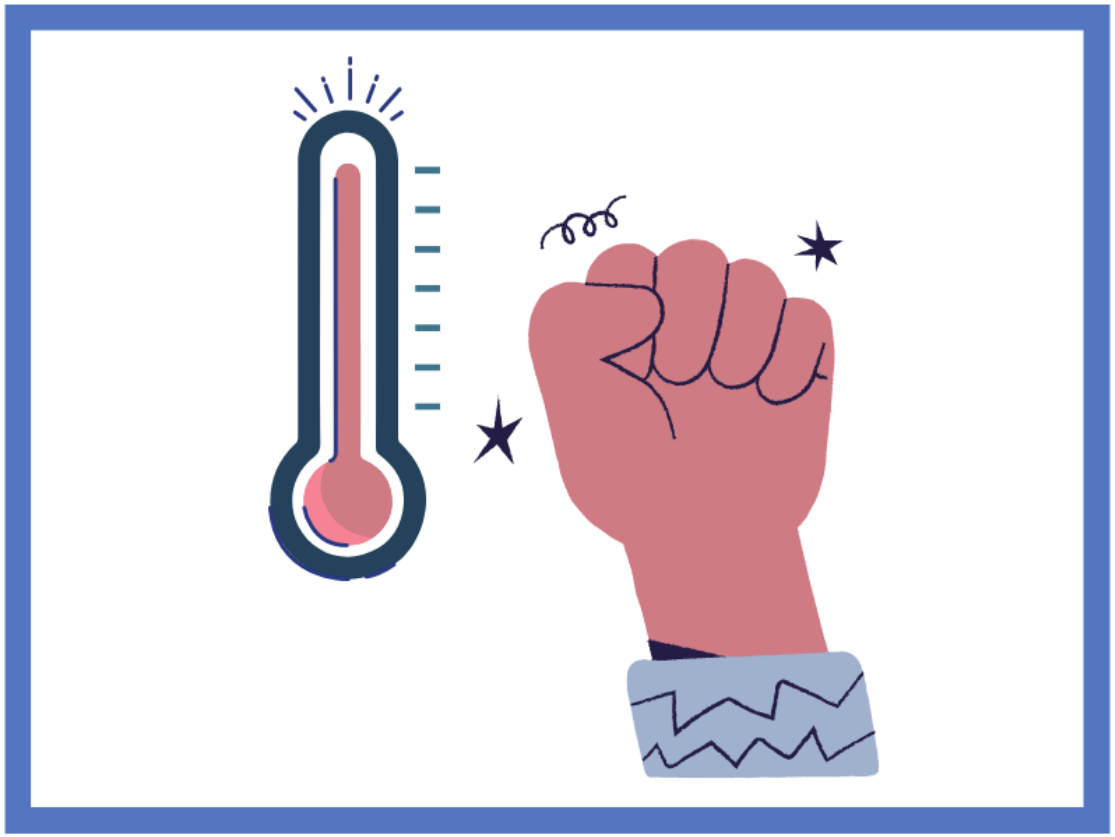
Potsdam Institute for Climate Impact Research
PIK

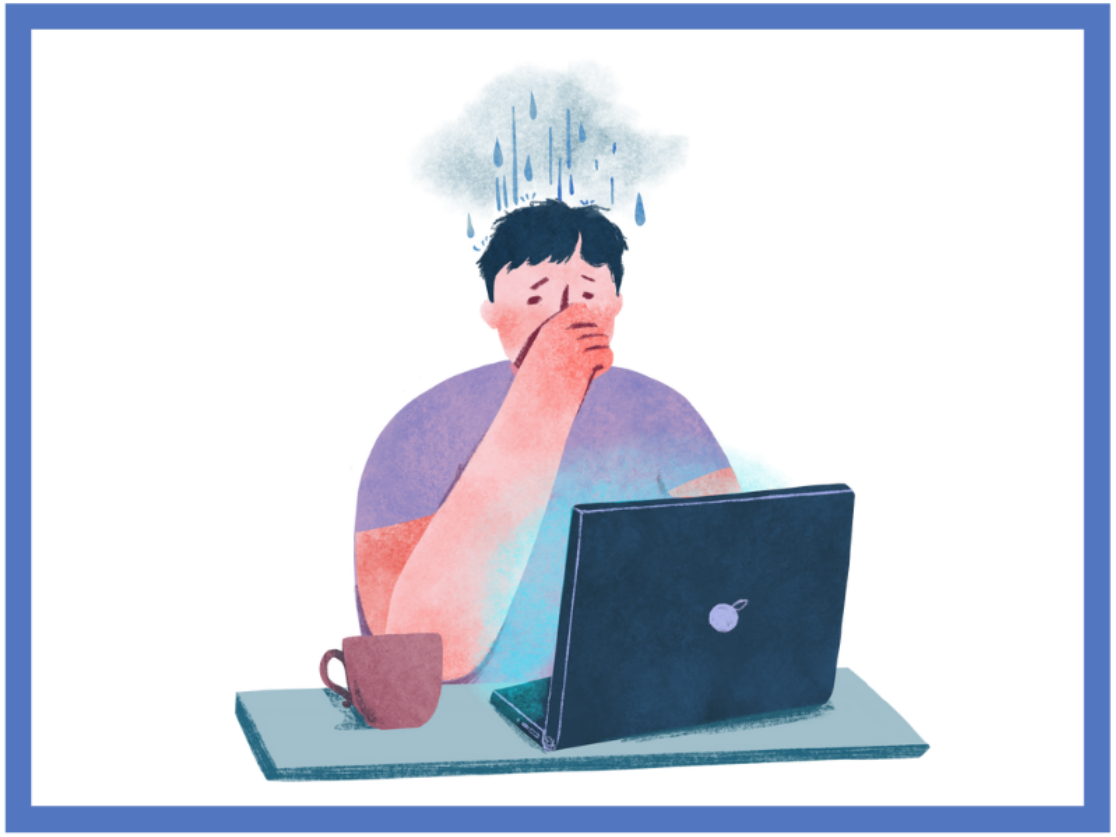
Temperature impacts on hate speech online: evidence from four billion tweets

Annika Stechemesser, Anders Levermann, Leonie Wenz

Motivation

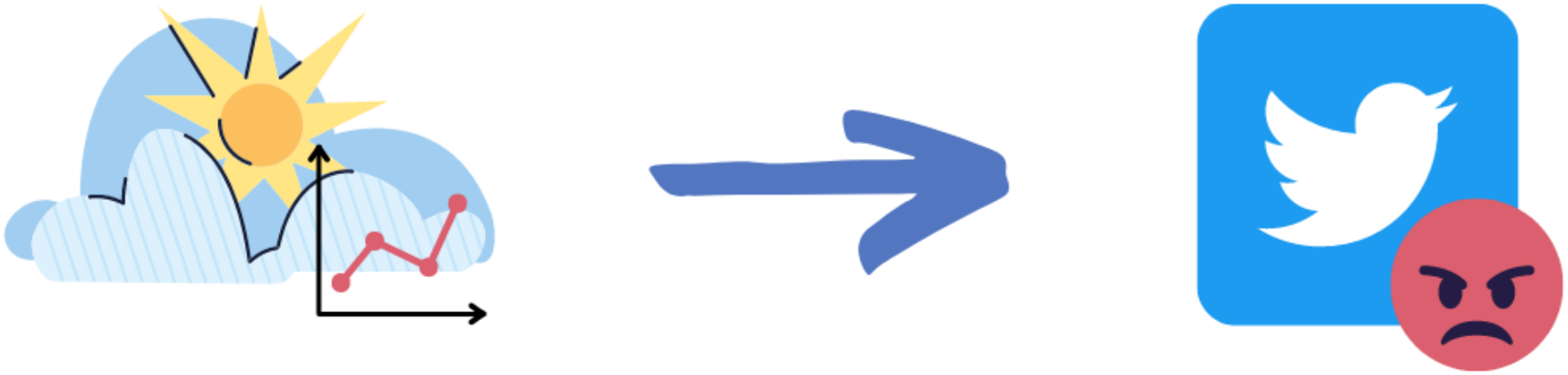


Ambient temperature has been identified as a potential cause for human conflict in a variety of studies.



Rapid digitalization of everyday life has led to a high frequency of interpersonal conflicts in the online world which aggravate mental health problems.

What is the impact of temperature on hate speech on the social media platform Twitter in the United States?



- 4 billion geolocated tweets from over 750 US cities (2014-2020)
- Supervised machine-learning approach to identify hate speech in tweets
- Binned fixed-effects panel regression model

Assembling the training data

Dataset	Labels and Counts	Total
HAR (Golbeck et al)	Harassment 5.285 No harassment 15.075	20.360
HATE (Davidson et al)	Hate 1430 Offensive language 19190 Neither 4163	24.783
Waseem and Hovy*	Racism 13 Sexism 1914 None 4183	6110

*Accessed in July 2020

Preprocessing of training and testing data

Remove characters not in the Latin Alphabet and stopwords

Preprocessing of training and testing data

Remove characters not in the Latin Alphabet and stopwords

Apply lemmatization and stemming

Preprocessing of training and testing data

Remove characters not in the Latin Alphabet and stopwords

Apply lemmatization and stemming

Split into training and testing data (75%/25%, balanced labels)

Preprocessing of training and testing data

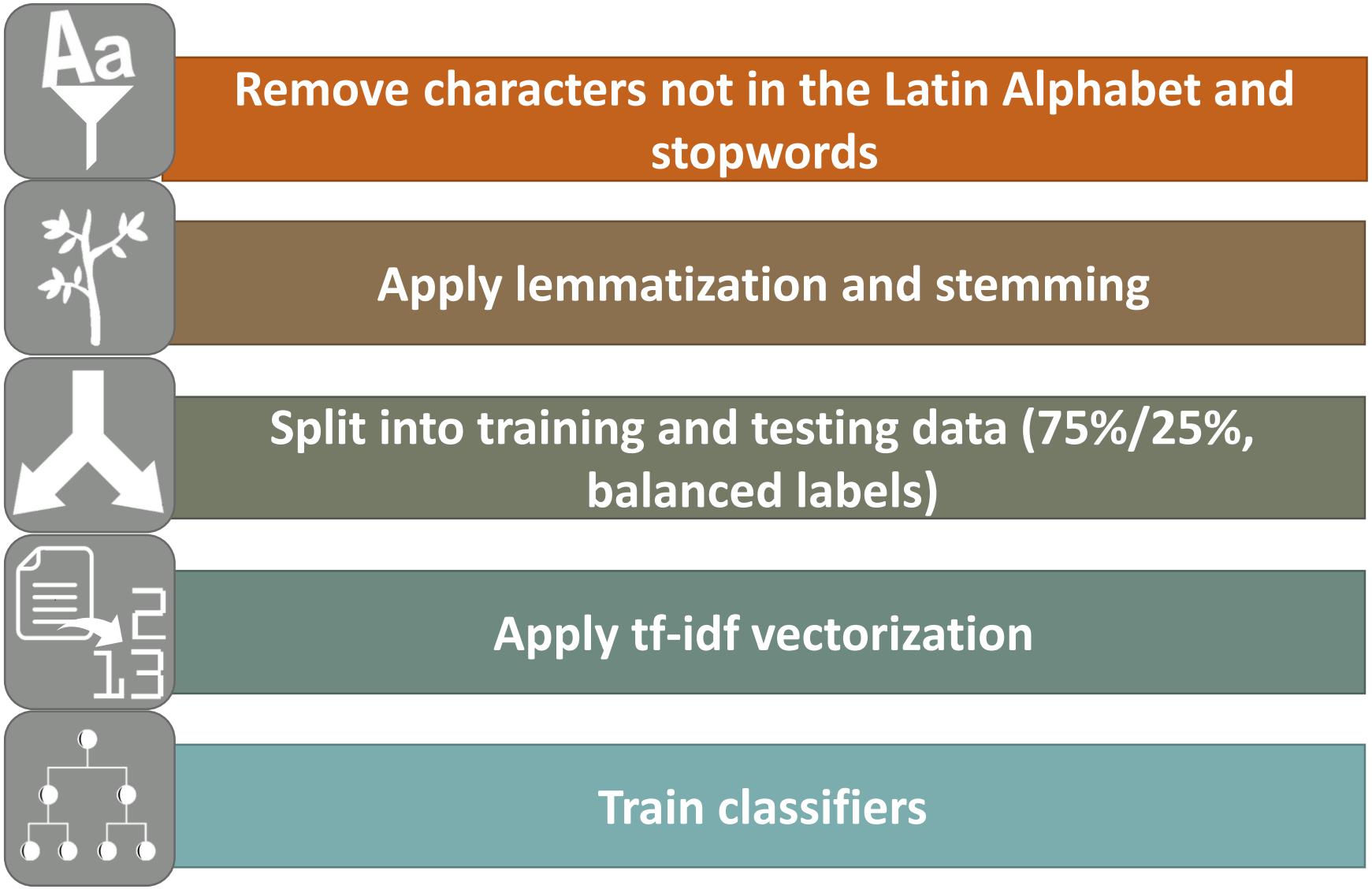
Remove characters not in the Latin Alphabet and stopwords

Apply lemmatization and stemming

Split into training and testing data (75%/25%, balanced labels)

Apply tf-idf vectorization

Preprocessing of training and testing data



Classifier	F1	Accuracy	Precision	Recall
Logistic Regression	Hate 0.85 No hate 0.86	0.85	0.9	0.83
Naïve Bayes	Hate 0.76 No hate 0.82	0.79	0.78	0.86
Random Forest	Hate 0.82 No hate 0.84	0.83	0.88	0.84
Stochastic Gradient Descent	Hate 0.83 No hate 0.83	0.83	0.90	0.78
Gradient Boosting Machine	Hate 0.86 No hate 0.84	0.85	0.96	0.75

Classifier	F1	Accuracy	Precision	Recall
Logistic Regression	Hate 0.85 No hate	0.85	0.9	0.83
Naïve Bayes			0.78	0.86
Random Forest			0.88	0.84
Stochastic Gradient Descent	Hate 0.83 No hate 0.83	0.83	0.90	0.78
Gradient Boosting Machine	Hate 0.86 No hate 0.84	0.85	0.96	0.75

Precision: What proportion of positive identifications was actually correct?

$$\frac{True\ Positives}{True\ Positives + False\ Positives}$$

Classifier	F1	Accuracy	Precision	Recall
Logistic Regression	Hate 0.85 No hate 0.86	0.85	0.9	0.83
Naïve Bayes	Hate 0.76 No hate 0.82	0.79	0.78	0.86
Random Forest	Hate 0.82 No hate 0.84	0.83	0.88	0.84
Stochastic Gradient Descent	Hate 0.83 No hate 0.83	0.83	0.90	0.78
Gradient Boosting Machine	Hate 0.86 No hate 0.84	0.85	0.96	0.75

Identification of hate tweets

75 Million

Labelled hate / non-hate tweets	
Hate?	Tweets
Yes	Tweet 1
No	Tweet 2
No	Tweet 3
Yes	Tweet 4

Testing data	
Hate?	Tweets
No	Tweet 2
Yes	Tweet 4

Training data	
Hate?	Tweets
Yes	Tweet 1
No	Tweet 3

Algorithm
Gradient Boosting Machine

Evaluation
Precision

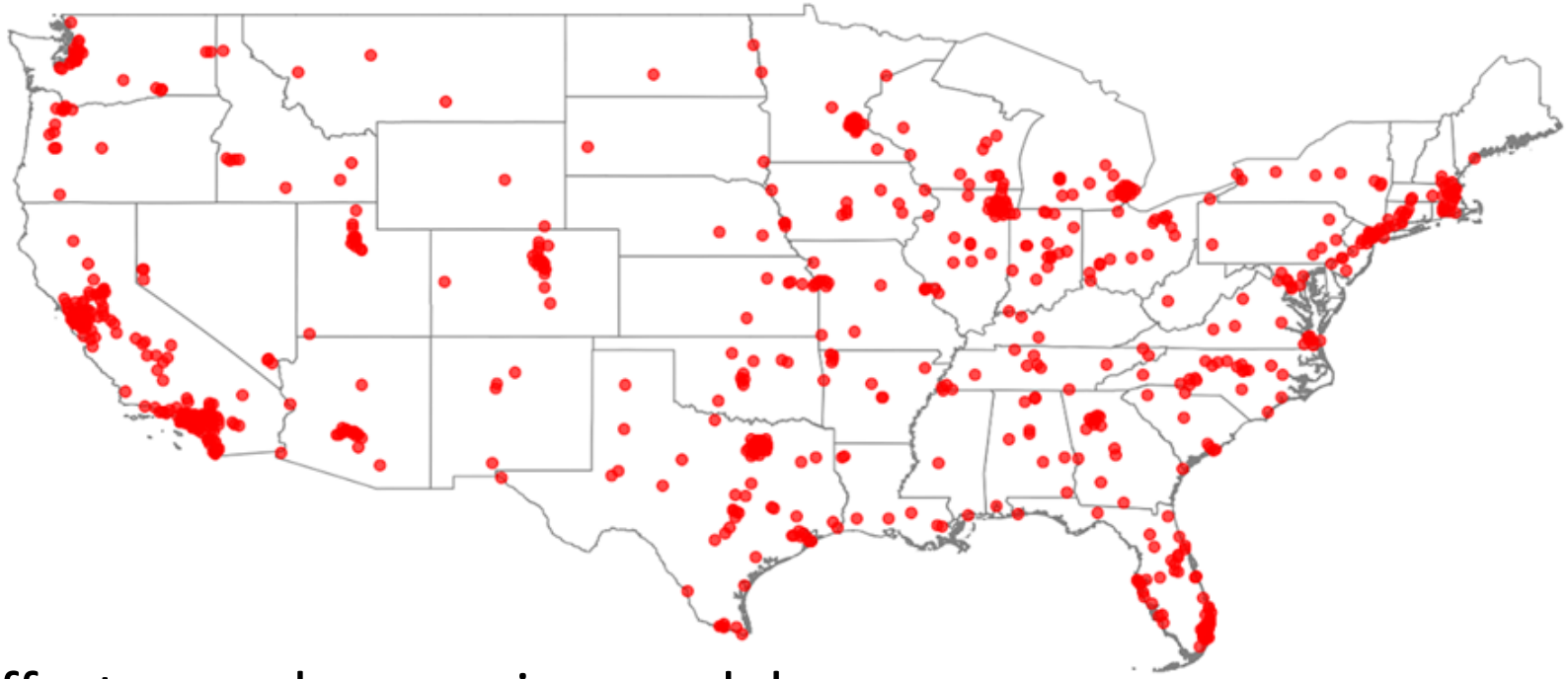
Classifier

Predictions

US1420 Data with Labels	
Hate?	US1420 Tweets
No	Tweet 1
...	...
Yes	Tweet 3.8 billion

New US1420 Tweets
Tweet 1
...
Tweet 3.8 Billion

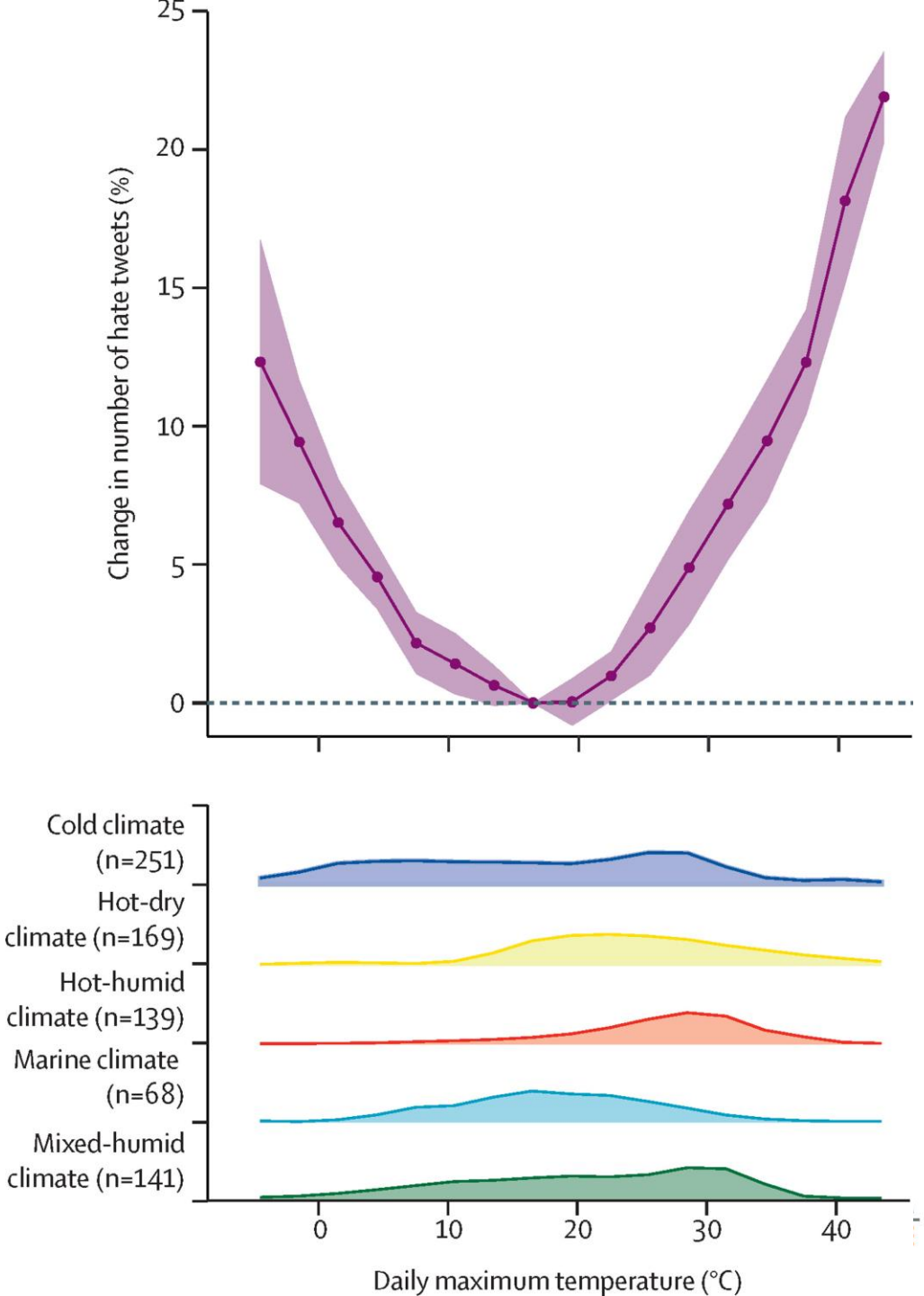
Aggregation and statistical analysis



- Binned fixed-effects panel regression model
- Dependent variable: log(daily hate tweets)
- Independent variable: Temperature in 3°C bins
- Controls for precipitation, cloud cover, windspeed, holidays, weekdays
- City:year fixed effects

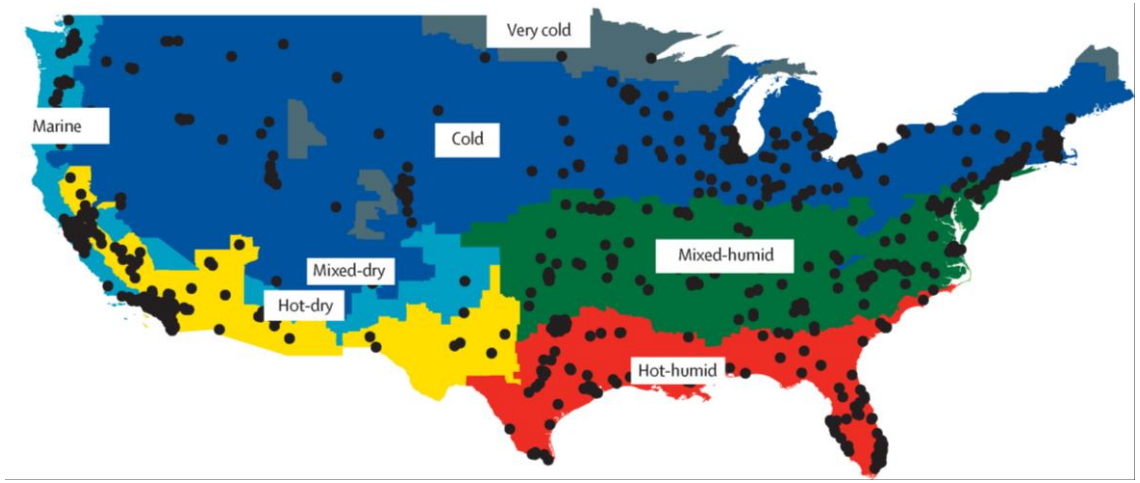
Strong nonlinear response of hate tweets to daily ambient temperature

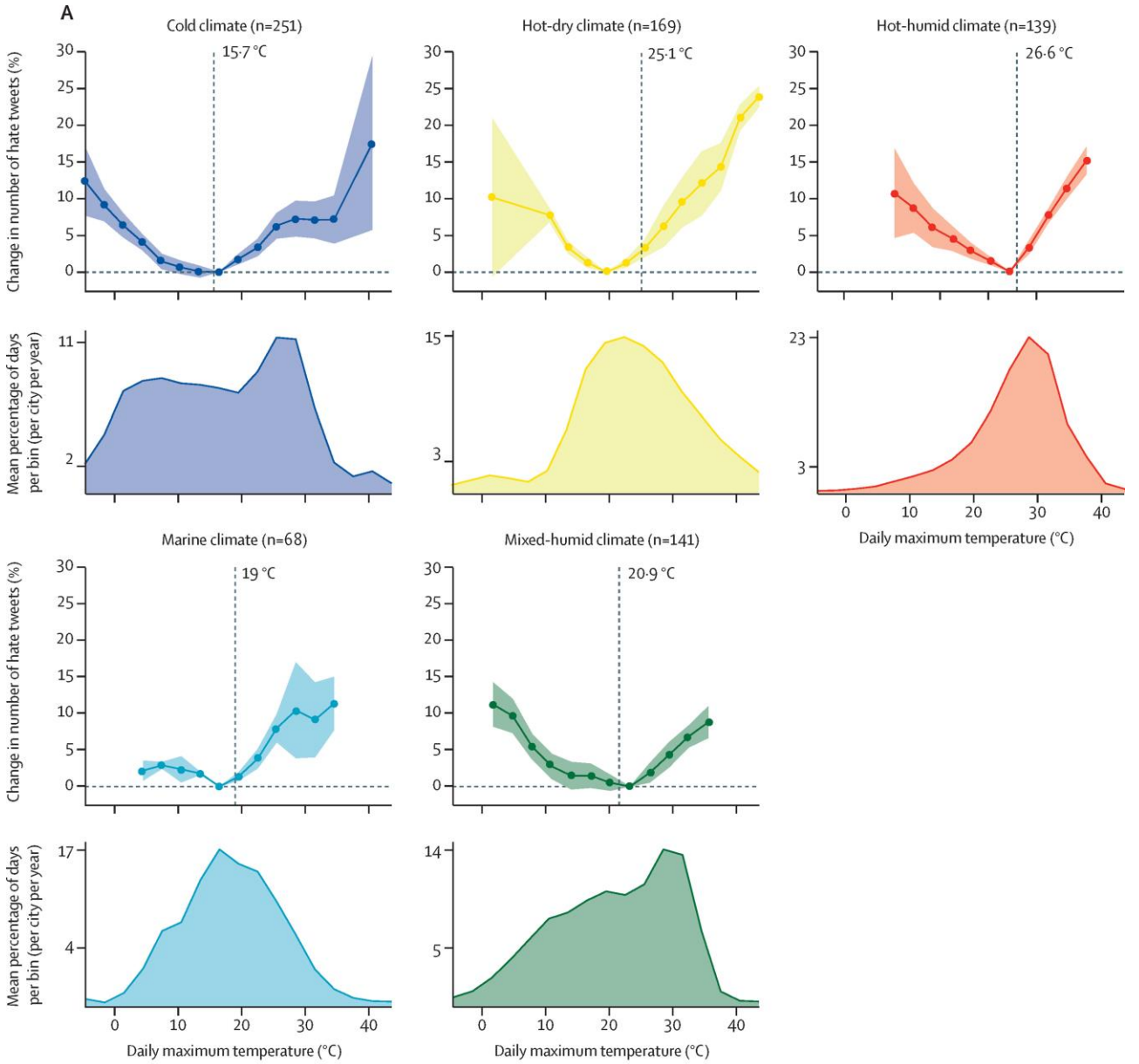
- fewest hate tweets between (15°C and 18°C)
- Sharp increases for temperatures warmer than 27°C and colder than 6°C
- On average, a city crosses 12.6 temperature bins in a year



Nonlinear response is consistent across climate zones

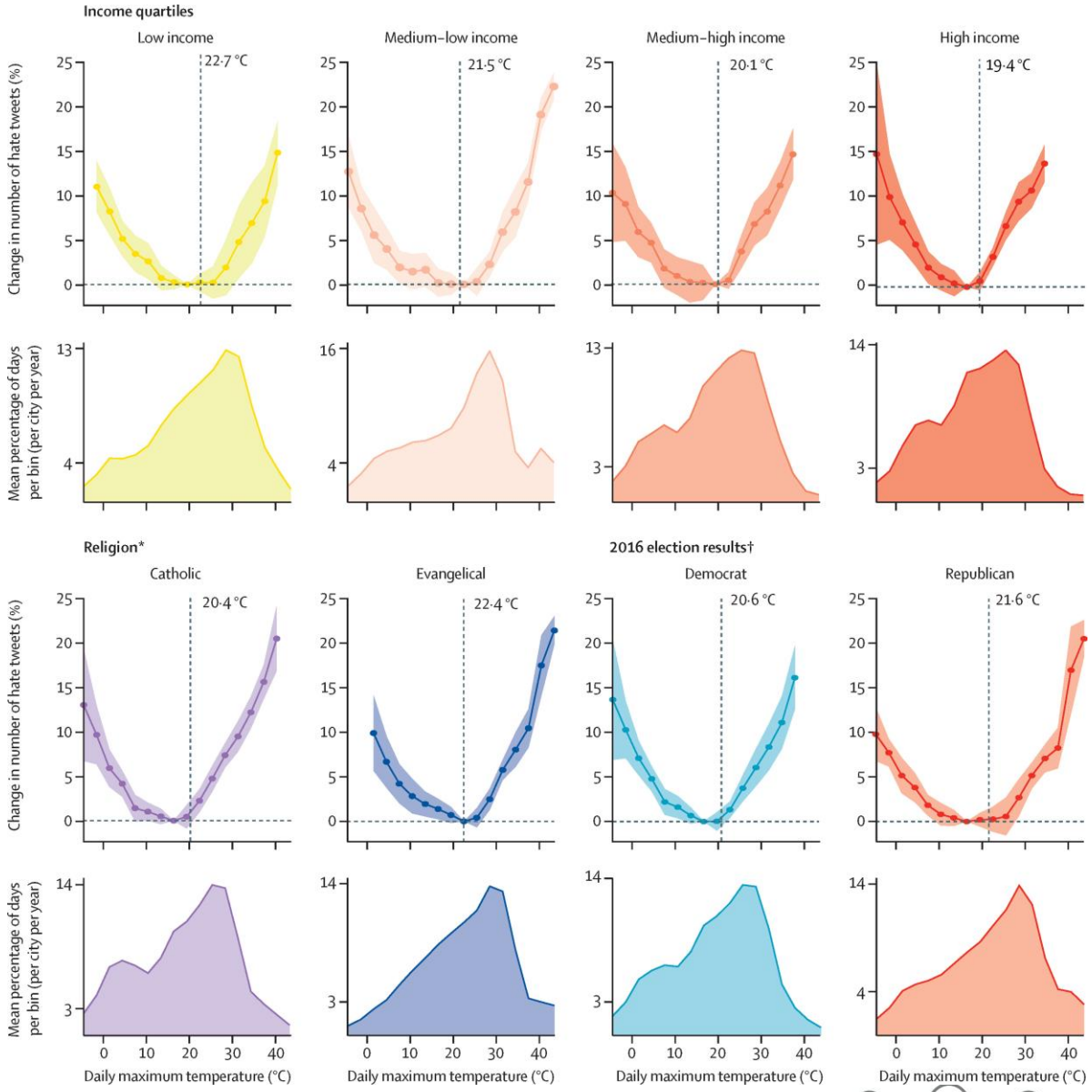
- Temperatures of more than 30°C (86°F) are linked with hate increases of at least 7% across all climate zones
- Evidence for limits to temperature adaptation





Increases for cold and hot temperatures across incomes and mindsets

- Evidence that aggressive tendencies are not a question of mindset but the temperature influence is more universal
- Even the highest income-quartile shows an increase in hate speech on hot days despite the ability to mitigate heat



Conclusions

- Extreme temperatures are associated with more hate on Twitter
- Evidence for limits to temperature adaptation
- Impacts of climate change on mental health

Thank you for your attention!

Full paper:

Stechemesser, Levermann and Wenz, Temperature impacts on hate speech online: evidence from 4 billion tweets. The Lancet Planetary Health, 09/22, doi: https://doi.org/10.1016/S2542-5196(22)00173-5

stechemesser@pik-potsdam.de