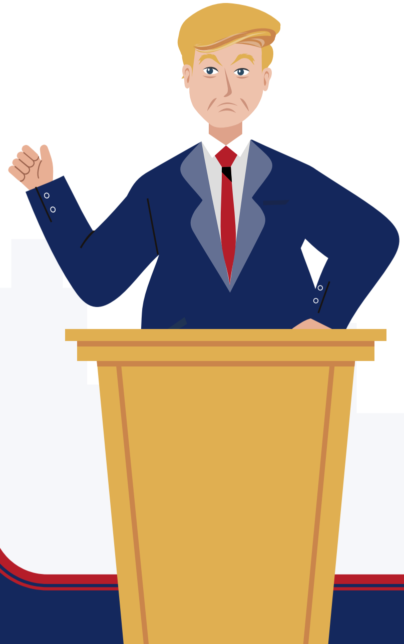


The USA representatives' articles analysis

Paulina Gacek, Filip Garbacki, Jakub Hulek

Project overview

Our aim was to collect a comprehensive dataset of articles and statements from various U.S. Congressmen and Congresswomen to explore key political patterns and get insights into political communication trends.

Considered Classification Tasks

1

Author Prediction

2

Political Affiliation Prediction

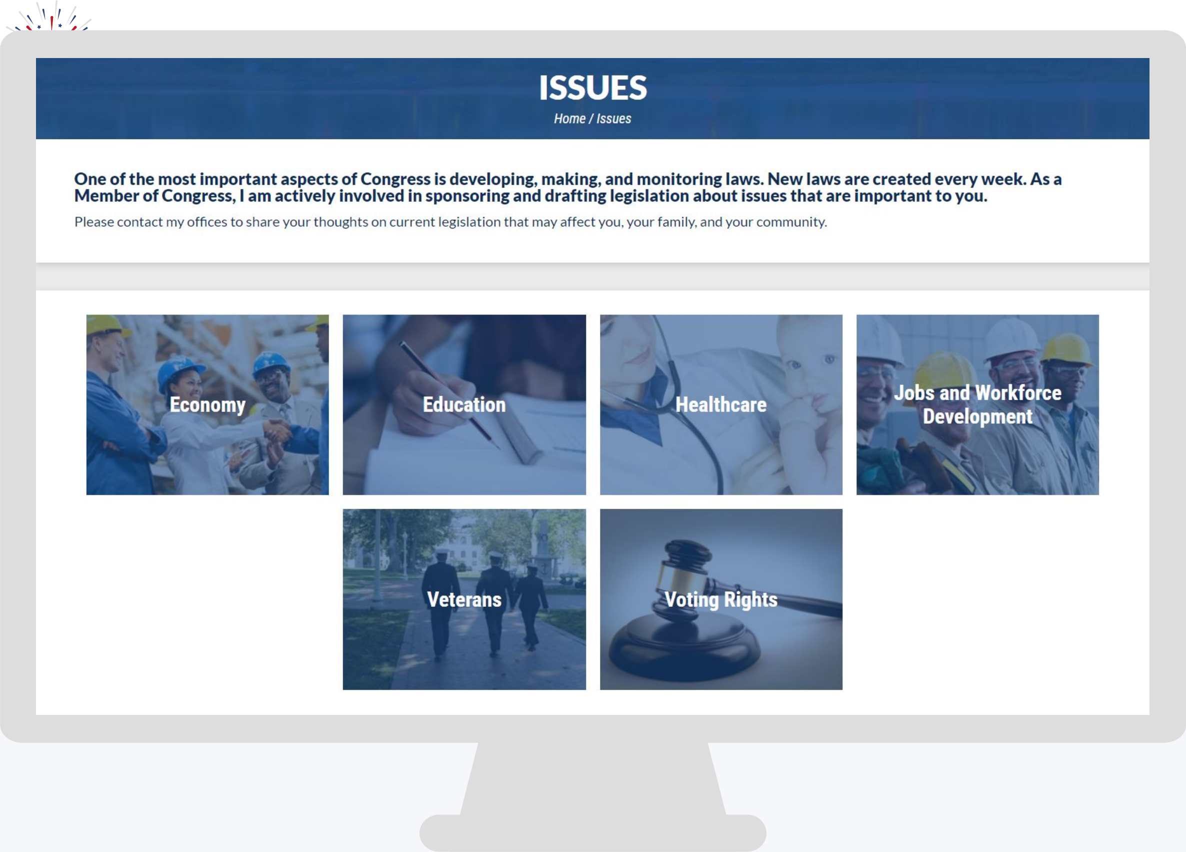
3

Topic Classification

01

Data Acquisition

Rep. Terri Sewell issues page



Rep. Terri Sewell articles on economy

More on Economy

Month | Year | -- Issues -- | Filter | Reset

October 2024

10/11/24 Rep. Sewell Delivers $1.6 Million in Federal Funding to Expand Mental Health and Substance Use Services in Jefferson County and Renovate the Loveman Village Daycare Facility

July 2024

7/11/24 Rep. Sewell Announces 13th Annual Job Fair

7/9/24 Rep. Sewell Announces $16.9 Million from President Biden's Bipartisan Infrastructure Law to Upgrade Montgomery Bus System

June 2024

6/24/24 Rep. Sewell Announces $14+ Million in Department of Transportation RAISE Grants for the Cities of York and Fairfield

6/18/24 Rep. Sewell Hosts Seniors Resource Roundtable in Linden

May 2024

5/29/24 Rep. Sewell Announces $16.9 Million in EPA Rebates for Clean School Buses in Alabama

5/15/24 Rep. Sewell Votes for FAA Reauthorization Act of 2024

5/1/24 Rep. Sewell Honors Award-Winning Small Businesses in Alabama's 7th Congressional District During National Small Business Week

April 2024

4/26/24 Rep. Sewell Continues Small Business Tour Through Alabama's 7th Congressional District Ahead of National Small Business Week

March 2024

3/27/24 Rep. Sewell Secures $13.1 Million in Direct Funding for Local Projects in the Fiscal Year 2024 Government Funding Package

« FIRST < PREV NEXT > LAST »

PAGE 1 OF 12

How to access and process 18000 pages in manageable time

Execution

Process for Representative #1

- Thread per page
- Thread per page
- Thread per page
- Thread per page
- Thread per page

200 threads

Process for Representative #2

- Thread per page
- Thread per page
- Thread per page
- Thread per page
- Thread per page

200 threads

This ensures that both whole processing power and whole internet bandwidth can be used.

Downloading and scrapping 5000 pages took only 6 minutes.

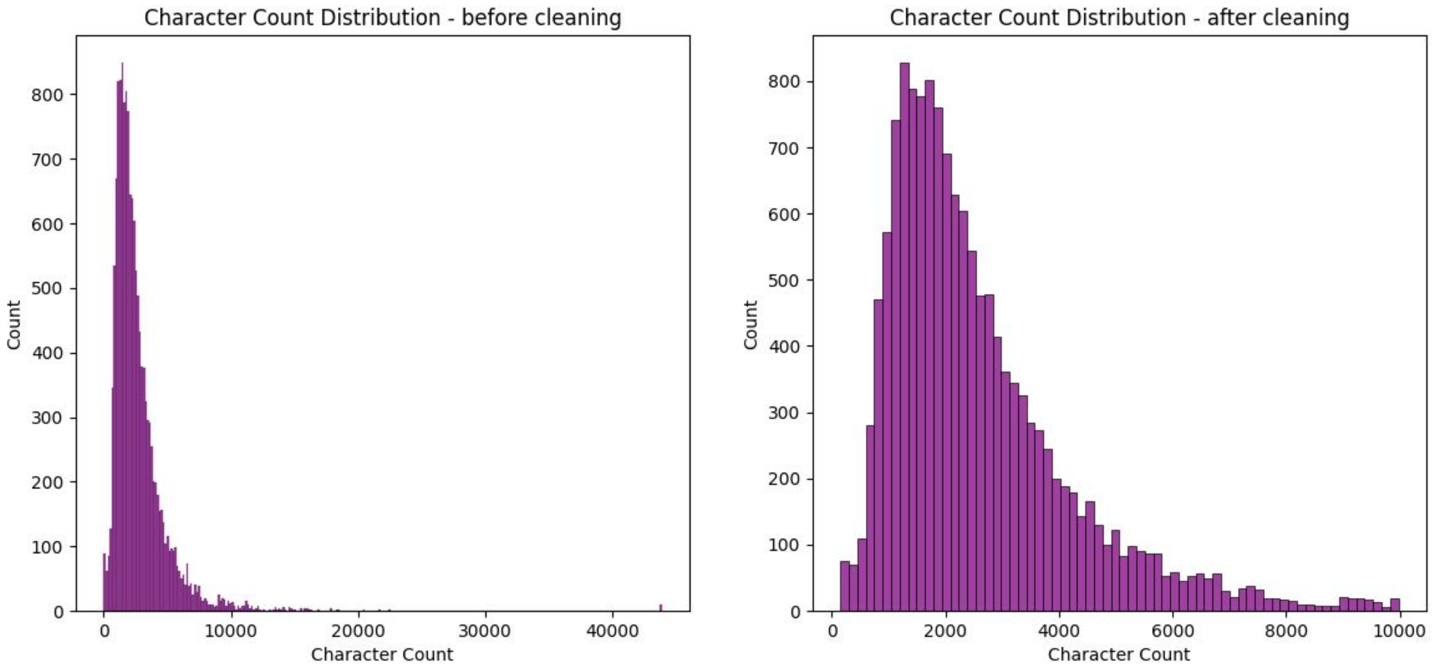
02

Data Preprocessing

1. Cleaning and formatting

1. Date format standardization
2. Article content cleaning
3. Removing faulty articles

2. Filtering by Character Count



in range (150; 10 000)

3. Topics standardization

Health Care and Social Security

Health; Healthcare;
Healthcare, Social Security, & Support Programs;
Social Security and Medicare;
Improving Access to Affordable Healthcare;
Social Security; Health Care;
What Rep. Flood is doing to expand healthcare options
...

Undefined

Back the Blue; Getting Things Done; Program Awards & Announcements; Crumbling Foundations; Congressman Larson's Committees
...

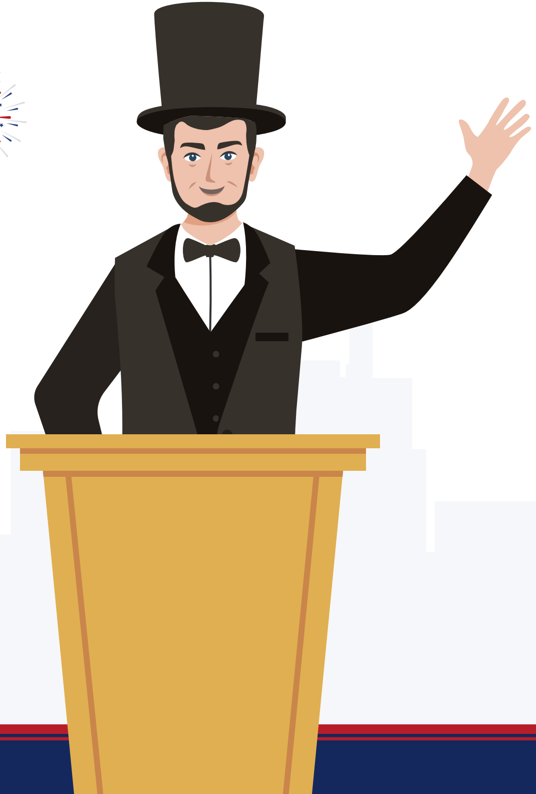
Final topics (from 358 to 19)

National Security, Defence, Foreign Affairs	3673
Health Care and Social Security	2930
Energy and Environment	2380
Jobs and the Economy	2333
Education	1411
Veterans and Military	1376
Government and Law	1015

Infrastructure and Transportation	823	2nd Amendment and Gun Violence	306
Agriculture	758	Supporting Seniors	146
Local issues	750	Pro-Life/Abortion and Family Values	141
Federal Budget and Taxes	667	Housing	115
Equality and Civil Rights	511	Disaster Relief & Preparedness	67
Science, Technology, & Telecommunications	337	Constitution	26

Articles in numbers

50 303	Collected articles
25 643	Cleaned up and UNIQUE articles
19 270	Articles with a single issue



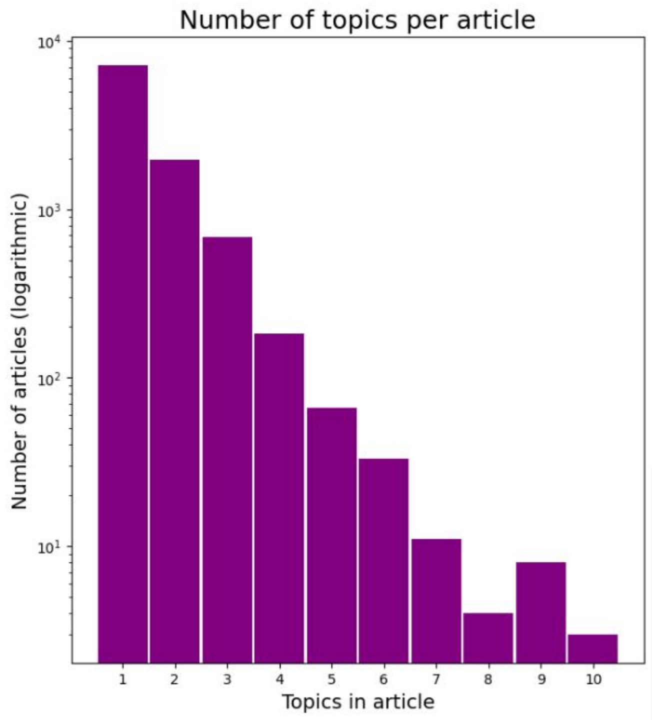
03

Explorative Data Analysis

Articles are complex!

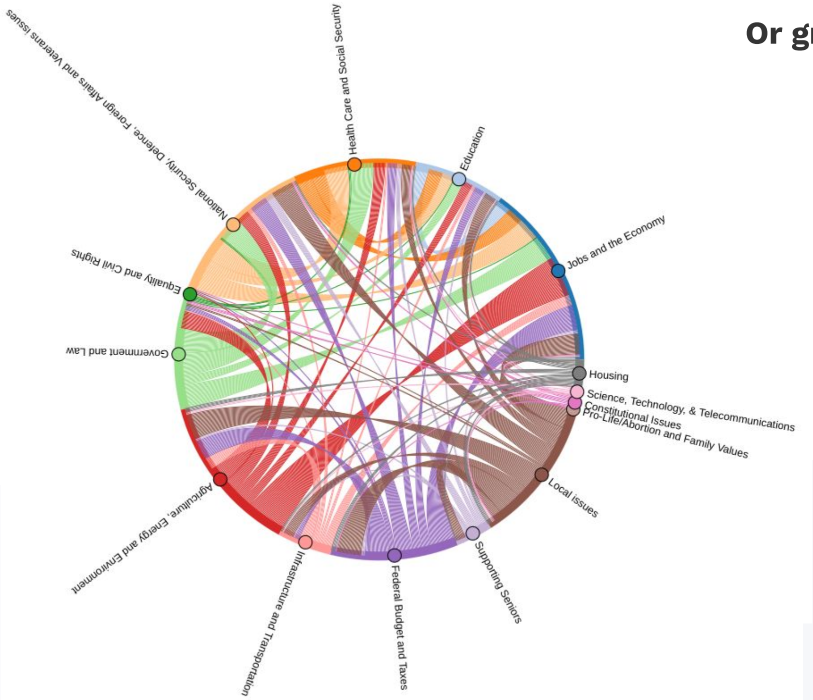
One article might regard multiple topics at once

How many topics appear only on their own?	
Constitutional Issues:	10.5%
Housing:	15.6%
Pro-Life/Abortion and Family Values:	16.7%
Supporting Seniors:	19.5%
Local issues:	32.0%
Federal Budget and Taxes:	39.5%
Government and Law:	40.7%
Science, Technology, & Telecommunications:	46.8%
Jobs and the Economy:	50.2%
Agriculture, Energy and Environment:	52.7%
Health Care and Social Security:	55.4%
National Security, Defence, Foreign Affairs and Veterans issues:	59.6%
Education:	59.9%
Infrastructure and Transportation:	60.4%
Equality and Civil Rights:	77.7%



Topics walk in pairs

Or groups, in fact



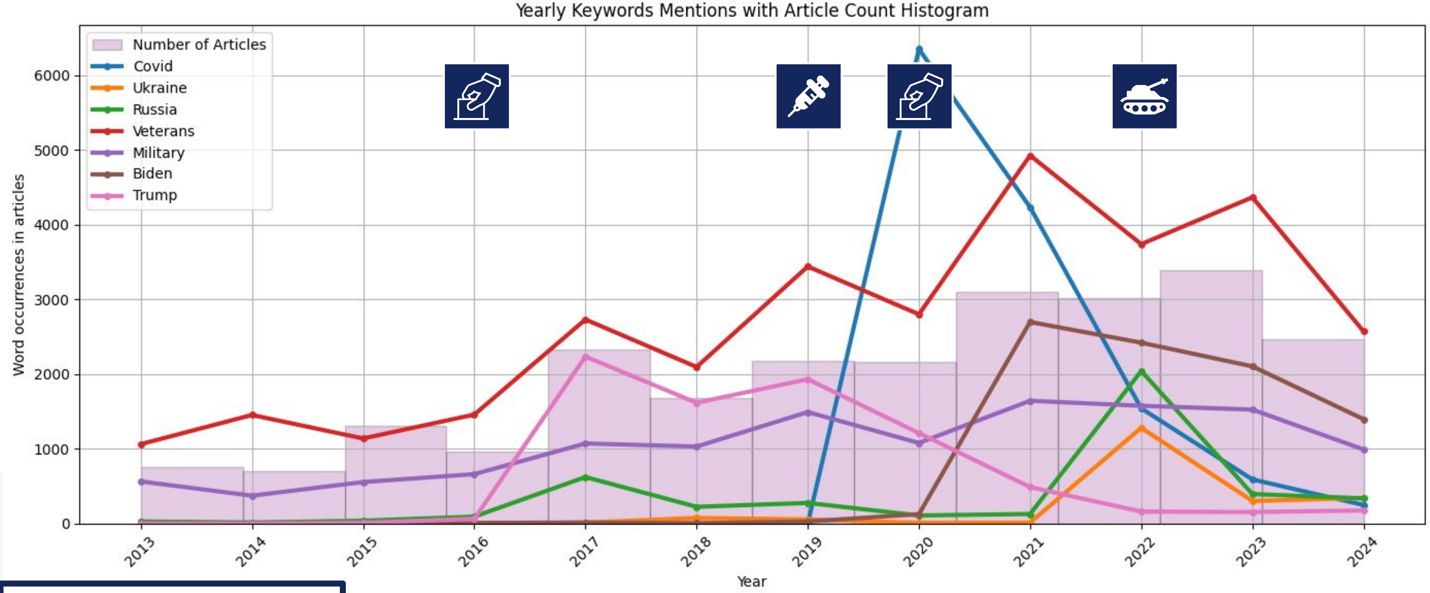
Those who support seniors also often mention healthcare, social security and veterans

Local issues are of all kinds - except law and government

If congressman tackles education, they won't be telling of farms, nuclear plants, nor environment

Everyone needs funding from the budget, but government's financing is discussed relatively little

What’s the buzz?



No matter what, veterans and army are always hot topics in the US Congress

04

The most common phrases

Bag of words, word clouds

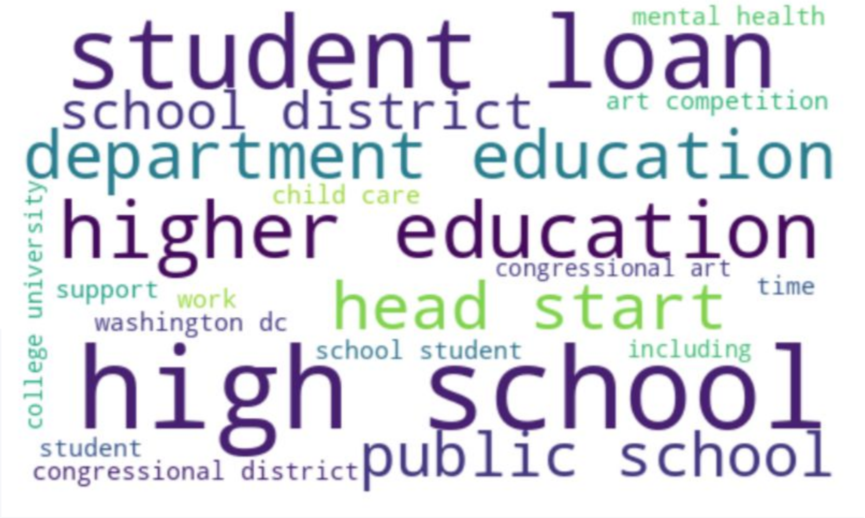
General-Purpose Words

★ act ★	★community★	year
★ state ★	★ program ★	american
today	member	house
american	rep	congressman

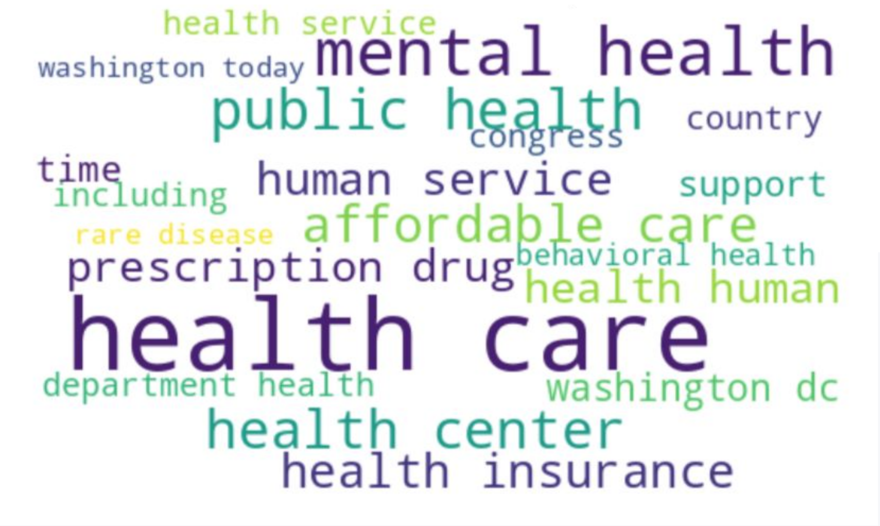
Extracted using BoW with
n-grams of length from 1 to 4

Topic-specific most common words

Education



Health Care and Social Security



(After removing general-purpose most common words)

05

Topic classification

Glove, Word2Vec and many more

Embedding - Part 1

Transfer Learning

Model	word2vec-google-news-300	fasttext-wiki-news-subwords-300	glove-twitter-200	glove-twitter-100	glove-twitter-50
Accuracy	85.68%	84.69%	84.65%	82.28%	77.35%
Training time	225s	223s	160s	80s	49s

Classification over 6 most common classes
(Identically structured neural networks, same train/test datasets, trained for 90 epochs)

Embedding - Part 1

Locally trained word2vec

Model	word2vec-google-news-300	Custom-300 Window = 5	Custom-150 Window = 5	Custom-150 Window = 7	Custom-150 Window = 9
Accuracy	85.68%	85.22%	85.71%	86.21%	86.18%
Training time	225s	228s	119s	119s	119s

Classification over 6 most common classes
(Identically structured neural networks, same train/test datasets, trained for 90 epochs)

Embedding - Part 2

Embedding approaches

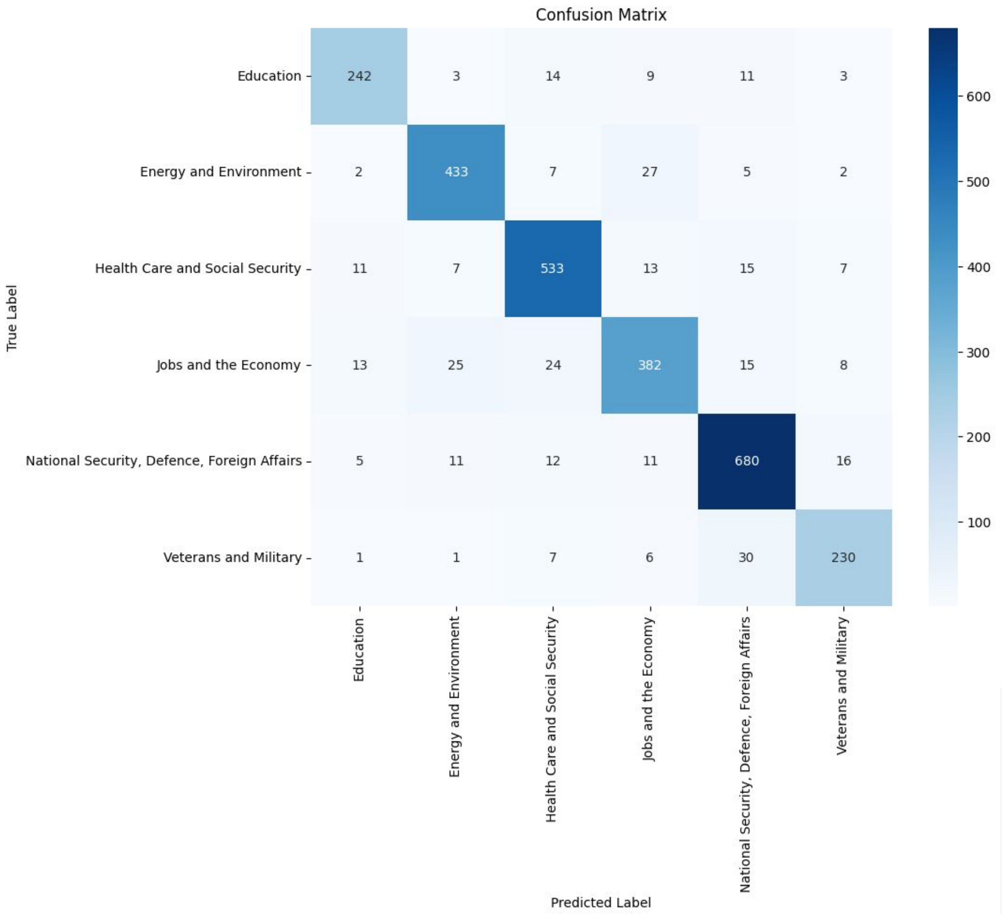
Bag of Words

Tested on various vocabulary sizes, only words with frequency above the threshold were taken into consideration

BERT Embedding

We used a sentence-transformers model: all-MiniLM-L6-v2.

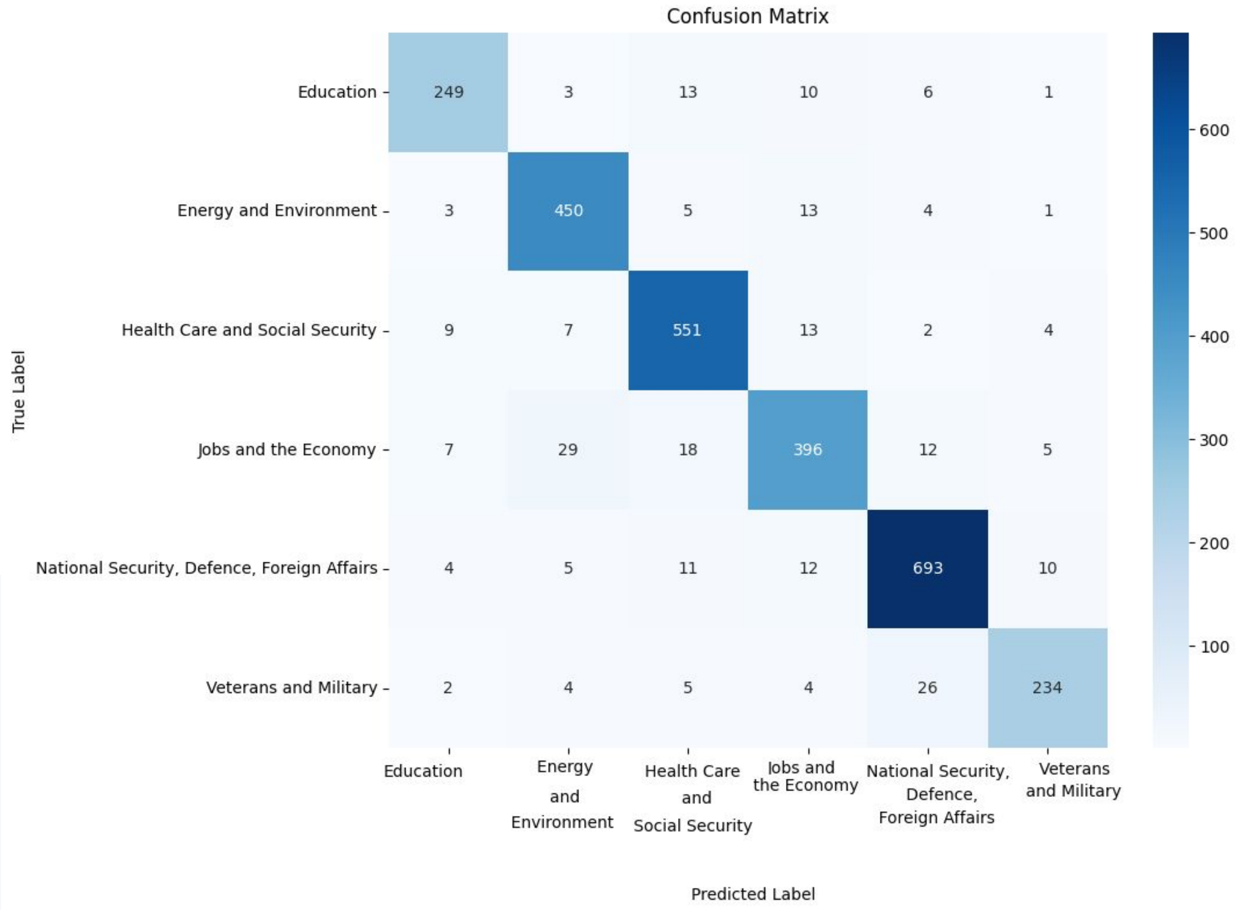
It maps sentences & paragraphs to a **384** dimensional dense vector space.



BoW Embedding

accuracy 90%
f1 90%
(on imbalanced test dataset)

3 hidden layers,
300 epochs



BERT Embedding

accuracy 91%
f1 91%
(on imbalanced test dataset)

3 hidden layers,
300 epochs

Confusion Matrix

True Label \ Predicted Label	2nd Amendment and Gun Violence	Agriculture	Education	Energy and Environment	Equality and Civil Rights	Federal Budget and Taxes	Government and Law	Health Care and Social Security	Infrastructure and Transportation	Jobs and the Economy	Local issues	National Security, Defence, Foreign Affairs	Science, Technology, & Telecommunications	Veterans and Military
2nd Amendment and Gun Violence	46	0	1	0	0	1	2	3	0	1	0	5	0	2
Agriculture	0	125	0	11	0	0	0	4	4	5	2	1	0	0
Education	0	0	253	0	2	0	2	4	1	10	0	7	0	3
Energy and Environment	0	1	4	409	1	1	3	3	1	30	3	16	1	3
Equality and Civil Rights	0	0	3	1	61	0	5	6	0	6	0	17	0	3
Federal Budget and Taxes	0	1	1	11	0	51	1	7	3	46	2	9	0	1
Government and Law	0	3	8	13	1	1	101	5	1	14	7	41	1	7
Health Care and Social Security	4	0	4	3	0	0	0	533	2	20	3	16	0	1
Infrastructure and Transportation	0	1	0	15	0	1	1	1	111	27	1	3	2	2
Jobs and the Economy	0	9	9	23	0	7	0	17	12	369	4	15	1	1
Local issues	0	0	8	5	0	0	4	7	0	11	100	11	0	4
National Security, Defence, Foreign Affairs	2	1	3	2	1	1	6	7	0	13	4	683	1	11
Science, Technology, & Telecommunications	0	0	3	0	0	0	0	2	1	12	1	7	40	1
Veterans and Military	0	0	1	0	0	0	1	3	0	3	0	34	0	233

Color scale: 0, 100, 200, 300, 400, 500, 600

BoW Embedding

81% accuracy on test dataset with 14 unbalanced topics

Confusion Matrix

True Label \ Predicted Label	2nd Amendment and Gun Violence	Agriculture	Education	Energy and Environment	Equality and Civil Rights	Federal Budget and Taxes	Government and Law	Health Care and Social Security	Infrastructure and Transportation	Jobs and the Economy	Local issues	National Security, Defence, Foreign Affairs	Science, Technology, & Telecommunications	Veterans and Military
2nd Amendment and Gun Violence	54	0	0	1	0	0	0	3	0	0	0	1	0	2
Agriculture	0	121	1	8	0	3	0	4	2	9	4	0	0	0
Education	1	0	244	1	3	2	2	5	3	11	5	2	0	3
Energy and Environment	0	3	3	390	0	14	5	5	15	29	2	6	1	3
Equality and Civil Rights	0	0	5	1	72	0	2	10	0	3	1	5	0	3
Federal Budget and Taxes	0	0	1	1	0	87	1	7	2	22	3	7	0	2
Government and Law	5	3	6	8	6	1	121	3	1	9	9	22	2	7
Health Care and Social Security	2	4	4	2	7	5	2	515	2	25	6	7	1	4
Infrastructure and Transportation	0	1	1	7	0	0	1	0	135	15	1	1	2	1
Jobs and the Economy	0	9	10	18	1	13	2	12	15	363	3	13	5	3
Local issues	2	2	10	2	1	6	2	4	3	8	102	7	0	1
National Security, Defence, Foreign Affairs	2	0	8	7	10	8	12	12	2	23	3	616	4	28
Science, Technology, & Telecommunications	0	0	3	1	4	0	0	0	0	8	0	3	47	1
Veterans and Military	0	0	3	0	0	2	1	6	1	4	2	16	0	240

BERT Embedding

81% accuracy on test dataset with 14 unbalanced topics

06

Key phrase extraction

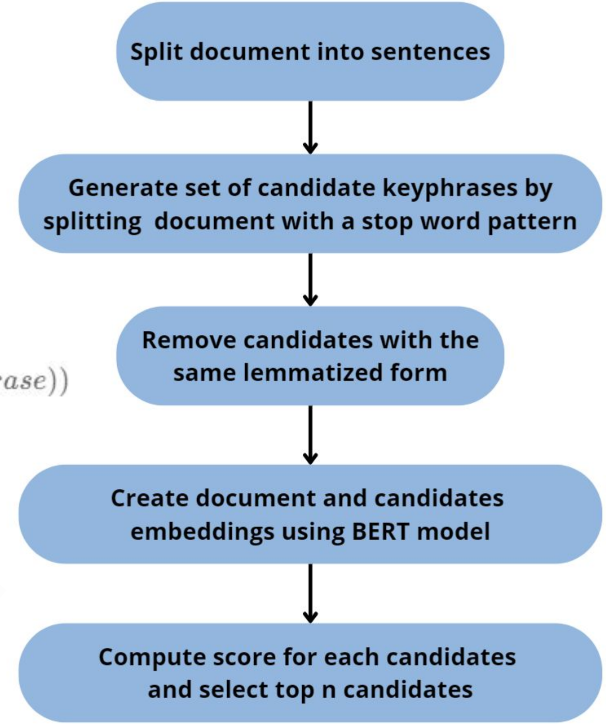
unsupervised approach

Keyphrase extraction algorithm

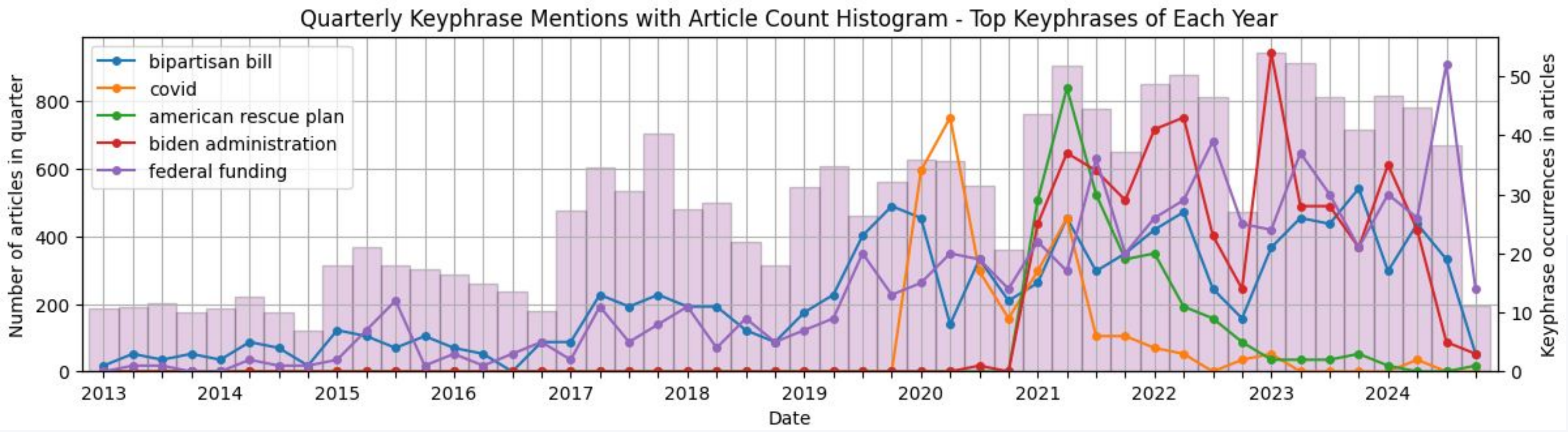
$$score(phrase) = similarity(embed(phrase), embed(doc)) \times prob(word\_count(phrase))$$

where:

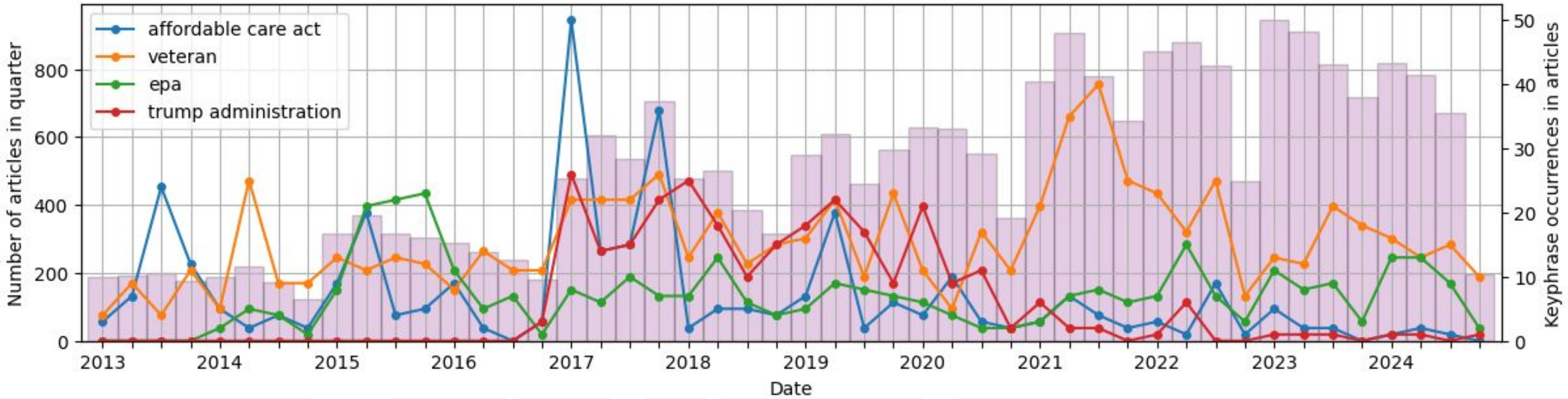
- doc - document text
- $phrase$ - candidate keyphrase



Time-Based Distribution of Articles Containing Selected Key Phrases (1)



Time-Based Distribution of Articles Containing Selected Key Phrases (2)



10 most frequently appearing key phrases between 2013 and 2024

Key phrase	Count
veteran	747
federal funding	660
bill	622
bipartisan bill	590
funding	569
national security	453
legislation	447
biden administration	445
representative	408
bipartisan legislation	403

Bonus visualization

Thank you!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**