Climate Change Education Engine

Machine Learning for the Betterment of Society

Author: Nabil Abbas

“We are the first generation to be able to ***end poverty****, and the last generation that can take steps to avoid the worst impacts of* ***climate change****. Future generations will judge us harshly if we fail to uphold our* ***moral and historical*** *responsibilities.”*

-Ban Ki-Moon

Secretary-General United Nations

Agenda

- Goals
- Strategy
- Data Sources
- EDA
- Classifier
- Topic Modelling
- What’s next?
- Future Considerations

Goals

- Create a **pipeline** to recommend custom content tailored towards a twitter user dependent on their online activity regarding the topic of “Climate Change”

- The Recommender is advanced because the engine uses LDA to model the subtopics a user discusses when discussing climate change.

Strategy

Classifier

Trained on **subreddit community text data**

Get data using pushshift.io API to get 173,000 subreddit comment on every post from the last 5 years.

Subreddit Pages: Climate & Climate Skeptics

Topic Model

Create **15 topics** trained on subreddit corpus.

Modify model to create optimal number of topics for corpus.

Recommender Engine

Natural Language Processing to clean climate change tweets per user.

Fit Recommender using Collaborative Filtering method

Users: Twitter Users

Items: Climate Subtopics

Data Sources

Twitter

Using Twint API tool obtain Climate Change tweets from users on Twitter.

Reddit

Using pushshift.io API tool obtain comment data from subreddits:

- Climateskeptics
- Climate

EDA - Problem Risen and Remedies

Slow API calls

- Distributed Systems: API calls made amongst 4 units to speed up data retrieval

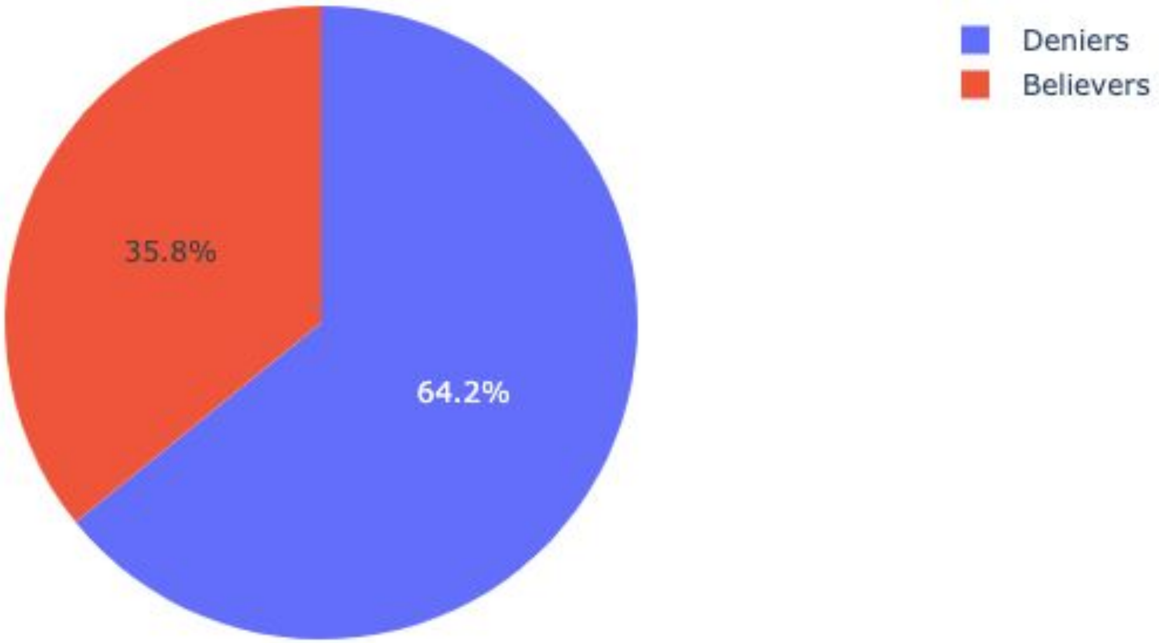
Reddit Data Class Imbalance

- 5 years indicate more comments for Deniers
- Undersampling: TomekLinks, NeighborhoodCleaningRule, ClusterCentroids, RandomUndersampler

Outside Noise

- Remove outside “noise” by removing negative scored comments from data set class

EDA - Comment Data Class Imbalance



Classification

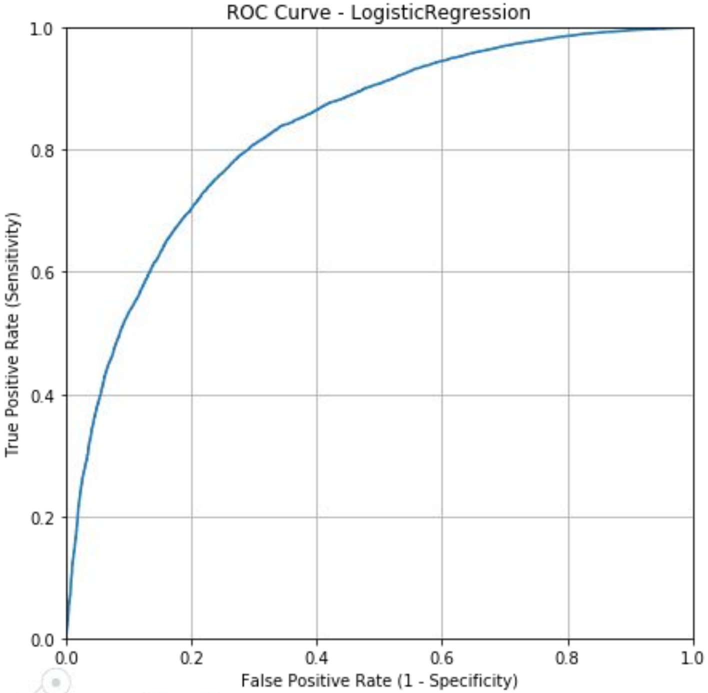
Dummy Metric	
F1	65 %
Recall	66 %
Precision	54 %
Accuracy	65 %

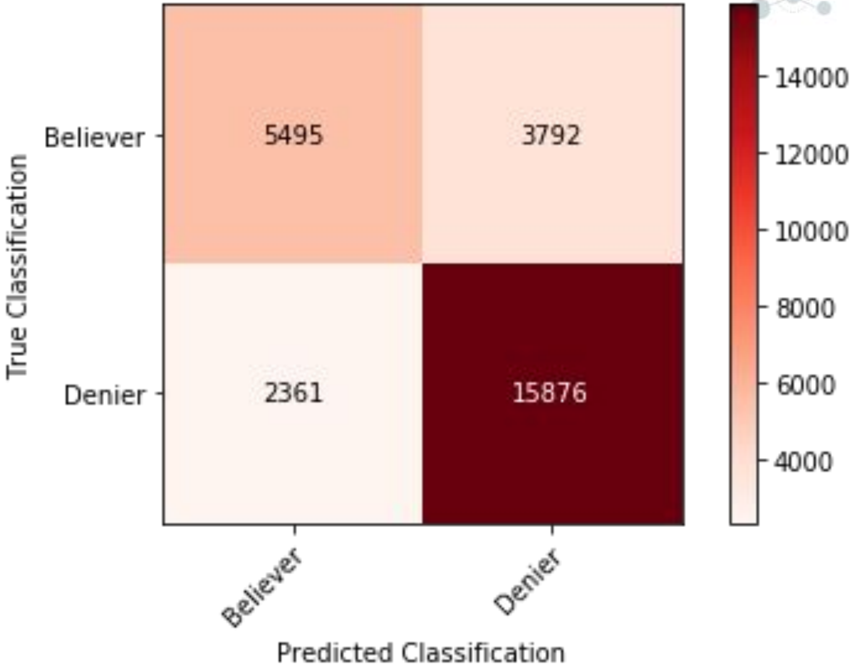
Best Classifier	**Logistic Regression**
F1	84 %
Recall	78 %
Precision	81 %
Accuracy	78 %

Overfit Concerns: Training Data	
F1	96 %
Recall	98 %
Precision	94 %
Accuracy	96 %

Classifiers			
Naive Baye’s	SVM	KNN	Decision Trees
Logistic Regression	XG Boost	AdaBoost	GradientBoost

Best Model: Logistic Regression

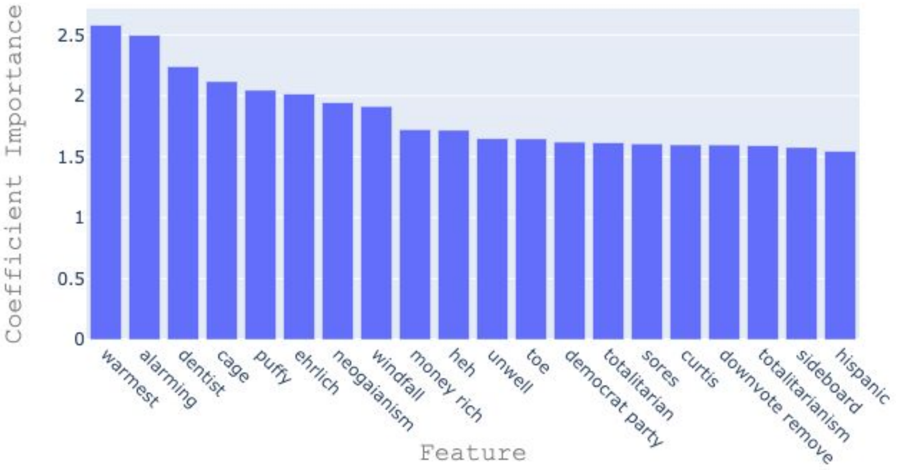




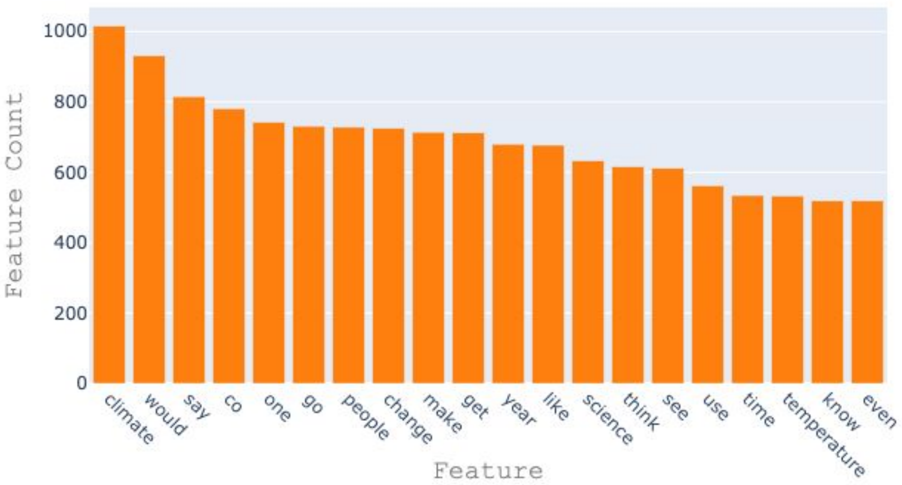
AUC: 0.832

Classification: Positive Class (Denier)

Positive Feature Importance

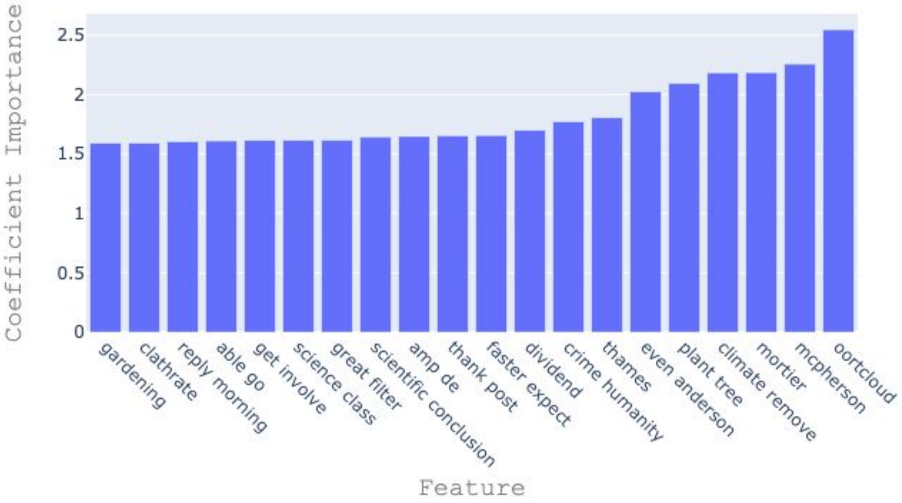


Positive Class Feature Count

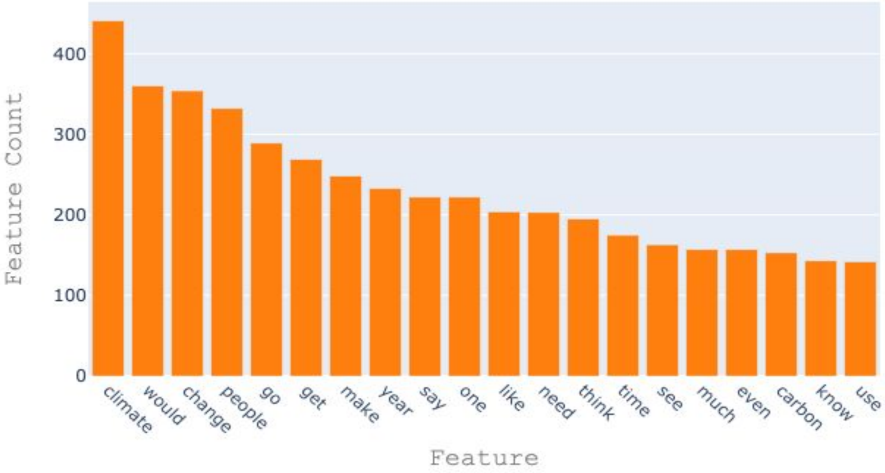


Classification: Negative Class (Believers)

Negative Feature Importance



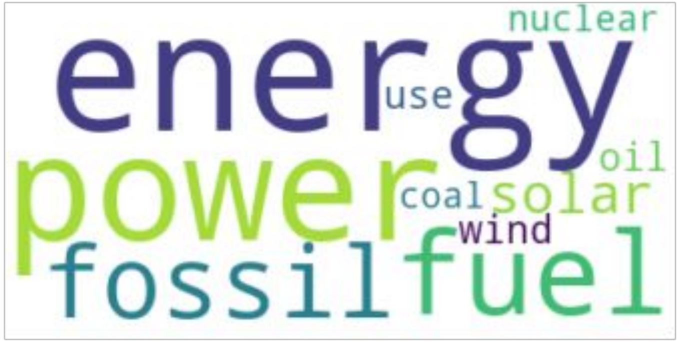
Negative Class Feature Count

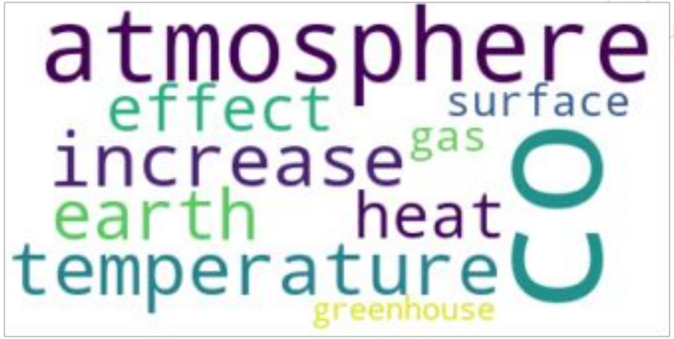


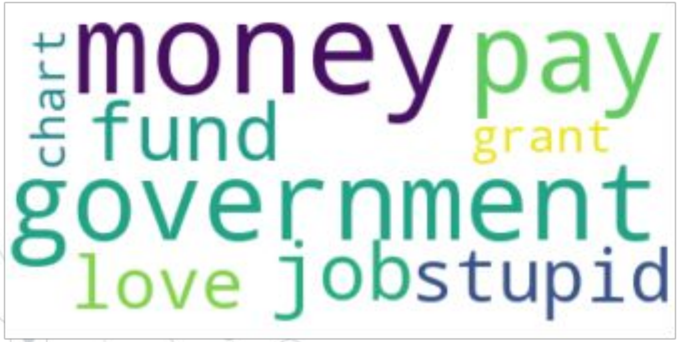
Topic Modeling

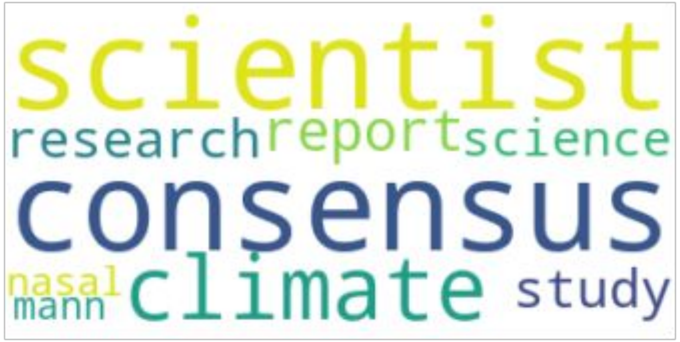
Topics Legend		
Climate Beliefs	Rising Temperatures	Thought Processes
Physics	Atmospheric Changes	Science
Global Warming	Government Involvement	Climate Reports
Energy	Water	Article / Link Discussion
Internet Conversations	Polar Ice	Legal

Topic Modeling - Key Topics





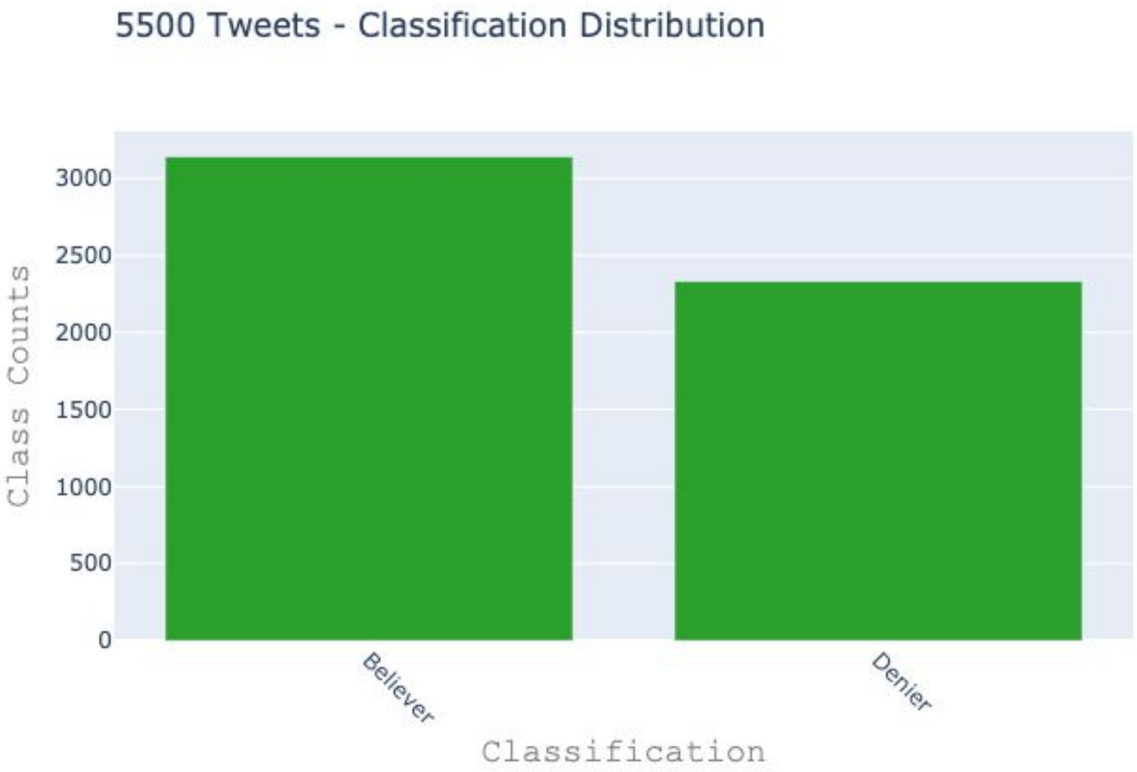


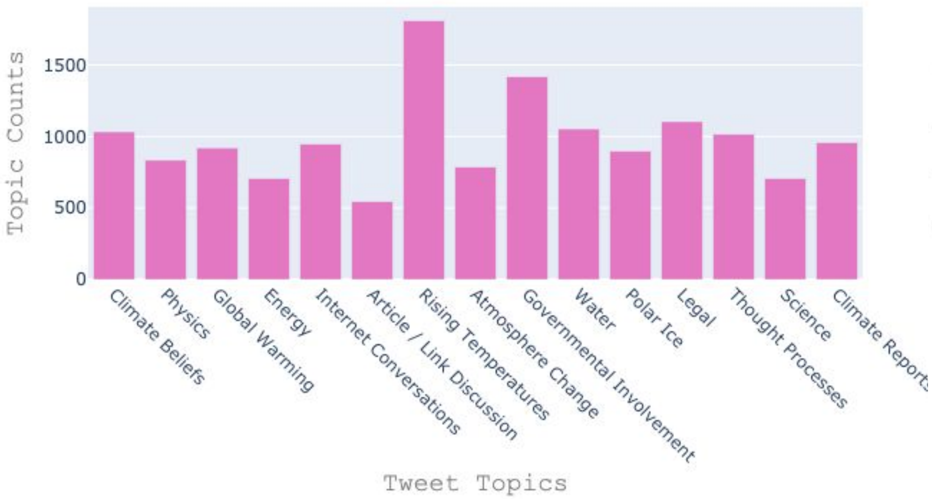


Preparing the Recommender

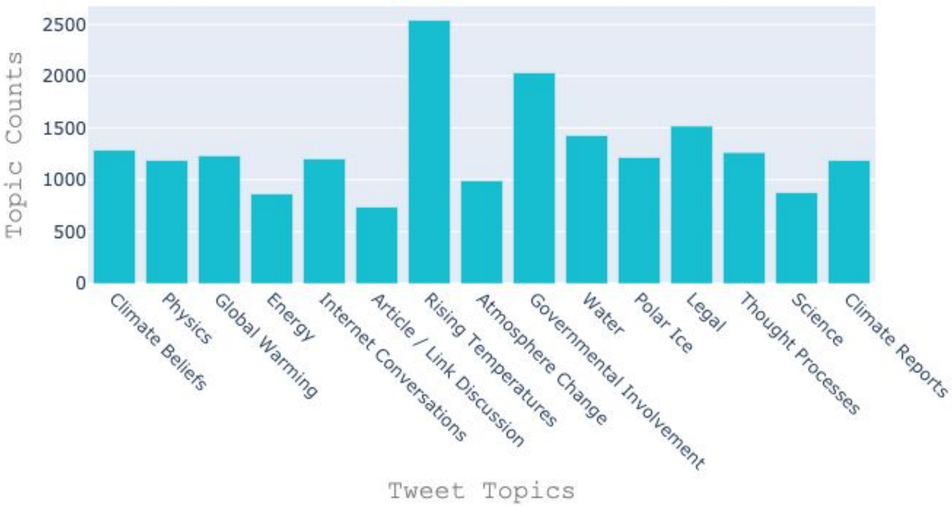
Twitter Data

Tweet Classification Distribution

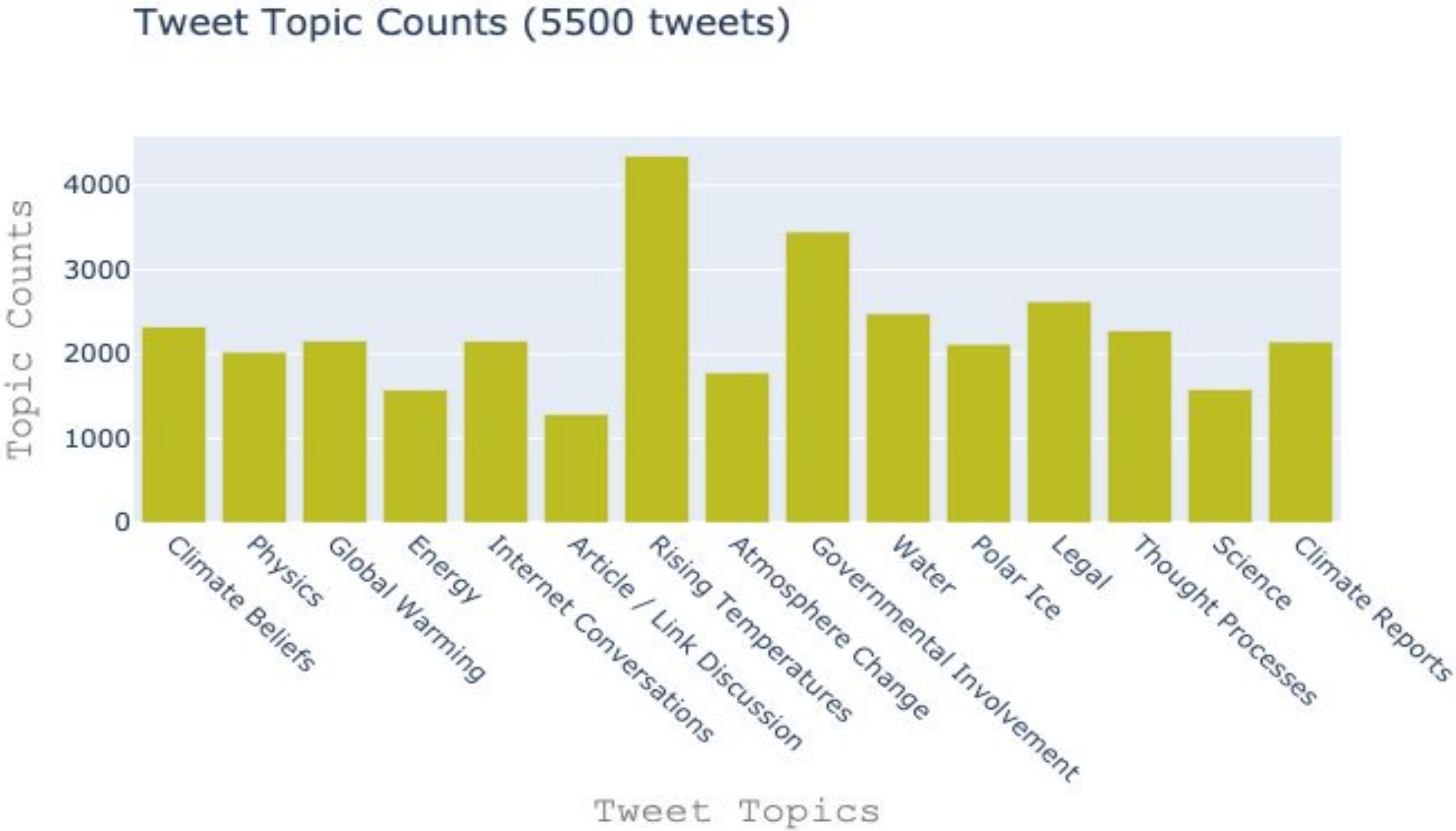




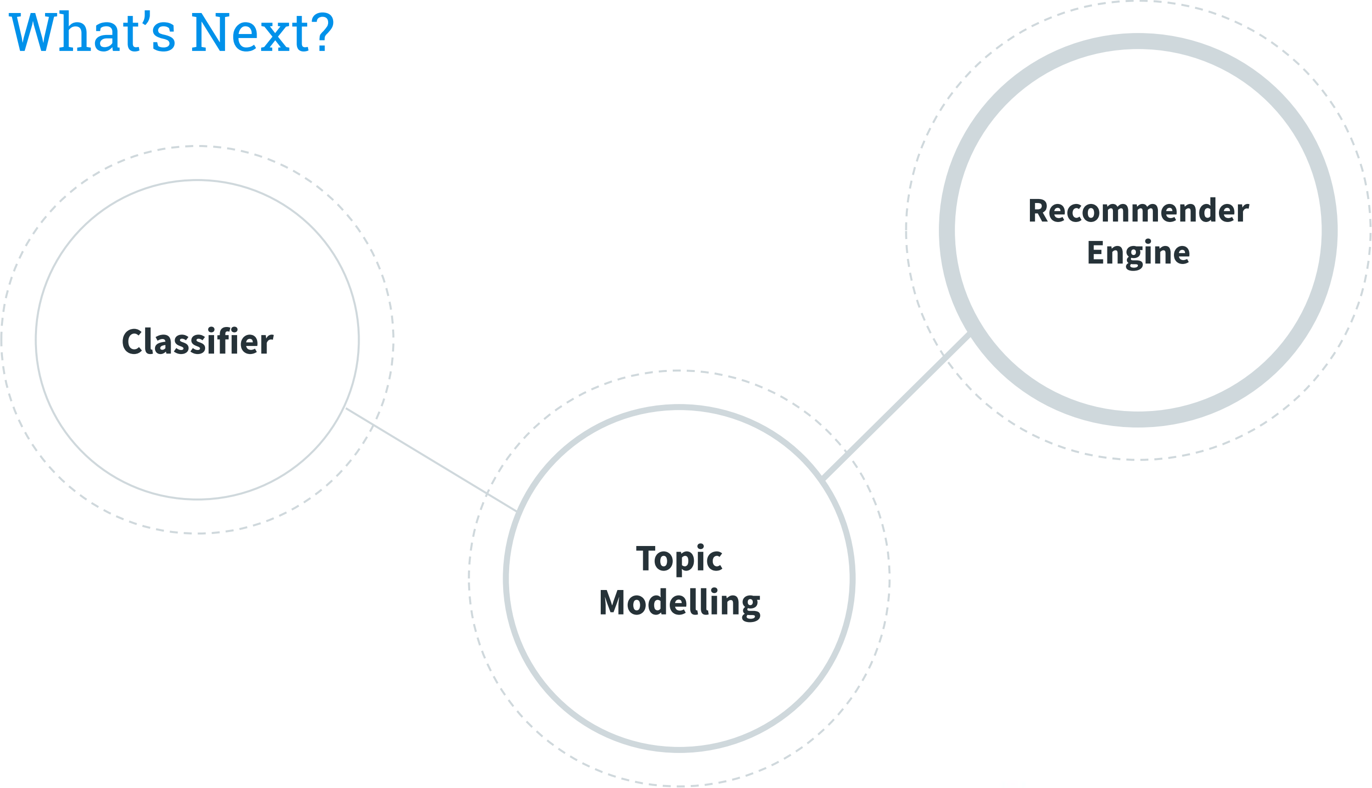
Positive Tweet Topic Counts (2300 tweets)



Negative Tweet Topic Counts (3100 tweets)


Tweet Topic Counts (5500 tweets)
Topic Counts
4000
3000
2000
1000
0
Climate Beliefs
Physics
Global Warming
Energy
Internet Conversations
Article / Link Discussion
Rising Temperatures
Atmosphere Change
Governmental Involvement
Water
Polar Ice
Legal
Thought Processes
Science
Climate Reports
Tweet Topics

What’s Next?



Classifier

Topic Modelling

Recommender Engine

What’s Next?

USER	TOPIC	CLASS	AGG - CLASS
1	a	1	**1**
1	a	0	
1	a	1	
1	b	0	**0**
1	b	1	
2	a	0	**0**
2	a	1	
2	b	1	**1**
2	c	0	**0**
2	c	0	

What’s Next?

- Finish data wrangling
- Fit Collaboratively Filtered Recommender

What’s Next? - Fitted Model

Climate Change					
	Policy	Energy	Education	Planet Life	Conversing
1	0	.2	1	.8	.4
2	1	.3	1	.5	0
3	.8	0	.8	1	.9
4	0	.6	1	.2	1
5	.7	1	1	1	.1

Future Considerations

Reddit Data

- Explore undersampling techniques

-Increase min_df when count vectorizing

-Test additional topic cluster combinations

Twitter Data

- Get more user climate change tweets

- Perform a Tweet/Reddit Comment comparison project to validate classifier

- Compile front end pipeline

Questions?