Climate Change Discourse Analysis on Reddit: Topics, Emotions and the Transformation of Perspectives

Luis Britz, Anja Huber, Felix Krause & Yvonne-Nadine Preda

universität wien

#ClimateChange #EmotionDetection
#Reddit #TopicDetection
#NLP #BERTopic

Scan me for more information, feedback, comments or code review :)

Goals & Research Questions

- General goal: Make large raw text data interpretable for (social) sciences
- Which topics are discussed on Reddit regarding climate change?
- How do emotions of the discourse develop over time?
- How do users' perspectives on climate change shift over time?

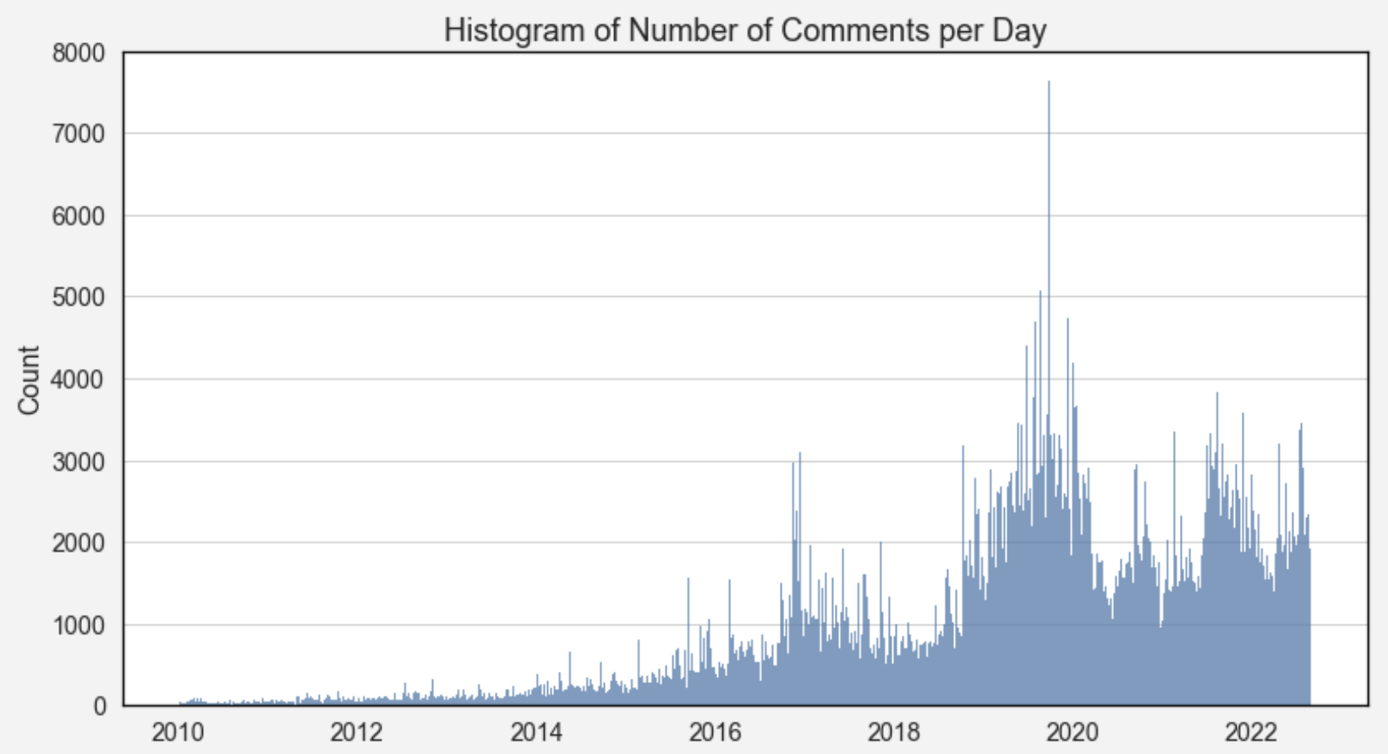
Dataset

- **Reddit posts and comments** with terms "climate" and "change" from 01/2010 to 08/2022
- **4,6m comments** with 10 features
- **620k posts** with 12 features
- Features contain IDs, subreddit name, timestamp, text and sentiment score

Data Understanding

Most Frequent Subreddits

politics	370.018
worldnews	351.195
askreddit	259.848
collapse	94.696
news	94.558
futurology	89.945
science	71.453
environment	70.444
canada	66.813
australia	60.239
conspiracy	50.951
unpopularopinion	49.178
climateskeptics	46.524
ukpolitics	43.179
changemyview	42.902
neoliberal	42.268
pics	42.233
europe	37.331
the_donald	34.106
canadapolitics	31.399


Histogram of Number of Comments per Day

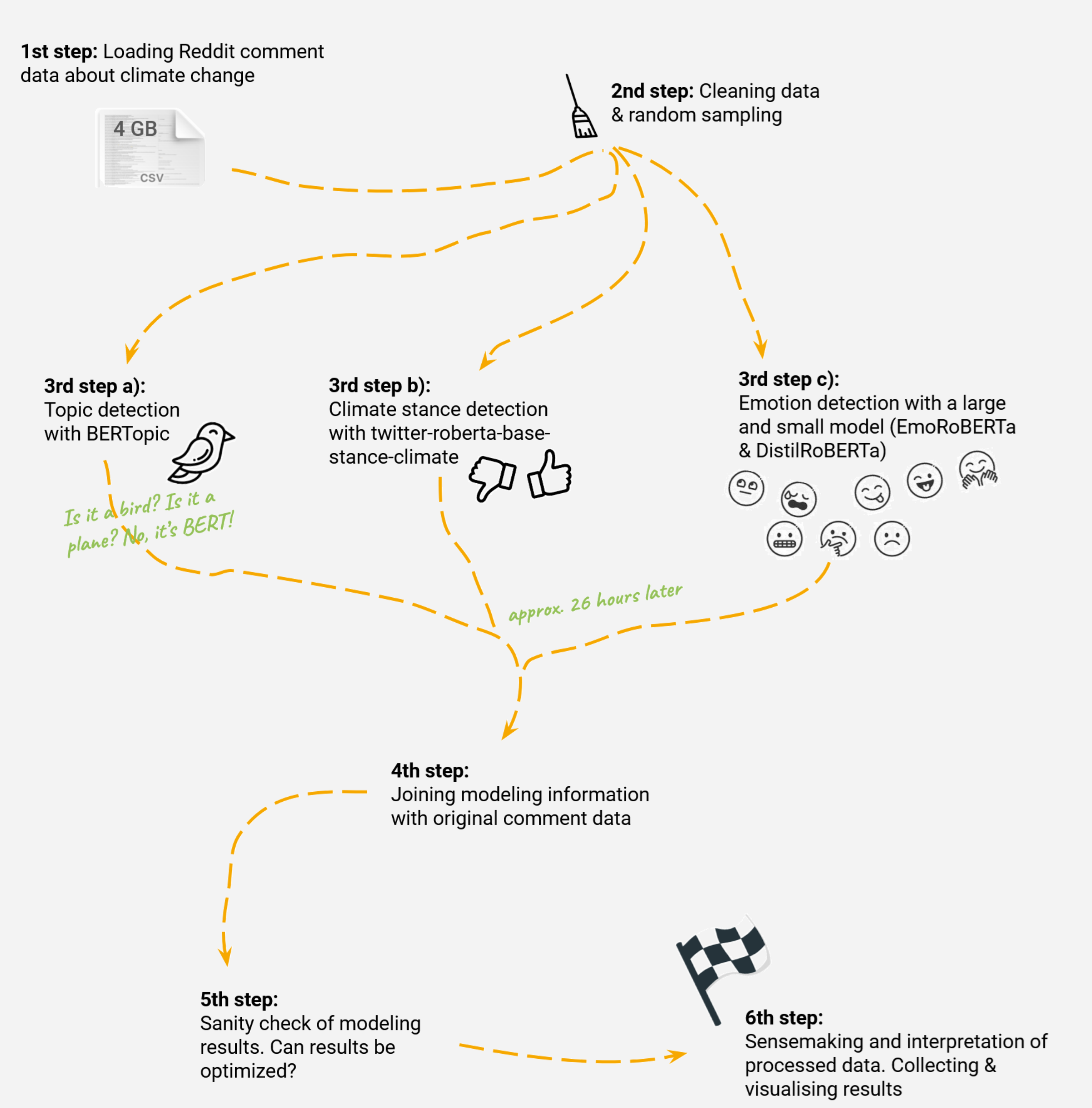
Subreddits with "bot" in Name (and # of Comments)

Bottown2	27.435	explainbothsides	208
bikinibottomtwitter	1.603	bottown22	166
subredditsummarybot	368	u_anticensor_bot	138
newsbotbot	274	mrrobot	122
botany	212	bottown	119

Data Preparation

- 85% of posts lacked any text → **drop all posts**
- **Timestamp conversion**
- **NA and duplicate removal**
- **Type conversion** into string values
- Large data set → **random sample** for each year (max 100k)
- Large **bot-subreddits** → removed via name
- **Text feature removal** (@mentions, hyperlinks, numeric symbols and HTML tags such as "&lt", "&gt", "&le", "&ge")
- **Language tagging** → removing non-english posts (~3%)
- No standard pre-processing steps like lemmatization or word stemming

Workflow



1st step: Loading Reddit comment data about climate change

2nd step: Cleaning data & random sampling

3rd step a): Topic detection with BERTopic

Is it a bird? Is it a plane? No, it's BERT!

3rd step b): Climate stance detection with twitter-roberta-base-stance-climate

3rd step c): Emotion detection with a large and small model (EmoRoBERTa & DistilRoBERTa)

approx. 26 hours later

4th step: Joining modeling information with original comment data

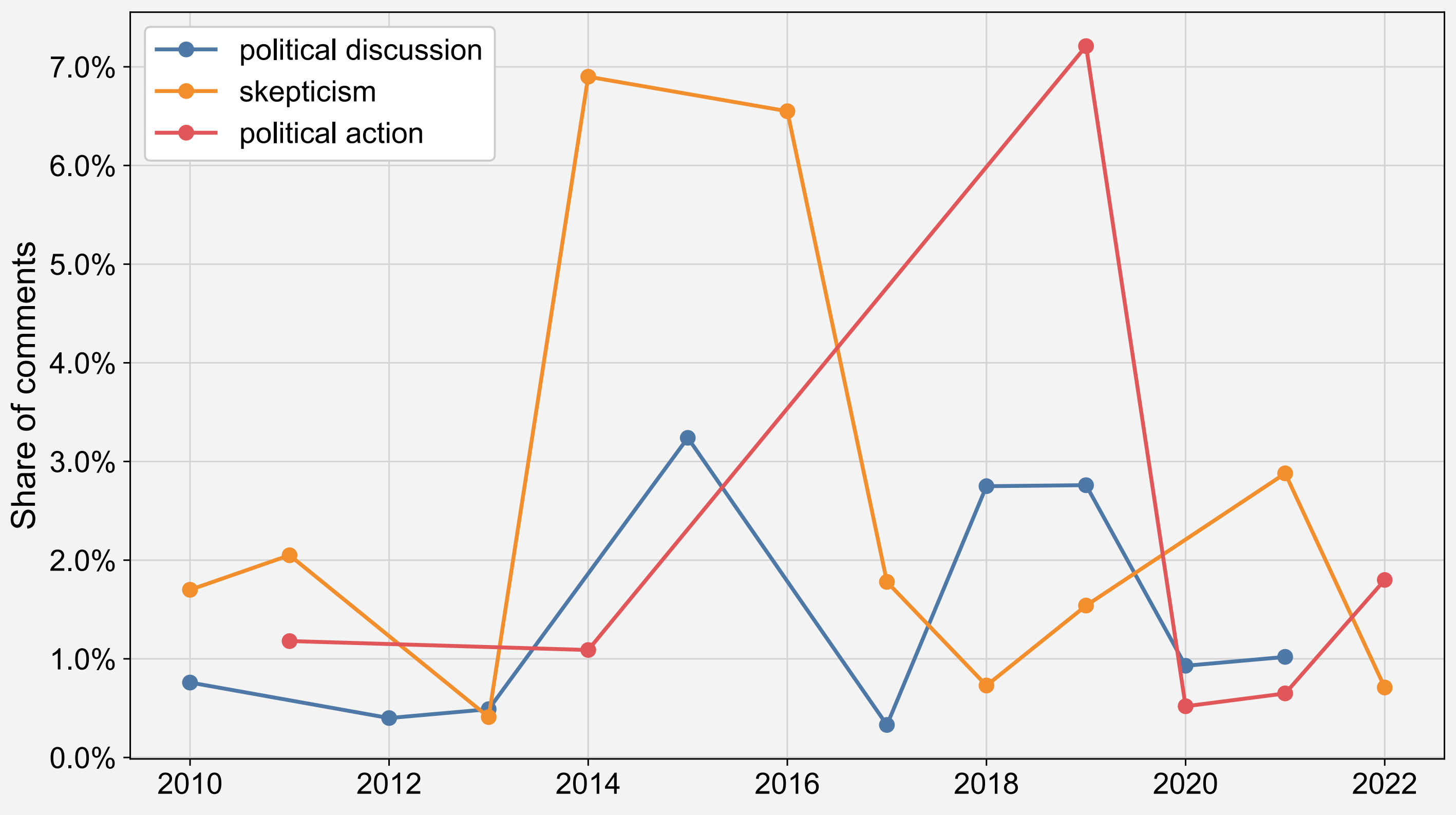
5th step: Sanity check of modeling results. Can results be optimized?

6th step: Sensemaking and interpretation of processed data. Collecting & visualising results

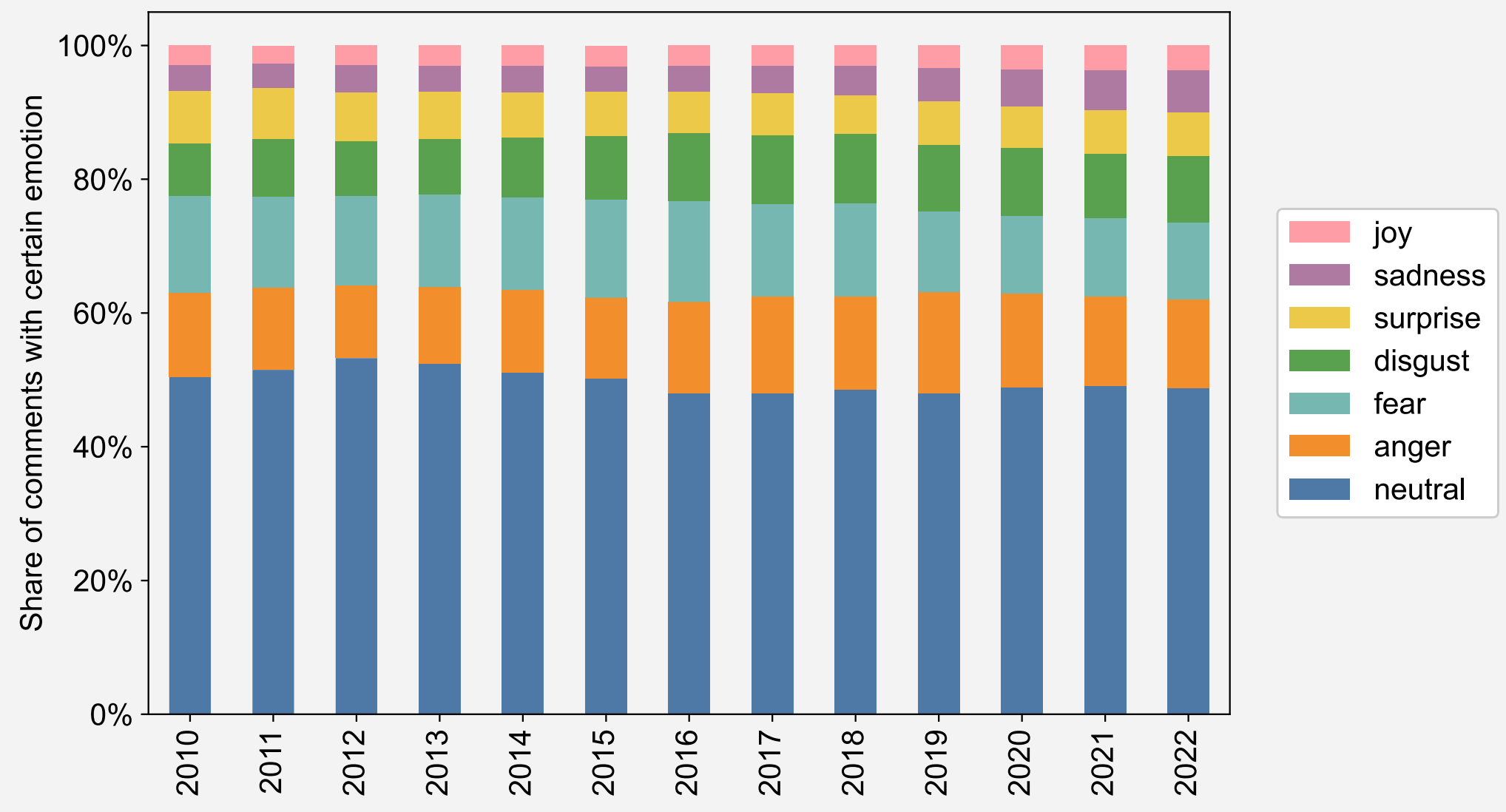
Findings & Results

- **Climate stance model** did not work well, "against" category was never classified
- **Emotions** do not fluctuate much over the years
- **Topics** are strongly influenced by large-scale events (e.g., elections, bushfires, mass shootings, COVID-19 pandemic) and are often specific to western world or in particular US
- **High-level topic categories**: Top 5 topics per year manually grouped in: "general", "individual responsibility", "political action", "political discussion", "scientific discussion", "skepticism"
- **Political discussions** show large fluctuations, probably caused by current political events
- **Political action discussions** drastically declined due to pandemic (e.g., no demonstrations possible, headlines dominated by COVID-19 → topic repressed from collective perception)
- **Climate change skepticism discussions**
  - 2014 to 2016 a lot of climate change scepticism happening (in line with results of own validation research on history of climate change deception)
  - Up to 2018 sharp decline of scepticism
  - 2020 slight increase driven by the pandemic

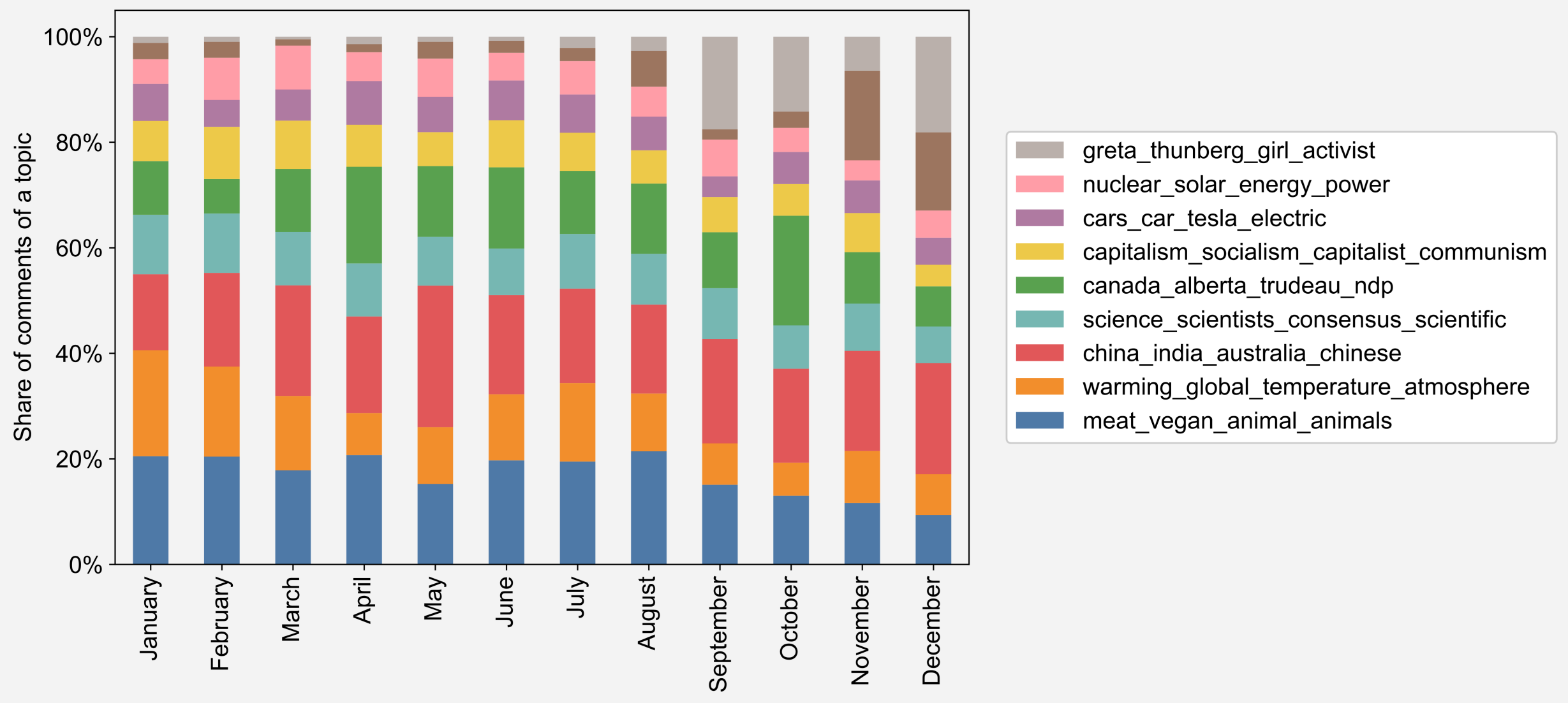
Share of Comments Grouped to High-Level Categories Over Time



Emotion of Comments over the Years



Topic Frequencies over the Months in 2019



About BERTopic

- Topic modeling technique using Google's BERT model for document embeddings
- Utilizes pre-trained BERT models for high-quality embeddings
- Applies UMAP for dimensionality reduction
- Uses HDBSCAN for clustering similar documents
- Enables intuitive exploration and visualization of clusters
- Extracts representative keywords for each topic
- Supports incremental learning for adding new documents
- Achieves state-of-the-art performance in NLP tasks

Learnings

- **Bias in data**: Need to keep in mind who is actively engaging in discussions on Reddit → bias in opinions created by subgroup of people using this platform actively (average user is male, less than 40 years old and from the US)
- **Little standard NLP pre-processing necessary**: State of the art models do it (e.g., stopwords removal) during the modeling process or even need some of the information for modeling (e.g., see BERTopic documentation)
- **Difficult interpretation**: Detected topics and emotions are difficult to interpret and require further manual, qualitative work. No "ready-to-use" results after modeling
- **Difficult results presentation**: Tough to present relevant findings out of final aggregated results (high dimensionality → years & topics)
- **State-of-the-art models need A LOT of processing power**
- **Group work can actually be fun** 🤓