive and time consuming. When gold standard labels are created, they are usually very limited in size (one thousand to several thousands), and this is for the case of only one label (i.e. stance). Our dataset has produced over 15 million silver labels for six labels/dimensions. We also note here that the dimension of topics, possibly, could not be created with gold labels, as the topics are not known in advance. So, it arises the question of quality versus quantity. Would a model trained on several thousands of gold-labeled tweets be superior of another model trained on millions of silver-labeled tweets? In the work of Tekumalla and Banda (2021), the authors do this experiment on Twitter data. Their gold-labeled set is consisted of 14,430 tweets and their silver-labeled set is consisted of seven datasets of sizes 100k, 200k, 300k, 500k, 1M, 2M, and 3M respectively. They showed that silver and gold standard datasets had similar performances. Using a BERT model, the silver standard dataset scored 99.51%, and the gold standard dataset 99.78%. They also demonstrated that with the increase in training size there is an increase in the performance. Another important factor that affects the quality of the silver labels is the performance of the algorithm that computed them. Our methods, according to our evaluations, have achieved high accuracy for all dimensions: 90% accuracy for gender when evaluating on a gold standard set, 87.89% accuracy for stance when doing cross-validation and 78.84% when evaluating on a gold standard set, 72.12% accuracy for sentiment when evaluating on gold standard set, and 97.4% for aggressiveness when doing cross-validation. In addition, Twitter forbids to share the text of the tweets and only allows to share tweet IDs. Since users delete their tweets or accounts, many tweets and especially older ones cannot be retrieved. In fact, we only retrieved 68.56% tweets from the Credibility of Climate Change Denial in Social Media dataset and 64.96% from the Climate Change Tweets IDs dataset. Another research study that tried to retrieve tweets from IDs, retrieved 66% of the original data (Cocos, Fiks, & Masino, 2017). Thus, a manually labeled gold standard dataset from tweets could not be reproduced, along with the original models of the research related to it. Consequently, for Twitter data, we can avoid manual annotation of small datasets as they are not backward compatible and have similar performance with big silver standard datasets. According to the previous, the Twitter Climate Change Dataset can be used to train new models with supervised methods.

Some example tweets together with the outcomes of our models for the dimensions of topic, sentiment, stance, aggressiveness and temperature deviation are visible in Table 10.

As a final note, there are three potential biases in the data. The queries used to retrieve them are in English, the data is a merge of three datasets, and the volume of tweets is increasing over time as Twitter grows in popularity. In order to deal with the biases, analyses on this data should be proportional and not absolute.

5. Conclusion

This work delivers a comprehensive dataset in both time and space, but also in volume. A total of 15,789,411 tweets related to climate change and global warming are included, spanning over 13 years. For 5,307,538 of them, the exact geolocation was reverse-engineered with high precision. The dataset was enriched with seven extra dimensions namely gender, stance, sentiment, aggressiveness, temperature, topics, and disasters. These dimensions were produced by testing and evaluating a plethora of state-of-the-art machine learning algorithms and methods, both supervised and unsupervised. The dataset is publicly available.

As a further study, we would like to use this dataset to unveil critical insights about climate change and global warming discourses on Twitter. We would also like to create our own manually labeled datasets for the dimensions of stance, sentiment, and aggressiveness and test if they produce better models than using Transfer Learning.

CRediT authorship contribution statement

Dimitrios Effrosynidis: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data

curation, Writing – original draft, Writing – review & editing, Visualization. **Alexandros I. Karasakalidis:** Methodology, Software, Validation, Writing – review & editing. **Georgios Sylaios:** Validation, Formal analysis, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Avi Arampatzis:** Conceptualization, Validation, Formal analysis, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by European Union's Horizon 2020 European Green Deal Research and Innovation Program (H2020-LC-GD-2020-4), grant number No. 101037643 – ILIAD (Integrated Digital Framework for Comprehensive Maritime Data and Information Services). The article reflects only authors' view and that the Commission is not responsible for any use that may be made of the information it contains.

References

Abdar, M., Basiri, M. E., Yin, J., Habibnezhad, M., Chi, G., Nemati, S., et al. (2020). Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets. *Renewable and Sustainable Energy Reviews*, *124*, Article 109781.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 annual conference of the north american chapter of the association for computational linguistics* (pp. 54–59).

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th international conference on computational linguistics* (pp. 1638–1649).

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87–93.

An, X., Ganguly, A. R., Fang, Y., Scyphers, S. B., Hunter, A. M., & Dy, J. G. (2014). Tracking climate change opinions from twitter data. In *Workshop on data science for social good* (pp. 1–6).

Anderson, A. A., & Huntington, H. E. (2017). Social media, science, and attack discourse: How Twitter discussions of climate change use sarcasm and incivility. *Science Communication*, *39*(5), 598–620.

Baylis, P. (2020). Temperature and temperament: Evidence from Twitter. *Journal of Public Economics*, *184*, Article 104161.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc..

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brulle, R. J., Carmichael, J., & Jenkins, J. C. (2012). Shifting public opinion on climate change: an empirical assessment of factors influencing concern over climate change in the US, 2002–2010. *Climatic Change*, *114*(2), 169–188.

Chen, E., Lerman, K., Ferrara, E., et al. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, *6*(2), Article e19273.

Chen, X., Zou, L., & Zhao, B. (2019). Detecting climate change deniers on twitter using a deep neural network. In *Proceedings of the 2019 11th international conference on machine learning and computing* (pp. 204–210).

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941.

Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association*, *24*(4), 813–821.

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLoS One*, *10*(8), Article e0136092.

Crossley, S., Dascalu, M., & McNamara, D. (2017). How important is size? An investigation of corpus size and meaning in both latent semantic analysis and latent Dirichlet allocation. In *The thirtieth international flairs conference*.

Dahal, B., Kumar, S. A., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, *9*(1), 1–20.

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(1), 2063–2067.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A comparison of pre-processing techniques for twitter sentiment analysis. In *International conference on theory and practice of digital libraries* (pp. 394–406). Springer.

El Barachi, M., AlKhatib, M., Mathew, S., & Oroumchian, F. (2021). A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *Journal of Cleaner Production*, Article 127820.

Fownes, J. R., Yu, C., & Margolin, D. B. (2018). Twitter and climate change. *Sociology Compass*, *12*(6), Article e12587.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 1–41.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12), 2009.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568–578.

Hahnel, U. J., Mumenthaler, C., & Brosch, T. (2019). Emotional foundations of the public climate change divide. *Climatic Change*, 1–11.

Holmberg, K., & Hellsten, I. (2015). Gender differences in the climate change communication on Twitter. *Internet Research*.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. (in press).

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media, vol. 8, no. 1*.

Jorgenson, A. K., Fiske, S., Hubacek, K., Li, J., McGovern, T., Rick, T., et al. (2019). Social science perspectives on drivers of and responses to global climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *10*(1), Article e554.

Karasakalidis, A., Effrosynidis, D., & Arampatzis, A. (2021). DUTH at SemEval-2021 task 7: Is conventional machine learning for humorous and offensive tasks enough in 2021? In *Proceedings of the 15th international workshop on semantic evaluation* (pp. 1125–1129).

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. In *Australasian joint conference on artificial intelligence* (pp. 488–499). Springer.

Kirilenko, A. P., Molodtsova, T., & Stepchenkova, S. O. (2015). People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Global Environmental Change*, *30*, 92–100.

Kirilenko, A. P., & Stepchenkova, S. O. (2014). Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, *26*, 171–182.

Koenecke, A., & Feliu-Fabà, J. (2019). Learning Twitter user sentiments on climate change with limited labeled data. arXiv preprint arXiv:1904.07342.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., et al. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), Article e1500779.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, *36*(2), 1–30.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 165–174).

Littman, J., & Wrubel, L. (2019). *Climate change tweets Ids*. Harvard Dataverse.

Loria, S. (2018). Textblob documentation. Release 0.15, 2.

Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in Adam. CoRR, abs/1711.05101.

Loureiro, M. L., & Alló, M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the UK and Spain. *Energy Policy*, *143*, Article 111490.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*.

Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P. R., et al. (2018). *1, An IPCC Special Report on the impacts of global warming of 1.5° C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (p. 32). Geneva, Switzerland: World Meteorological Organization.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, 299–313.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2), 103–134.

Palani, S., Rajagopal, P., & Pancholi, S. (2021). T-BERT–Model for sentiment analysis of micro-blogs integrating topic model and BERT. arXiv preprint arXiv:2106.01097.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Philo, G., & Happer, C. (2013). *Communicating climate change and energy security: new methods in understanding audiences*. Routledge.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Twenty-fourth international joint conference on artificial intelligence*.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer.

Rohde, R. A., & Hausfather, Z. (2020). The berkeley earth land/ocean temperature record. *Earth System Science Data*, *12*(4), 3469–3479.

Samantray, A., & Pin, P. (2019). Credibility of climate change denial in social media. *Palgrave Communications*, *5*(1), 1–8.

Shukla, P., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Pörtner, H., Roberts, D., et al. (2019). *IPCC, 2019: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. Intergovernmental Panel on Climate Change (IPCC).

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Sisco, M. R., Bosetti, V., & Weber, E. U. (2017). When do extreme weather events generate attention to climate change? *Climatic Change*, *143*(1), 227–241.

Sit, M. A., Koylu, C., & Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane irma. *International Journal of Digital Earth*.

Stojanovski, D., Strezoski, G., Madjarov, G., & Dimitrovski, I. (2015). Twitter sentiment analysis using deep convolutional neural network. In *Hybrid artificial intelligent systems*.

Sugg, J. W. (2021). Exploratory geovisualization of the character and distribution of American climate change beliefs. *Weather, Climate, and Society*, *13*(1), 67–82.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, 298–310.

Tekumalla, R., & Banda, J. M. (2021). Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Computing and Applications*, 1–9.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Williams, H. T., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138.

Yeo, S. K., Handlos, Z., Karambelas, A., Su, L. Y.-F., Rose, K. M., Brossard, D., et al. (2017). The influence of temperature on# ClimateChange and# GlobalWarming discourses on Twitter. *Journal of Science Communication*, *16*(5), A01.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 233–242).

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, *9*(2), 1–29.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., et al. (2016). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2105–2114).

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*(2), 379–398.