Political Polarization

CSCI 577

Matt Jensen

May 18, 2023

Outline

- Hypothesis
- Sources
- Data Workup
- Experiments
- Remaining Work
- Questions

Hypothesis

Hypothesis

Political polarization is rising, and news articles are a proxy measure.

Why might we expect this?

Why might we expect this?

Mostly anecdotal experience.

Why might we expect this?

Mostly anecdotal experience.

Evidence is mixed in the literature [1,2,3].

Why might we expect this?

Mostly anecdotal experience.

Evidence is mixed in the literature [1,2,3].

Our goal is whether, not why.

Sub-hypothesis

Sub-hypothesis

- The polarization is not evenly distributed across publishers.

Sub-hypothesis

- The polarization is not evenly distributed across publishers.
- The polarization is not evenly distributed across political specturm.

Sub-hypothesis

- The polarization is not evenly distributed across publishers.
- The polarization is not evenly distributed across political specturm.
- The polarization increases near elections.

Sub-sub-hypothesis

Sub-sub-hypothesis

- Similarly polarized publishers link to each other.

Sub-sub-hypothesis

- Similarly polarized publishers link to each other.
- 'Mainstream' media uses more neutral titles.

Sub-sub-hypothesis

- Similarly polarized publishers link to each other.
- 'Mainstream' media uses more neutral titles.
- Highly polarized publications don't last as long.

Data Sources

Data Sources

- Memeorandum:
- AllSides:
- HuggingFace:
- ChatGPT:

Data Sources

- Memeorandum: **stories**
- AllSides:
- HuggingFace:
- ChatGPT:

Data Sources

- Memeorandum: **stories**
- AllSides: **bias**
- HuggingFace:
- ChatGPT:

Data Sources

- Memeorandum: **stories**
- AllSides: **bias**
- HuggingFace: **sentiment**
- ChatGPT:

Data Sources

- Memeorandum: **stories**
- AllSides: **bias**
- HuggingFace: **sentiment**
- ChatGPT: **election dates**

Memeorandum

Memeorandum

Memeorandum

- News aggregation site.

Memeorandum

- News aggregation site.
- Was really famous before Google News.

Memeorandum

- News aggregation site.
- Was really famous before Google News.
- Still aggregates sites today.

Memeorandum

Memeorandum

- I still use it.

Memeorandum

- I still use it.
- I like to read titles.

Memeorandum

- I still use it.
- I like to read titles.
- Publishers block bots.

Memeorandum

- I still use it.
- I like to read titles.
- Publishers block bots.
- Simple html to parse.

Memeorandum

- I still use it.
- I like to read titles.
- Publishers block bots.
- Simple html to parse.
- Headlines from 2006 forward.

Memeorandum

- I still use it.
- I like to read titles.
- Publishers block bots.
- Simple html to parse.
- Headlines from 2006 forward.
- Automated, not editorialized.

AllSides

AllSides

AllSides

- Rates publications as left, center or right.

AllSides

- Rates publications as left, center or right.
- Ratings combine:
  - blind bias surveys.
  - editorial reviews.
  - third party research.
  - community voting.

AllSides

AllSides

- One of the only bias apis.

AllSides

- One of the only bias apis.
- Ordinal ratings [-2: very left, 2: very right].

AllSides

- One of the only bias apis.
- Ordinal ratings [-2: very left, 2: very right].
- Covers 1400 publishers + some blog and authors.

AllSides

- One of the only bias apis.
- Ordinal ratings [-2: very left, 2: very right].
- Covers 1400 publishers + some blog and authors.
- Easy format and semi-complete data.

HuggingFace

HuggingFace

HuggingFace

- Deep learning library.

HuggingFace

- Deep learning library.
- Lots of pretrained models.

HuggingFace

- Deep learning library.
- Lots of pretrained models.
- Easy, off the shelf word/sentence embeddings and text classification models.

HuggingFace

- Language models are .

HuggingFace

- Language models are **HOT**.

HuggingFace

- Language models are **HOT**.
- Literally 5 lines of python.

HuggingFace

- Language models are **HOT**.
- Literally 5 lines of python.
- The dataset needed more features.

HuggingFace

- Language models are **HOT**.
- Literally 5 lines of python.
- The dataset needed more features.
- Testing different model performance was easy.

HuggingFace

- Language models are **HOT**.
- Literally 5 lines of python.
- The dataset needed more features.
- Testing different model performance was easy.
- Lots of pretrained classification tasks.

Data Collection

Data Collection

Stories

```python
day = timedelta(days=1)
cur = date(2005, 10, 1)
end = date.today()
while cur <= end:
    cur = cur + day
    save_as = output_dir / f"{cur.strftime('%y-%m-%d')}.html"
    url = f"https://www.memeorandum.com/{cur.strftime('%y%m%d')}/
    r = requests.get(url)
    with open(save_as, 'w') as f:
        f.write(r.text)
```

Data Collection

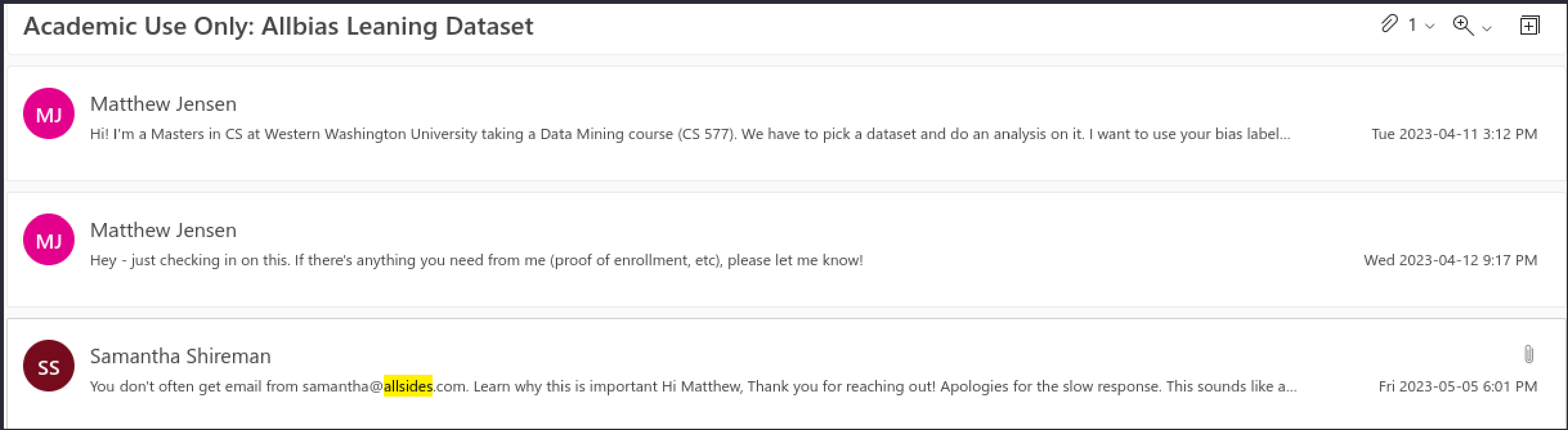
Bias **hard**

```python
...
bias_html = DATA_DIR / 'allsides.html'
parser = etree.HTMLParser()
tree = etree.parse(str(bias_html), parser)
root = tree.getroot()
rows = root.xpath('//table[contains(@class,"views-table")]/tbody/

ratings = []
for row in rows:
    rating = dict()
    ...
```

Data Collection

Bias **easy**



Data Collection

Embeddings

```
# table = ...
tokenizer = AutoTokenizer.from_pretrained("roberta-base")
model = AutoModel.from_pretrained("roberta-base")

for chunk in table:
    tokens = tokenizer(chunk, add_special_tokens = True, truncati
    outputs = model(**tokens)
    embeddings = outputs.last_hidden_state.detach().numpy()
    ...
```

Data Collection

Classification Embeddings

```
...
outputs = model(**tokens)[0].detach().numpy()
scores = 1 / (1 + np.exp(-outputs))  # Sigmoid
class_ids = np.argmax(scores, axis=1)
for i, class_id in enumerate(class_ids):
    results.append({"story_id": ids[i], "label" : model.config.id
...
```

Data Structures

Stories

Data Structures

Stories

Data Structures

Stories

- Top level stories.
    - title, author, publisher, url, date.

Data Structures

Stories

- Top level stories.
  - title, author, publisher, url, date.
- Related discussion.
  - publisher, url.
  - uses 'parent' story as a source.

Data Structures

Stories

- Top level stories.
  - title, author, publisher, url, date.
- Related discussion.
  - publisher, url.
  - uses 'parent' story as a source.
- Story stream changes constantly (dedup. required).

Data Structures

Stories

published_at date	title varchar	name varchar	author varchar	url varchar
2017-01-18	FBI, 5 other agencies probe possible cov	McClatchy Washington Bureau	NULL	http://www.mcclatchydc.com/news/politics
2017-01-18	FBI, other agencies probing possible Rus	The Hill	Mark Hensch	http://thehill.com/policy/national-secur
2017-01-18	Assange lawyer: Manning commutation does	The Hill	Joe Uchill	http://thehill.com/policy/cybersecurity/
2017-01-18	Earnest: GOP intellectually dishonest on	The Hill	Jordan Fabian	http://thehill.com/policy/technology/314
2017-01-18	The Betsy DeVos Hearing Was an Insult to	Esquire	Charles P. Pierce	http://www.esquire.com/news-politics/pol
2017-01-18	Betsy DeVos Fight Demonstrates Donald Tr	The Daily Beast	Matt Lewis	http://www.thedailybeast.com/articles/20
2017-01-18	De Blasio: Don't ‘overstate the threat’	Politico	Eliza Shapiro	http://www.politico.com/states/new-york/
2017-01-18	Betsy DeVos Cites Grizzly Bears During G	NBC News	Alastair Jamieson	http://www.nbcnews.com/news/us-news/bets
2017-01-18	Betsy DeVos apparently ‘confused’ about	Washington Post	Valerie Strauss	http://www.washingtonpost.com/news/answe
2017-01-18	Trump inauguration time, how to watch, a	Vox	Tara Golshan	http://www.vox.com/policy-and-politics/2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2021-11-11	White Supremacy on Trial: From Rittenhou	Democracy Now	NULL	https://www.democracynow.org/2021/11/11/
2021-11-11	Want to prevent future cases like Kyle R	Washington Examiner	Zachary Faria	https://www.washingtonexaminer.com/opini
2021-11-11	'I hope it's a fever that will break': G	Politico	David Siders	https://www.politico.com/news/2021/11/11
2021-11-11	How Likely Is a Democratic Comeback Next	New York Times	Kyle Kondik	https://www.nytimes.com/2021/11/11/opini
2021-11-11	Too dumb to check: Biden to improve stan	HotAir	Ed Morrissey	https://hotair.com/ed-morrissey/2021/11/
2021-11-11	San Francisco police officer dies of cov	Washington Post	Andrea Salcedo	https://www.washingtonpost.com/nation/20
2021-11-11	Iran-Backed Militants Storm US Embassy i	Washington Free Beacon	Adam Kredo	https://freebeacon.com/national-security
2021-11-11	Biden calls Satchel Paige ‘the great neg	Fox News	Jessica Chasmar	https://www.foxnews.com/politics/biden-s
2021-11-11	Can We Talk About Critical Race Theory?	New York Times	Jay Caspian Kang	https://www.nytimes.com/2021/11/11/opini
2021-11-11	Shelby of Alabama is said to plan $5 mil	Washington Post	Michael Scherer	https://www.washingtonpost.com/politics/

? rows (>9999 rows, 20 shown) 5 columns

Data Structures

Stories

parent_id int64	url varchar	publisher varchar
5666868610266459443	http://thinkprogress.org/federal-investi	thinkprogress.org
5666868610266459443	http://washingtonmonthly.com/2017/01/18/	Washington Monthly
5666868610266459443	http://www.thedailybeast.com/cheats/2017	The Daily Beast
5666868610266459443	http://www.balloon-juice.com/2017/01/18/	Balloon Juice
5666868610266459443	http://littlegreenfootballs.com/article/	littlegreenfootballs.com
5666868610266459443	http://mashable.com/2017/01/18/hack-paid	Mashable
5666868610266459443	http://digbysblog.blogspot.com/2017/01/n	Hullabaloo
5666868610266459443	http://hotair.com/archives/2017/01/18/mc	HotAir
5666868610266459443	http://www.motherjones.com/kevin-drum/20	Mother Jones
5666868610266459443	http://occupydemocrats.com/2017/01/18/ci	Occupy Democrats
.	.	.
.	.	.
.	.	.
8561148704463968832	http://www.politico.com/story/2016/10/do	Politico
-9146493099307035426	http://rightwingnews.com/top-news/hiding	rightwingnews.com
-9146493099307035426	http://ijr.com/wildfire/2016/10/723674-a	IJR
-9146493099307035426	http://www.breitbart.com/big-government/	Breitbart
-9146493099307035426	http://mobile.wnd.com/2016/10/congress-l	WorldNetDaily
656449242448385764	http://www.theguardian.com/commentisfree	The Guardian
656449242448385764	http://www.theatlantic.com/business/arch	The Atlantic
656449242448385764	http://www.engadget.com/2016/10/28/faceb	Engadget
656449242448385764	http://dailycaller.com/2016/10/28/facebo	The Daily Caller
656449242448385764	http://consumerist.com/2016/10/28/facebo	The Consumerist

? rows (>9999 rows, 20 shown) 3 columns

Data Structures

Stories

metric	value
total stories	299714
total related	960111
publishers	7031
authors	34346
max year	2023
min year	2005

Data Selection

Stories

Data Selection

Stories

- Clip the first and last full year of stories.

Data Selection

Stories

- Clip the first and last full year of stories.
- Remove duplicate stories (big stories span multiple days).

Data Selection

Stories

- Clip the first and last full year of stories.
- Remove duplicate stories (big stories span multiple days).
- Convert urls to tld to link to publishers.

Data Selection

Publishers

Data Selection

Publishers

- Combine subdomains of stories.
  - blog.washingtonpost.com and washingtonpost.com are considered the same publisher.
  - This could be bad. For example: opinion.wsj.com != wsj.com.

Data Selection

Publishers

- Combine subdomains of stories.
  - blog.washingtonpost.com and washingtonpost.com are considered the same publisher.
  - This could be bad. For example: opinion.wsj.com != wsj.com.
- Find common name of publisher.

Data Selection

Related

Data Selection

Related

- Select only stories with publishers whose story had been a 'parent' ('original publishers').
    - Eliminates small blogs and non-original news.

Data Selection

Related

- Select only stories with publishers whose story had been a 'parent' ('original publishers').
  - Eliminates small blogs and non-original news.
- Eliminate publishers without links to original publishers.
  - Eliminate silo'ed publications.
  - Link matrix is square and low'ish dimensional.

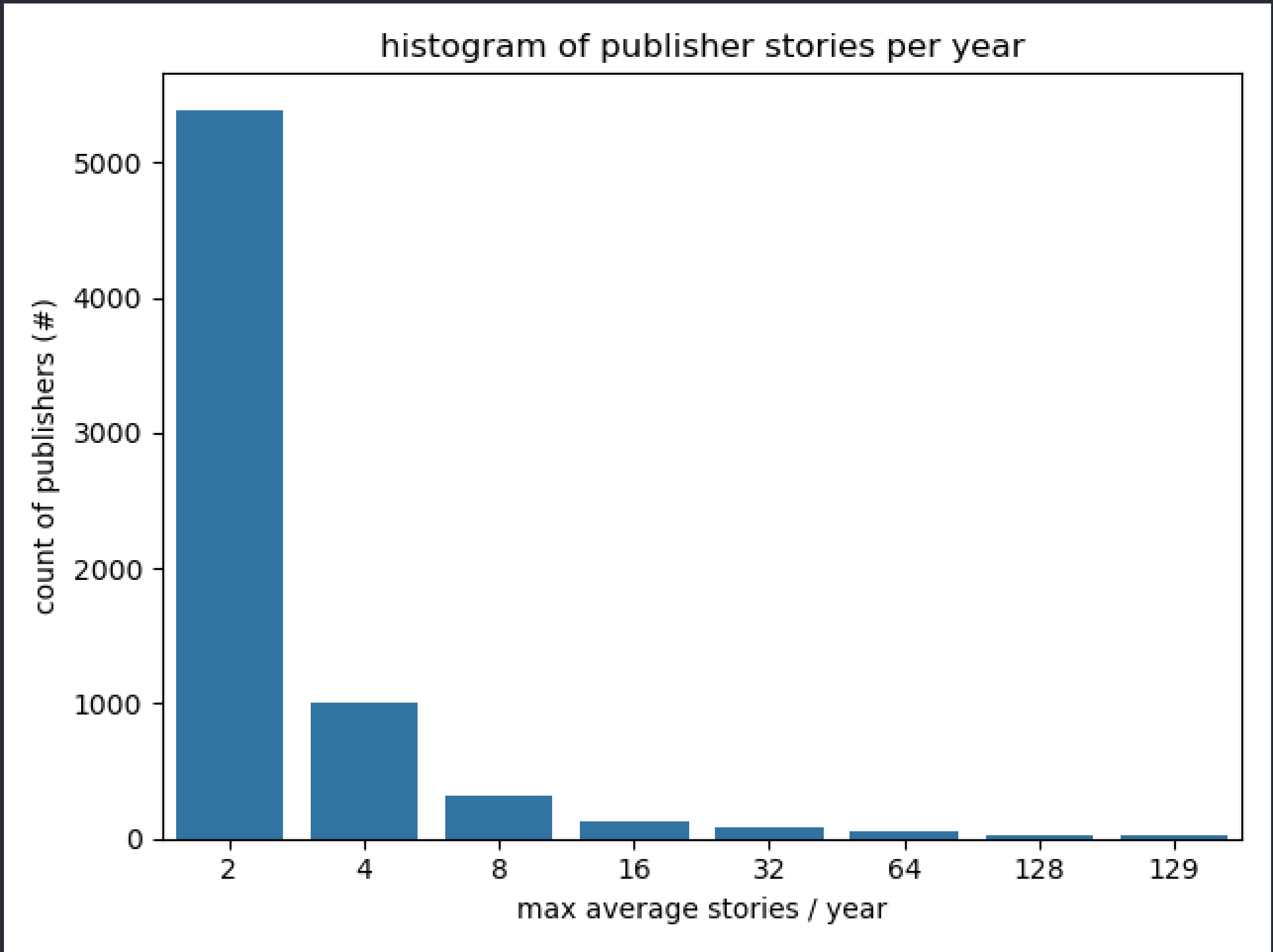
Data Selection

Post Process

metric	value
total stories	251553
total related	815183
publishers	223
authors	23809
max year	2022
min year	2006

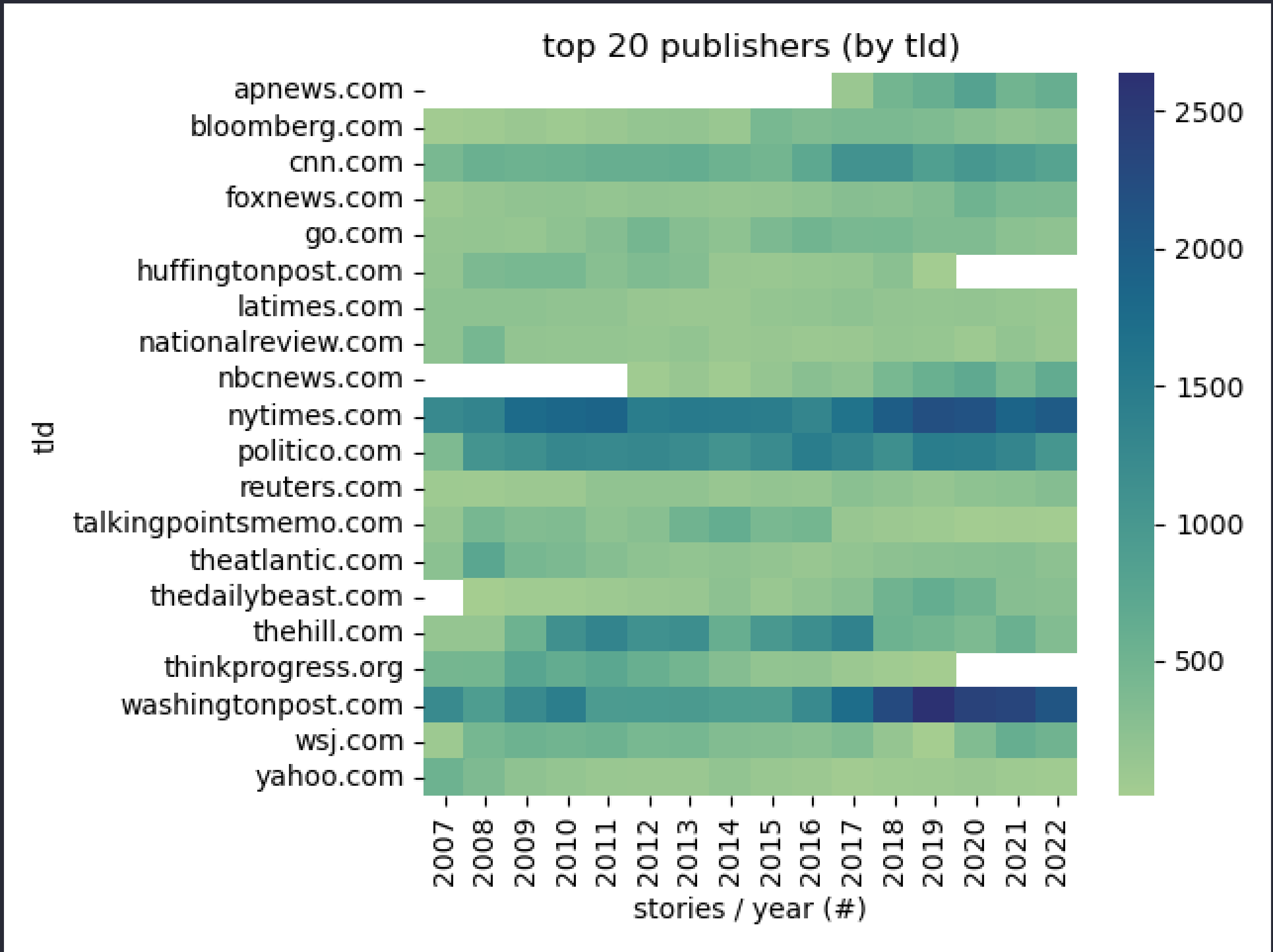
Descriptive Stats

Stories Per Publisher



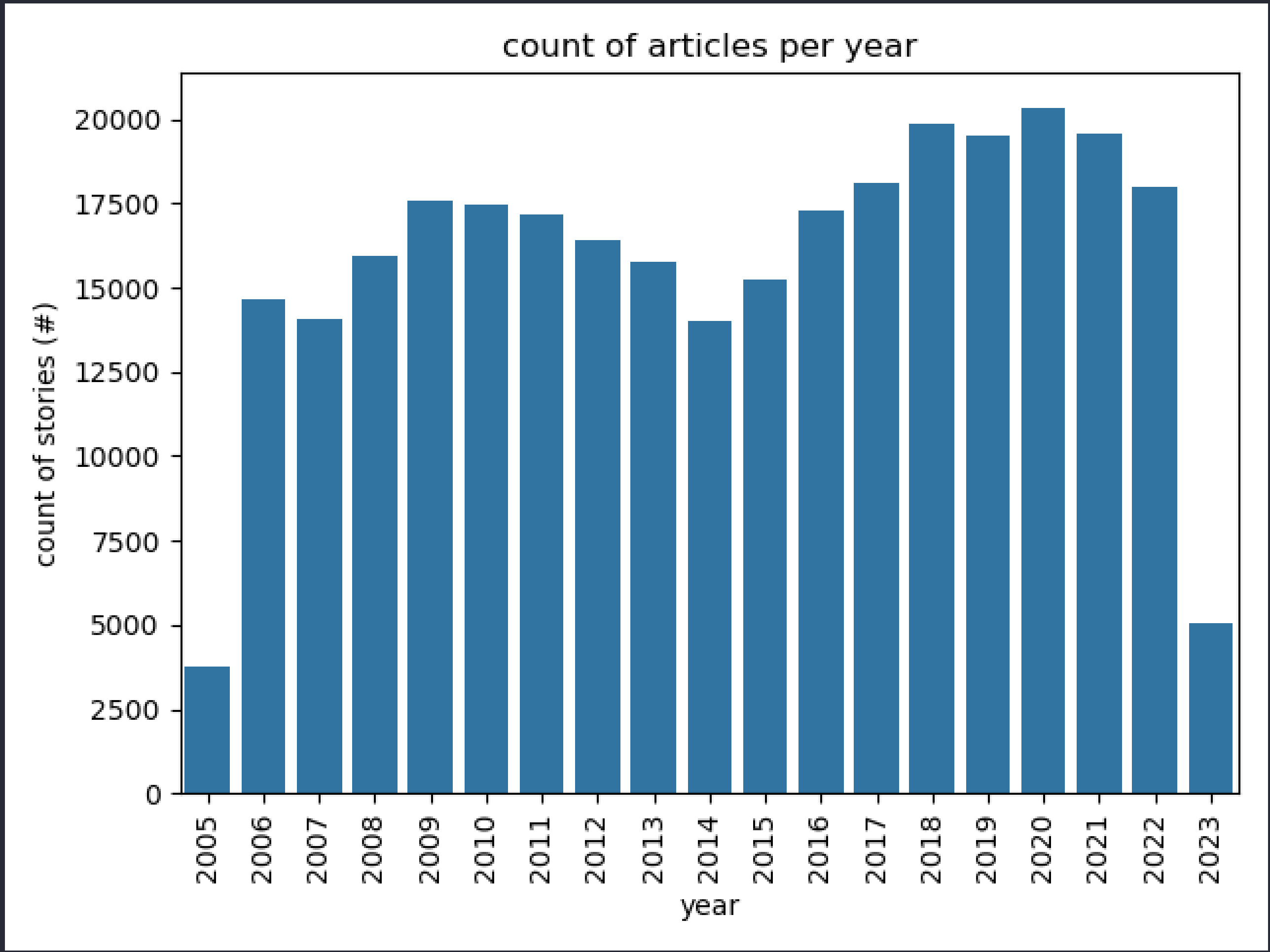
Descriptive Stats

Top Publishers



Descriptive Stats

Articles Per Year



Descriptive Stats

Common TLDs

tld	publishers
com	8660
org	1693
gov	636
net	245
edu	221
uk	138
us	66
ca	34
au	30
mil	30
tv	22
news	19
eu	16
nu	15
int	14

Data Structures

Bias

Data Structures

Bias

Data Structures

Bias

- Per publisher.
  - name,
  - label/ordinal value.
  - agree/disagree vote by community.

Data Structures

Bias

- Per publisher.
  - name,
  - label/ordinal value.
  - agree/disagree vote by community.
- Name could be semi-automatically joined to stories.

Data Structures

Bias

publisher varchar	label varchar	ordinal int32	agree int64	disagree int64
Karol Markowicz	right	2	43	47
'The Conversation' Contributor	left-center	-1	14	11
'The Fulcrum' Contributor	center	0	8	13
A Project for America	center	0	27	17
AARP	center	0	1955	4106
ABC News (Online)	left-center	-1	40679	20159
ACLU	left-center	-1	2496	3604
AJ+	left	-2	760	278
AZ Central	center	0	337	625
Aaron Rupar	left	-2	213	112
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Yes! Magazine	left	-2	496	270
York Dispatch	center	0	1	1
Yuma Sun	center	0	0	0
Zack Beauchamp	left-center	-1	27	33
Zeeshan Aleem	left	-2	22	19
ZeroHedge	right-center	1	258	199
azcentral	center	0	6	6
nj.com	left-center	-1	26	29
redefinED	center	0	186	107
theSkimm	left-center	-1	16	13
1582 rows (20 shown)				5 columns

Data Selection

Bias

Data Selection

Bias

- Keep all ratings.

Data Selection

Bias

- Keep all ratings.
- Join datasets on publisher name.
    - Started with 'jaro winkler similarity' then manually from there (look up Named Entity Recognition).

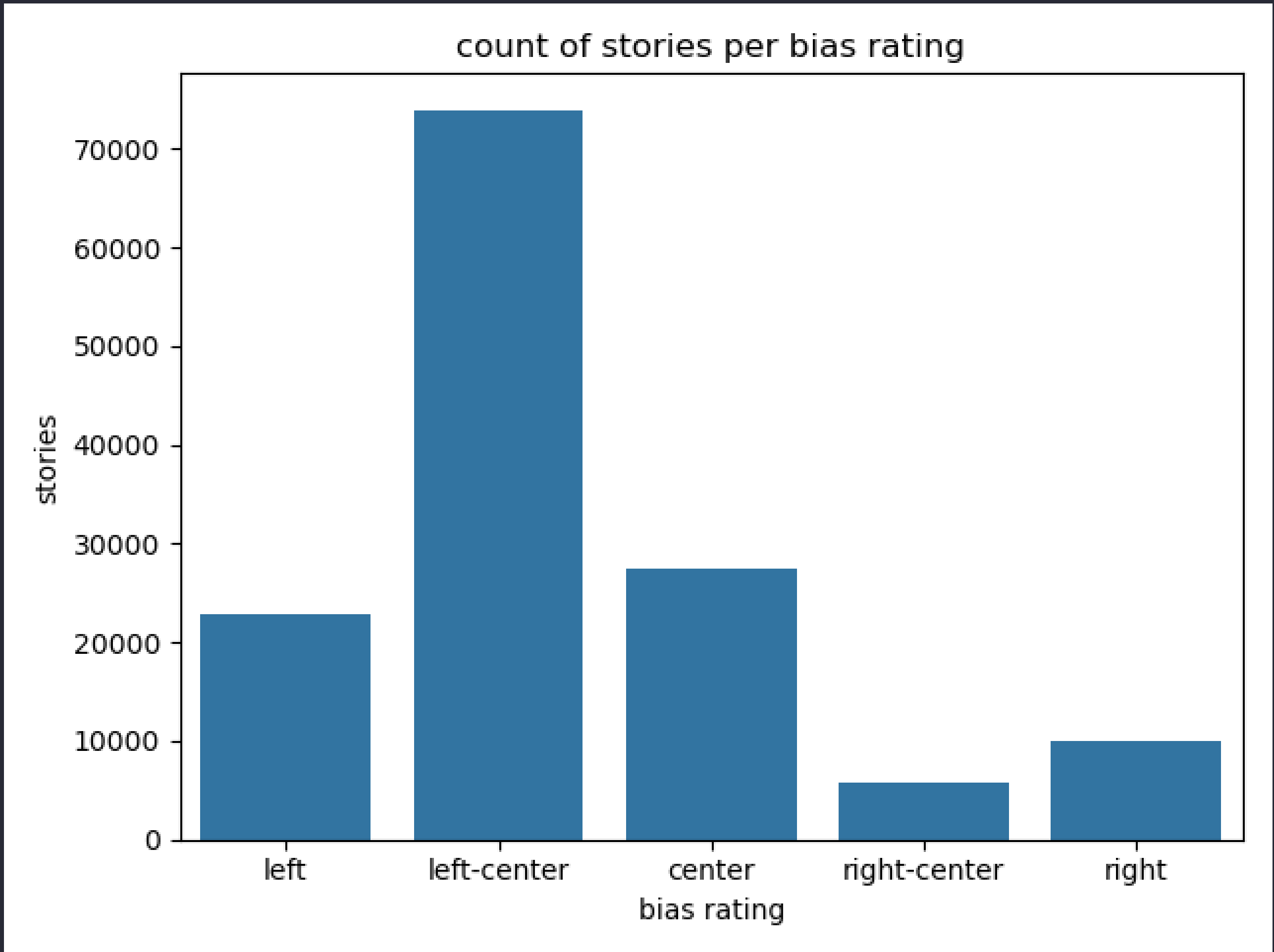
Data Selection

Bias

- Keep all ratings.
- Join datasets on publisher name.
    - Started with 'jaro winkler similarity' then manually from there (look up Named Entity Recognition).
- Use numeric values.
    - [left: -2, left-center: -1, ...].
    - Possibly scale ordinal based on agree/disagree ratio.

Data

Bias



count of stories per bias rating
stories
70000
60000
50000
40000
30000
20000
10000
0
left
left-center
center
right-center
right
bias rating

Data

Bias

bias varchar	ordinal int32	publishers int64	stories int64
left	-2	20	22839
left-center	-1	27	73934
center	0	33	27426
right-center	1	7	5726
right	2	14	9924

Data Structures

Embeddings

Data Structures

Embeddings

Data Structures

Embeddings

- Per story title.
  - sentence embedding (n, 384) - **BERT**.
  - sentiment classification (n, 1) - **RoBERTa base**.
  - emotional classification (n, 1) - **RoBERTa Go-Emotions**.

Data Structures

Embeddings

- Per story title.
  - sentence embedding (n, 384) - **BERT**.
  - sentiment classification (n, 1) - **RoBERTa base**.
  - emotional classification (n, 1) - **RoBERTa Go-Emotions**.
- ~ 1 hour of inference time to map story titles and descriptions.

Data Selection

Embeddings

Data Selection

Embeddings

- Word embeddings were too complicated.

Data Selection

Embeddings

- Word embeddings were too complicated.
- Kept argmax of classification prediction ([0.82, 0.18] -> LABEL_0).

Data Selection

Embeddings

- Word embeddings were too complicated.
- Kept argmax of classification prediction ([0.82, 0.18] -> LABEL_0).
- For publisher based analysis, averaged sentence embeddings for all stories.

Data

Embeddings

label	stories	publishers
positive	87830	223
negative	163723	223

Data

Embeddings

label	stories	publishers
neutral	124257	223
anger	34124	223
fear	36756	223
sadness	27449	223
disgust	17939	222
surprise	5710	216

Experiments

Experiments

1. **clustering** on link similarity.
2. **classification** on link similarity.
3. **classification** on sentence embedding.
4. **classification** on sentiment analysis.
5. **regression** on emotional classification over time and publication.

Experiment 1

clustering on link similarity.

Experiment 1

Setup

Experiment 1

Setup

- Create one-hot encoding of links between publishers.

Experiment 1

Setup

- Create one-hot encoding of links between publishers.
- Cluster the encoding.

Experiment 1

Setup

- Create one-hot encoding of links between publishers.
- Cluster the encoding.
- Expect similar publications in same cluster.

Experiment 1

Setup

- Create one-hot encoding of links between publishers.
- Cluster the encoding.
- Expect similar publications in same cluster.
- Use PCA to visualize clusters.

Experiment 1

Encoding schemes

Experiment 1

One-hot Encoding

publisher	nytimes	wsj	newsweek	...
nytimes	1	1	1	...
wsj	1	1	0	...
newsweek	0	0	1	...
...	...	...	...	...

Experiment 1

n-Hot Encoding

publisher	nytimes	wsj	newsweek	...
nytimes	11	1	141	...
wsj	1	31	0	...
newsweek	0	0	1	...
...	...	...	...	...

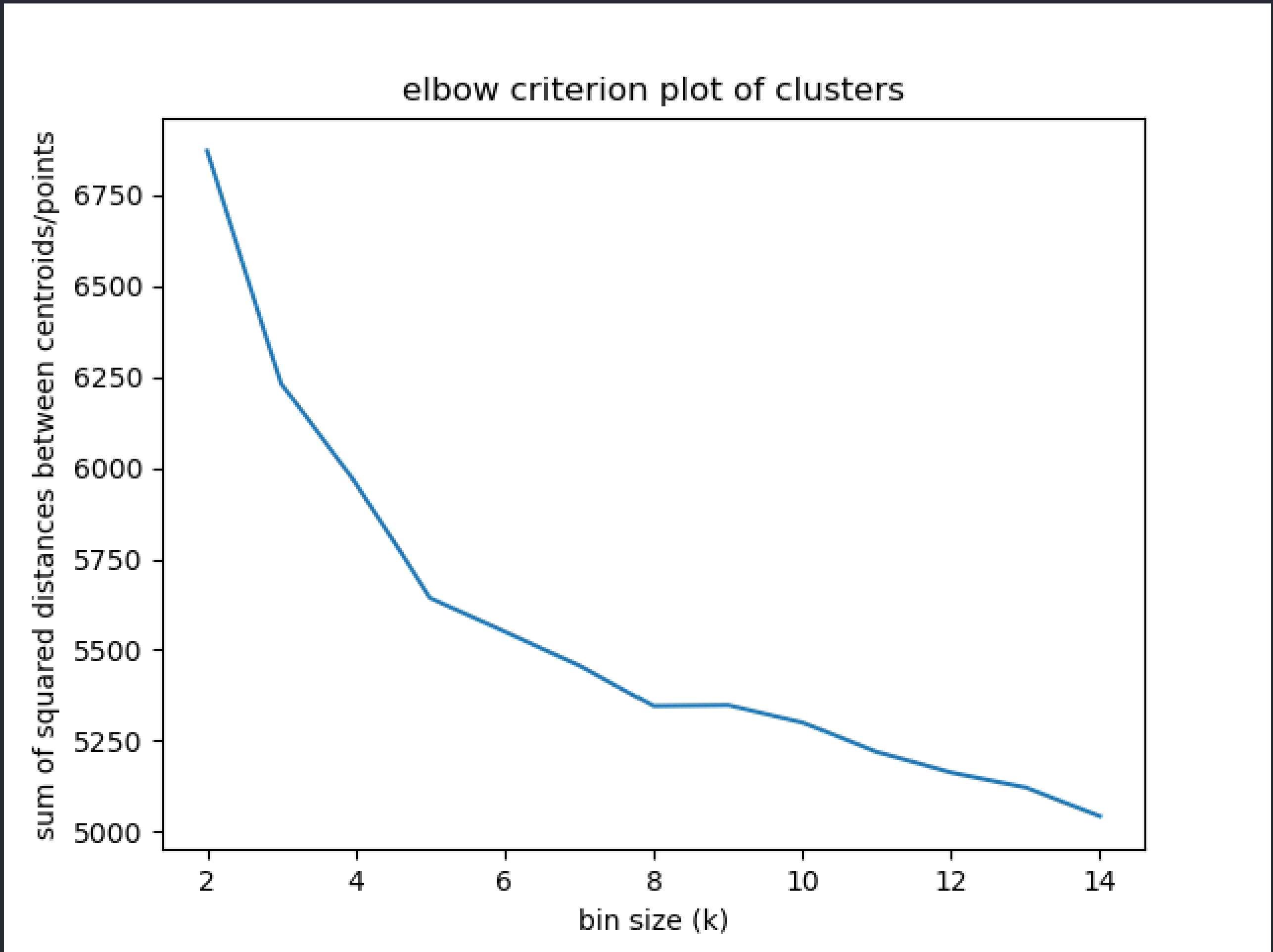
Experiment 1

Normalized n-Hot Encoding

publisher	nytimes	wsj	newsweek	...
nytimes	0	0.4	0.2	...
wsj	0.2	0	0.4	...
newsweek	0.0	0.0	0.0	...
...	...	...	...	...

Experiment 1

Elbow criterion

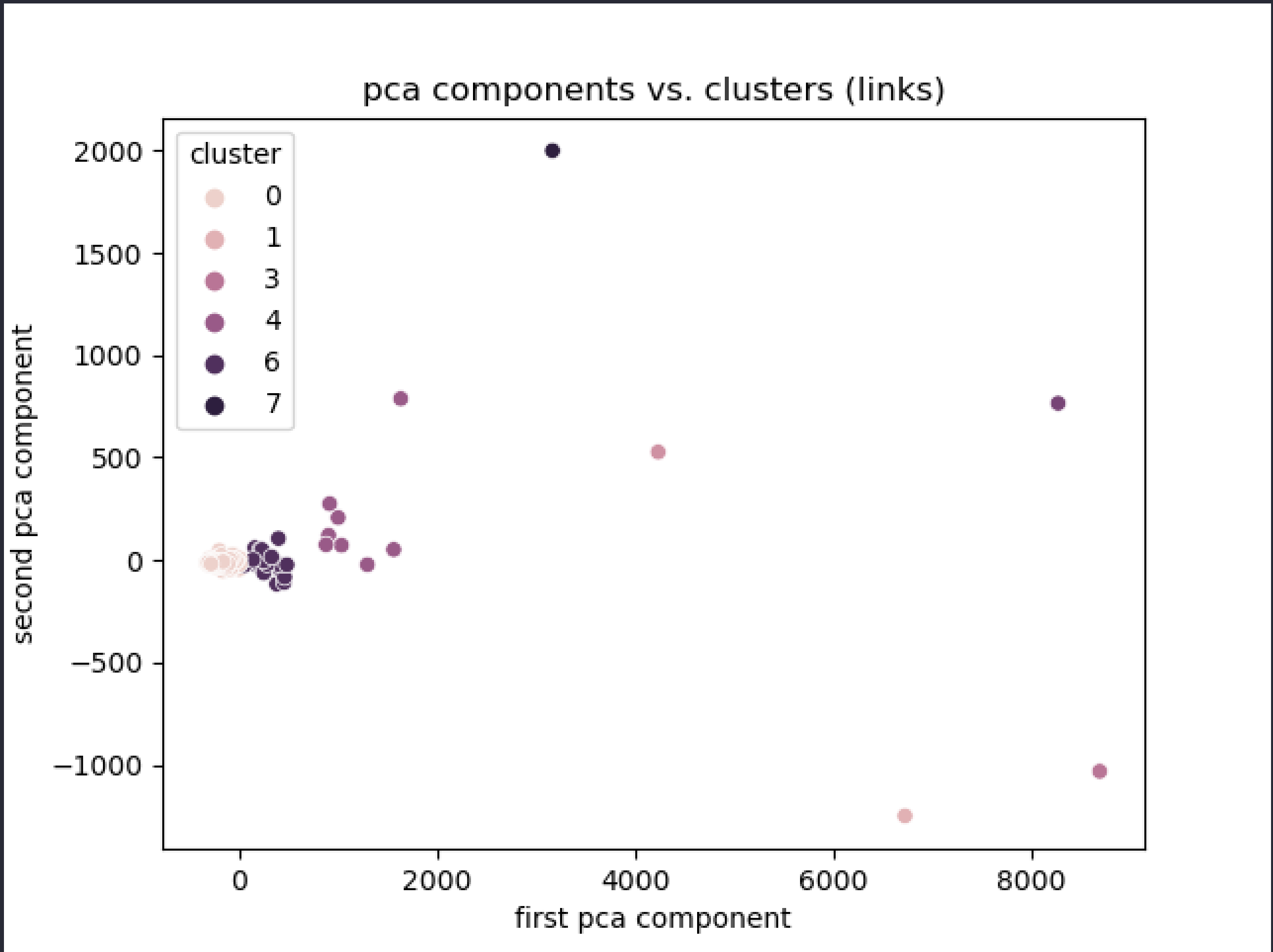


Experiment 1

Comparing encoding schemes

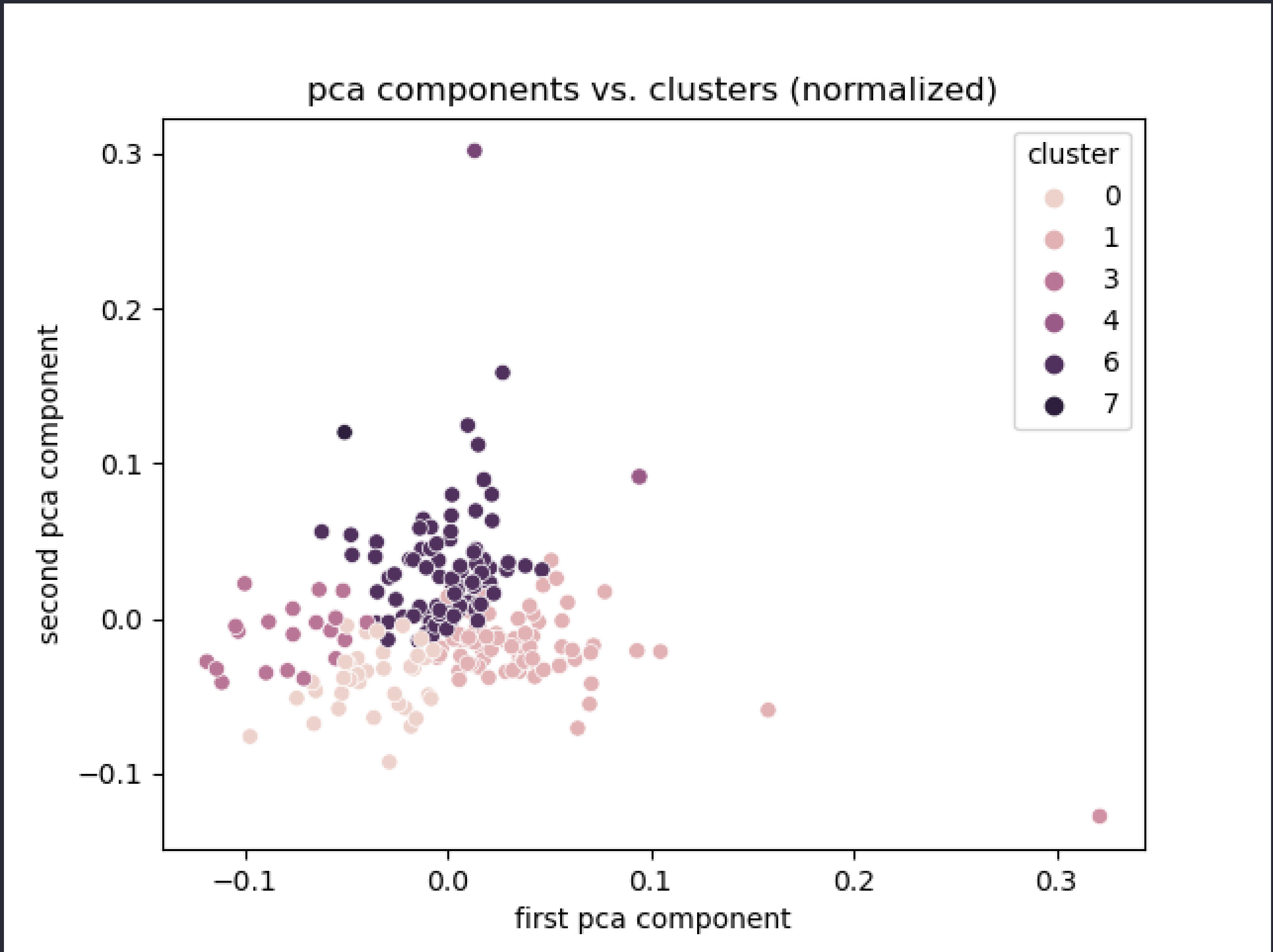
Experiment 1

Link Magnitude



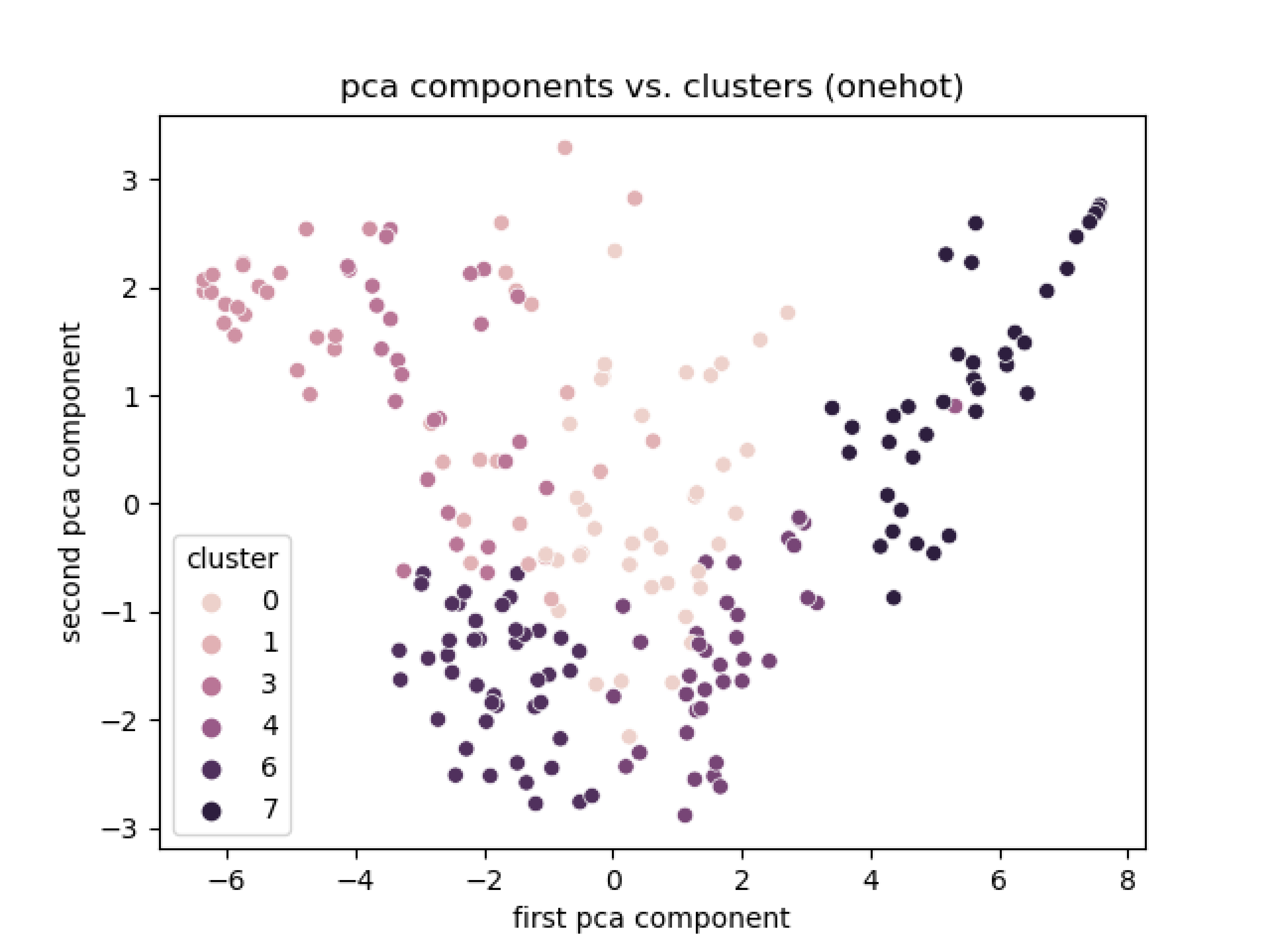
Experiment 1

Normalized



Experiment 1

One-Hot



Experiment 1

Discussion

Experiment 1

Discussion

- One-hot seems to reflect the right features.

Experiment 1

Discussion

- One-hot seems to reflect the right features.
- Found clusters, but meaning is arbitrary.
  - map to PCA results nicely.

Experiment 1

Discussion

- One-hot seems to reflect the right features.
- Found clusters, but meaning is arbitrary.
  - map to PCA results nicely.
- Limitation: need the link encoding to cluster.
  - Smaller publishers might not link very much.

Experiment 1

Discussion

- One-hot seems to reflect the right features.
- Found clusters, but meaning is arbitrary.
    - map to PCA results nicely.
- Limitation: need the link encoding to cluster.
    - Smaller publishers might not link very much.
- TODO: Association Rule Mining.
    - 'Basket of goods' analysis to group publishers.

Experiment 2

classification on link similarity.

Experiment 2

Setup

Experiment 2

Setup

- Create features:
  - Publisher frequency.
  - Reuse link encodings.

Experiment 2

Setup

- Create features:
  - Publisher frequency.
  - Reuse link encodings.
- Create classes:
  - Join bias classifications.

Experiment 2

Setup

- Create features:
  - Publisher frequency.
  - Reuse link encodings.
- Create classes:
  - Join bias classifications.
- Train classifier.

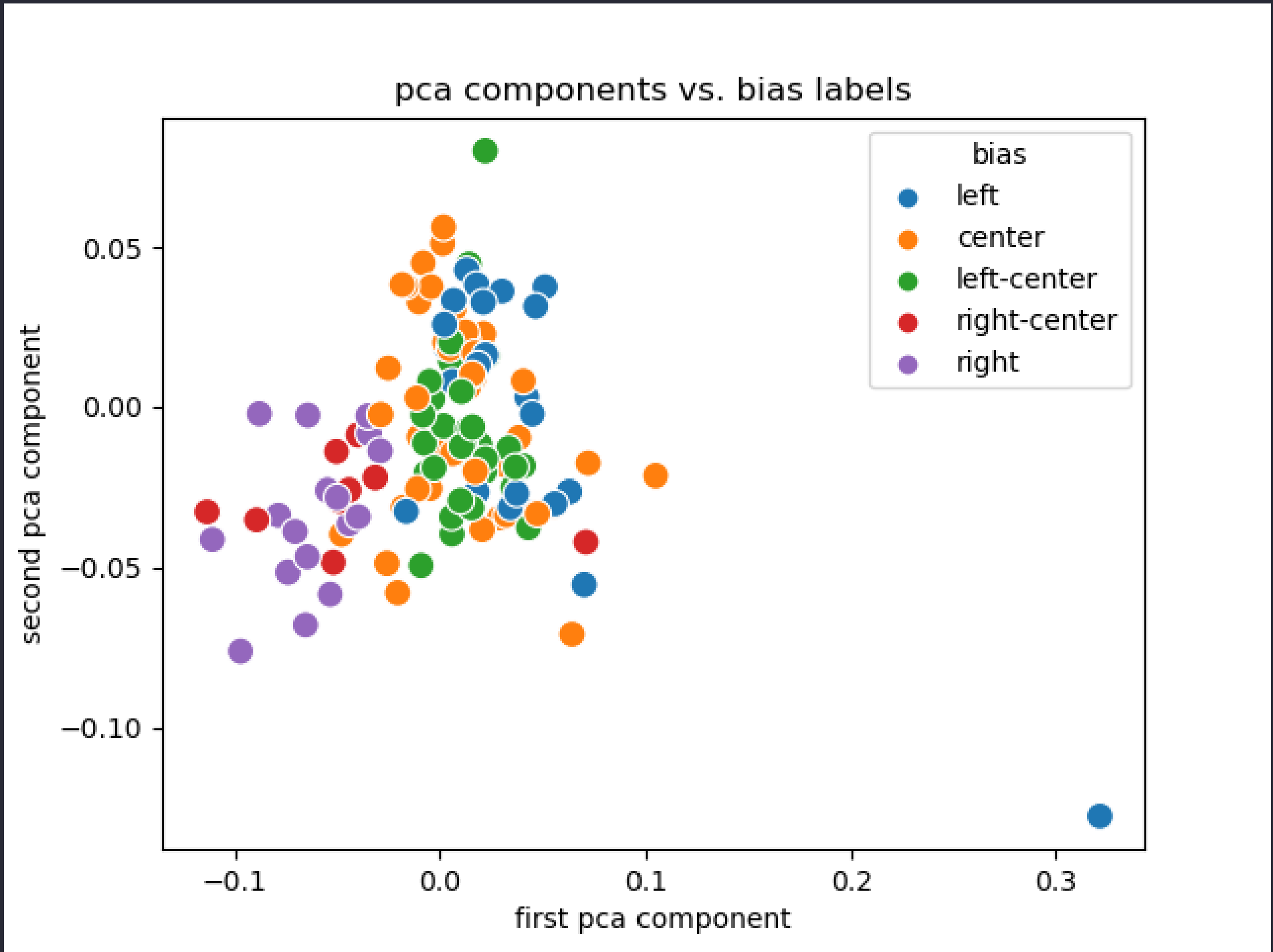
Experiment 2

Descriptive stats

metric	value
publishers	1582
labels	6
left	482
center	711
right	369
agree range	[0.0-1.0]

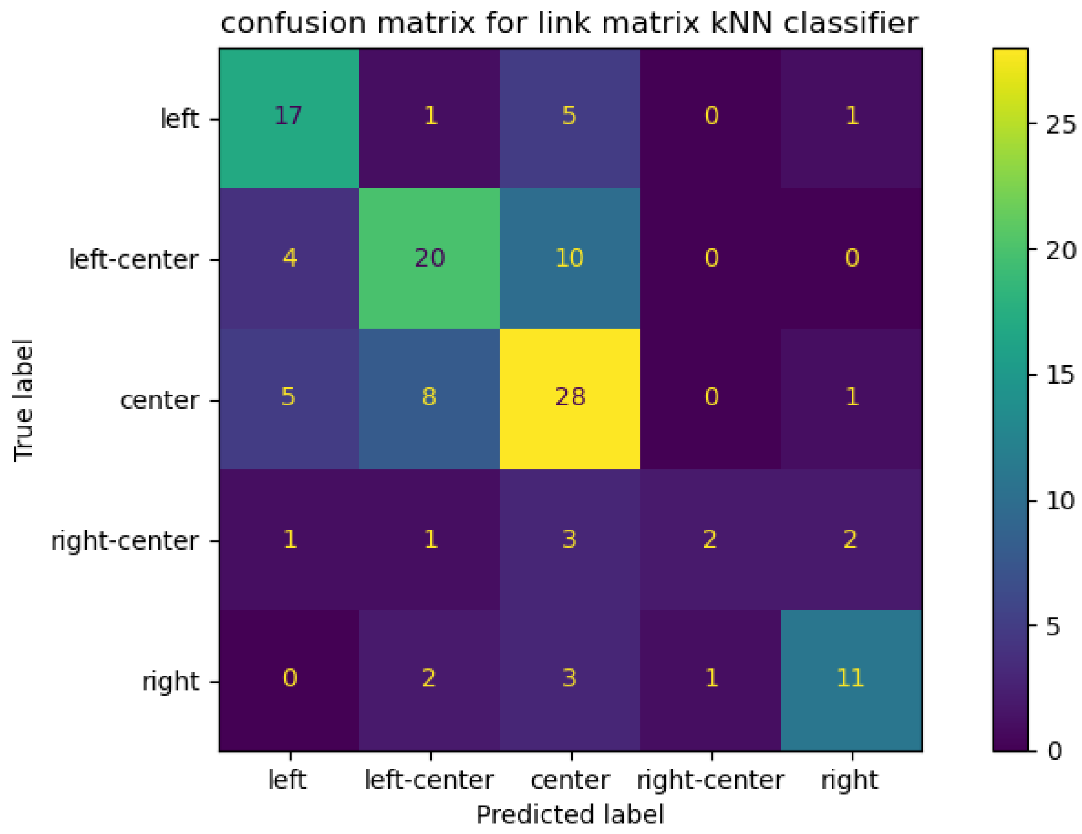
Experiment 2

Results



Experiment 2

Results



Experiment 2

Discussion

Experiment 2

Discussion

- Link encodings (and their PCA) are useful.
  - Labels are (sort of) separated and clustered.
  - Creating them for smaller publishers is trivial.

Experiment 2

Discussion

- Link encodings (and their PCA) are useful.
  - Labels are (sort of) separated and clustered.
  - Creating them for smaller publishers is trivial.
- Hot diagonal confusion matrix is good.

Experiment 2

Discussion

- Link encodings (and their PCA) are useful.
  - Labels are (sort of) separated and clustered.
  - Creating them for smaller publishers is trivial.
- Hot diagonal confusion matrix is good.
- Need to link more publisher data to get good test data.

Experiment 2

Limitations

Experiment 2

Limitations

- Dependent on accurate rating.

Experiment 2

Limitations

- Dependent on accurate rating.
- Ordinal ratings weren't available.

Experiment 2

Limitations

- Dependent on accurate rating.
- Ordinal ratings weren't available.
- Dependent on accurate joining across datasets.

Experiment 2

Limitations

- Dependent on accurate rating.
- Ordinal ratings weren't available.
- Dependent on accurate joining across datasets.
- Entire publication is rated, not authors.

Experiment 2

Limitations

- Dependent on accurate rating.
- Ordinal ratings weren't available.
- Dependent on accurate joining across datasets.
- Entire publication is rated, not authors.
- Don't know what to do with community rating.

Experiment 3

classification on sentence embedding.

Experiment 3

Setup

Experiment 3

Setup

- Generate sentence embedding for each title.

Experiment 3

Setup

- Generate sentence embedding for each title.
- Rerun PCA analysis on title embeddings.

Experiment 3

Setup

- Generate sentence embedding for each title.
- Rerun PCA analysis on title embeddings.
- Use kNN classifier to map embedding features to bias rating.

Experiment 3

Embeddings Primer

Experiment 3

Embedding Steps

1. Extract titles.
2. Tokenize titles.
3. Pick pretrained language model.
4. Generate embeddings from tokens using model.

Experiment 3

Tokens

The sentence:

"Spain, Land of 10 P.M. Dinners, Asks if It's Time to Reset Clock"

Tokenizes to:

```
['[CLS]', 'spain', ',', 'land', 'of', '10', 'p', '.', 'm', '.',
    'dinners', ',', 'asks', 'if', 'it', "'", 's', 'time', 'to',
    'reset', 'clock', '[SEP]']
```

Experiment 3

Tokens

The sentence:

"NPR/PBS NewsHour/Marist Poll Results and Analysis"

Tokenizes to:

```
['[CLS]', 'npr', '/', 'pbs', 'news', '##ho', '##ur', '/', 'maris'
    '##t', 'poll', 'results', 'and', 'analysis', '[SEP]', '[PAD]'
    '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]']
```

Experiment 3

Embeddings

- Using a BERT (Bidirectional Encoder Representations from Transformers) based model.
- Input: tokens.
- Output: dense vectors representing 'semantic meaning' of tokens.

Experiment 3

Embeddings

The tokens:

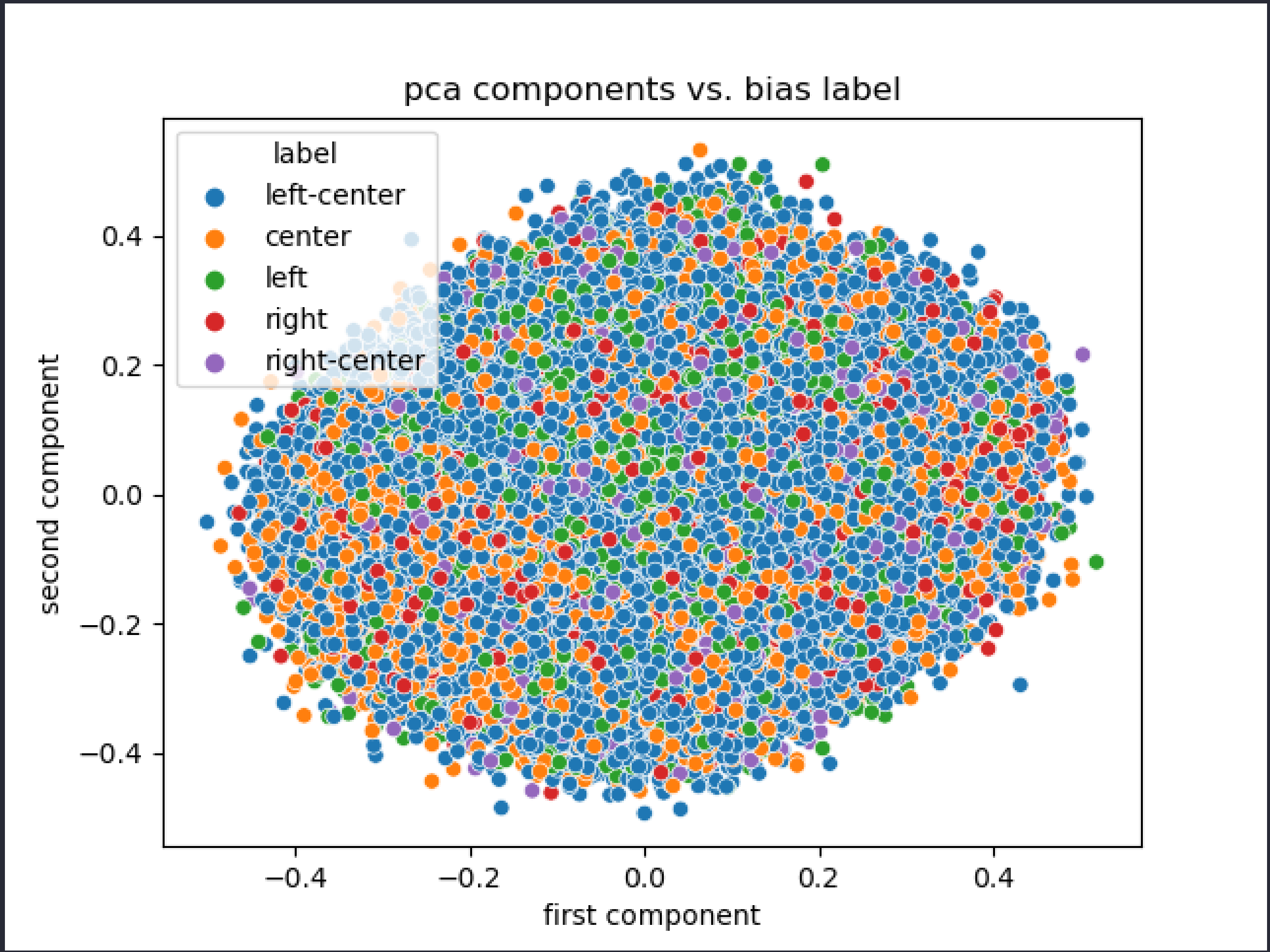
```
['[CLS]', 'npr', '/', 'pbs', 'news', '##ho', '##ur', '/', 'maris'
    '##t', 'poll', 'results', 'and', 'analysis', '[SEP]', '[PAD]'
    '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]']
```

Embeds to a vector (1, 384):

```
array([[ 0.12444635, -0.05962477, -0.00127911, ...,  0.13943022,
        -0.2552534 , -0.00238779],
       [ 0.01535596, -0.05933844, -0.0099495 , ...,  0.48110735,
         0.1370568 ,  0.3285091 ],
       [ 0.2831368 , -0.4200529 ,  0.10879617, ...,  0.15663117,
        -0.29782432,  0.4289513 ],
       ...,
```

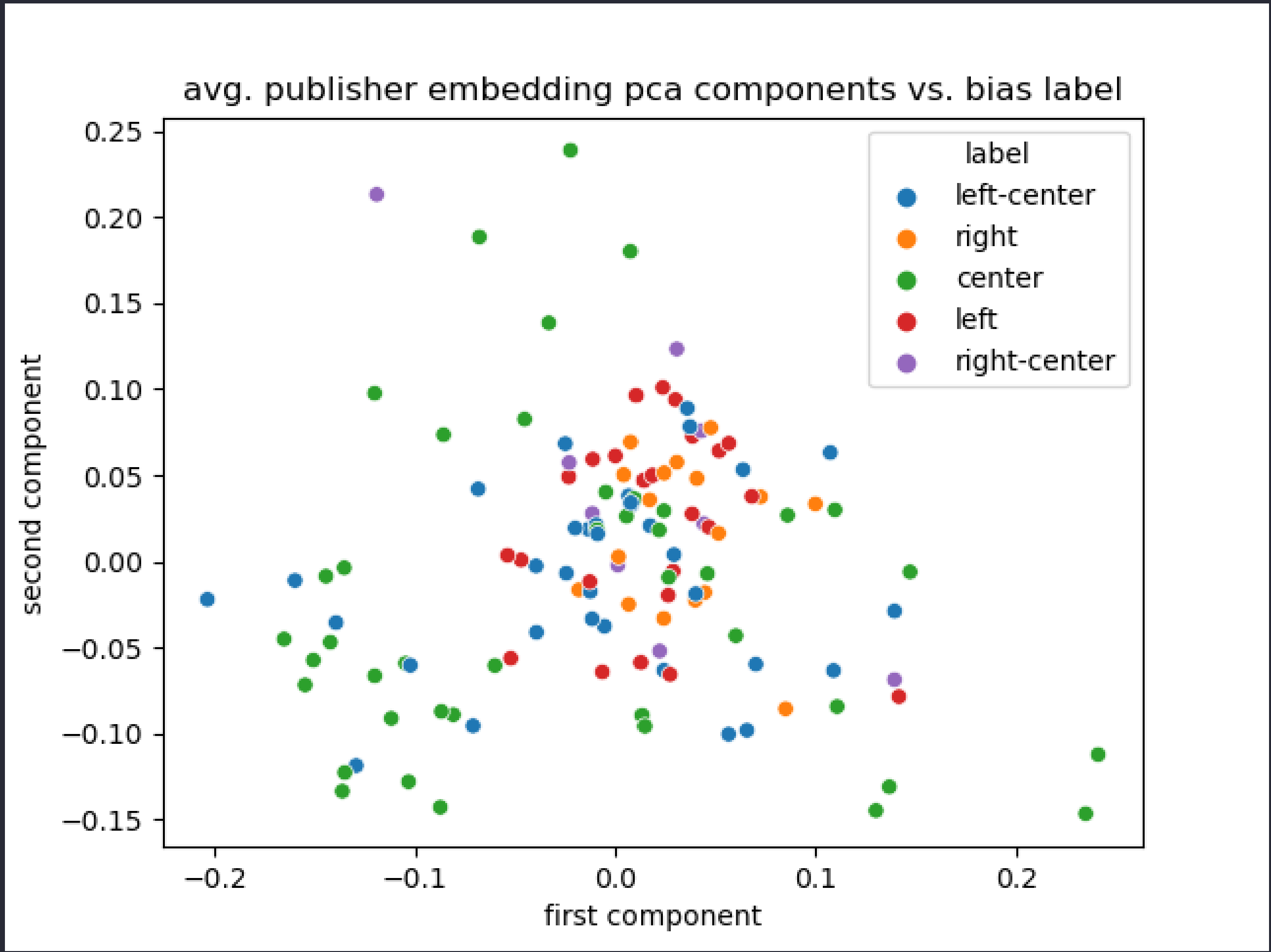
Experiment 3

Results



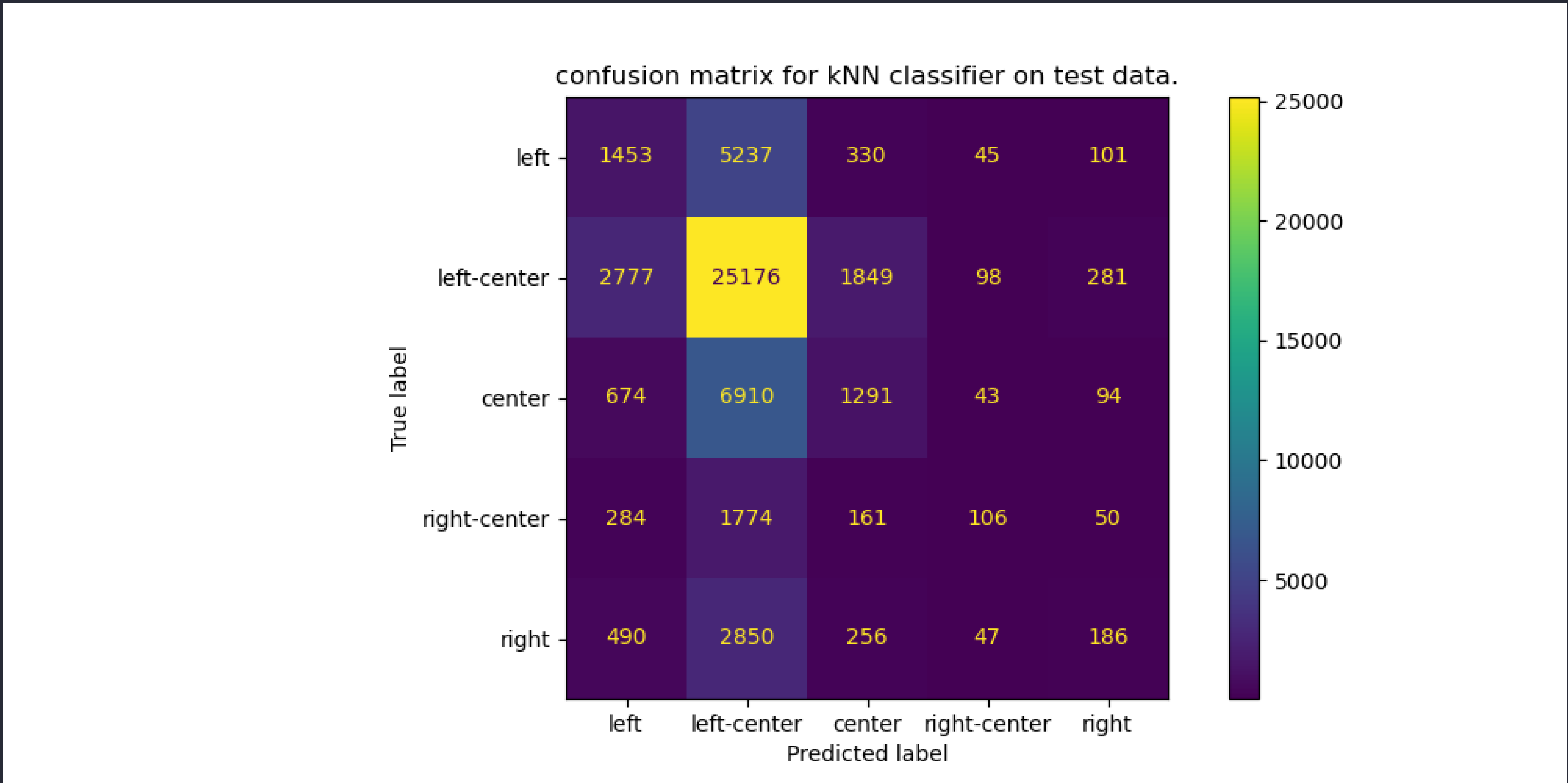
Experiment 3

Results



Experiment 3

Results



Experiment 3

Discussion

Experiment 3

Discussion

- Embedding space is hard to condense with PCA.

Experiment 3

Discussion

- Embedding space is hard to condense with PCA.
- Maybe the classifier is learning to guess 'left-ish'?

Experiment 3

Discussion

- Embedding space is hard to condense with PCA.
- Maybe the classifier is learning to guess 'left-ish'?
- Does DL work better on sparse inputs?

Experiment 4

classification on sentiment analysis.

Experiment 4

Setup

Experiment 4

Setup

- Use pretrained language classifier.

Experiment 4

Setup

- Use pretrained language classifier.
- Previously: Mapped twitter posts to tokens, to embedding, to ['positive', 'negative'] labels.

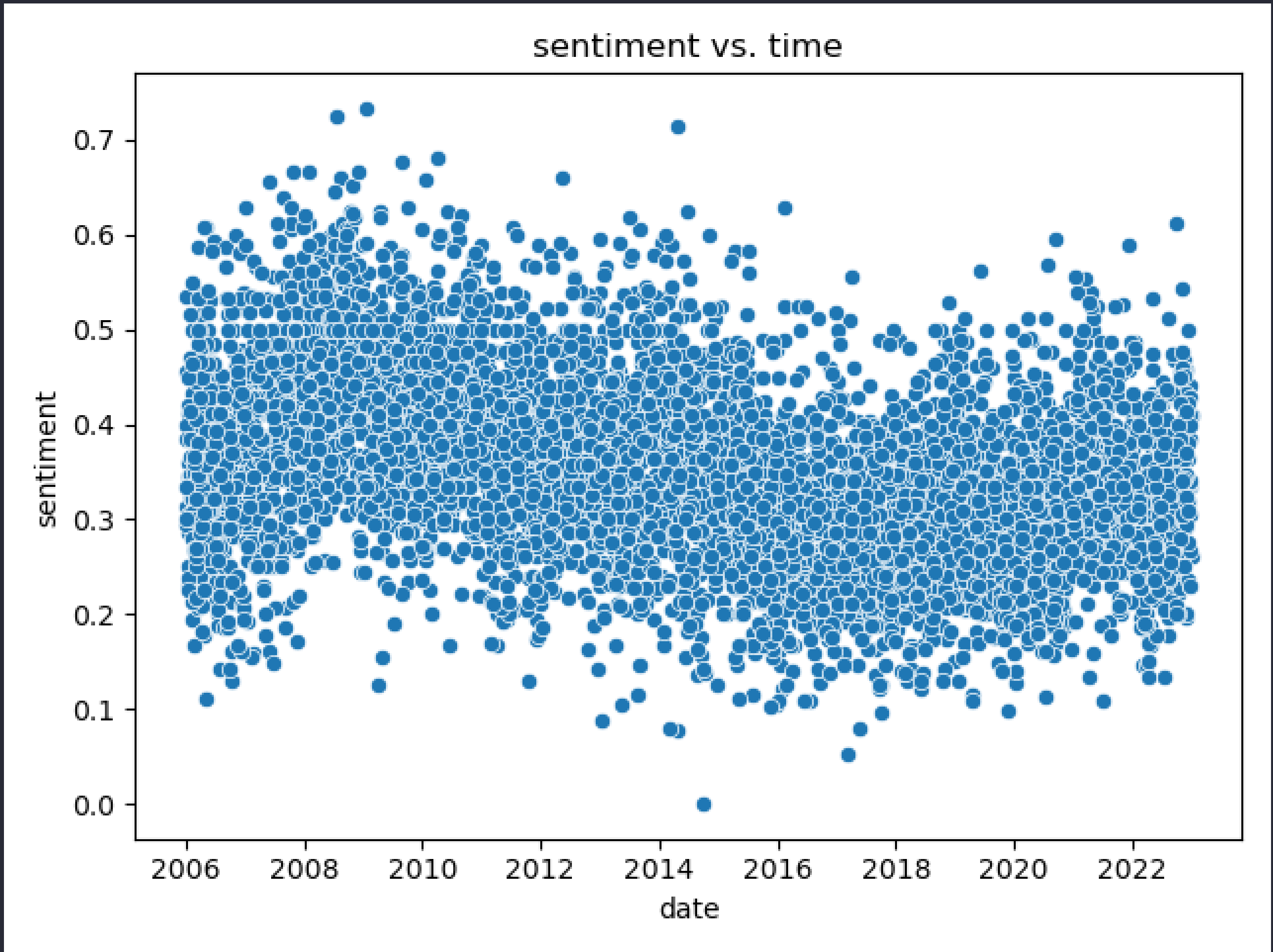
Experiment 4

Setup

- Use pretrained language classifier.
- Previously: Mapped twitter posts to tokens, to embedding, to ['positive', 'negative'] labels.
- Predict: rate of neutral titles decreasing over time.

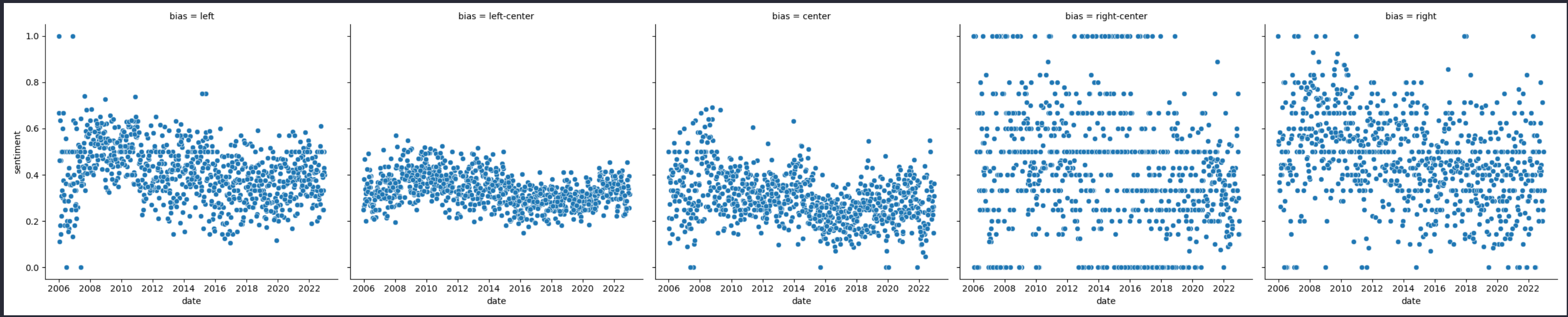
Experiment 4

Results



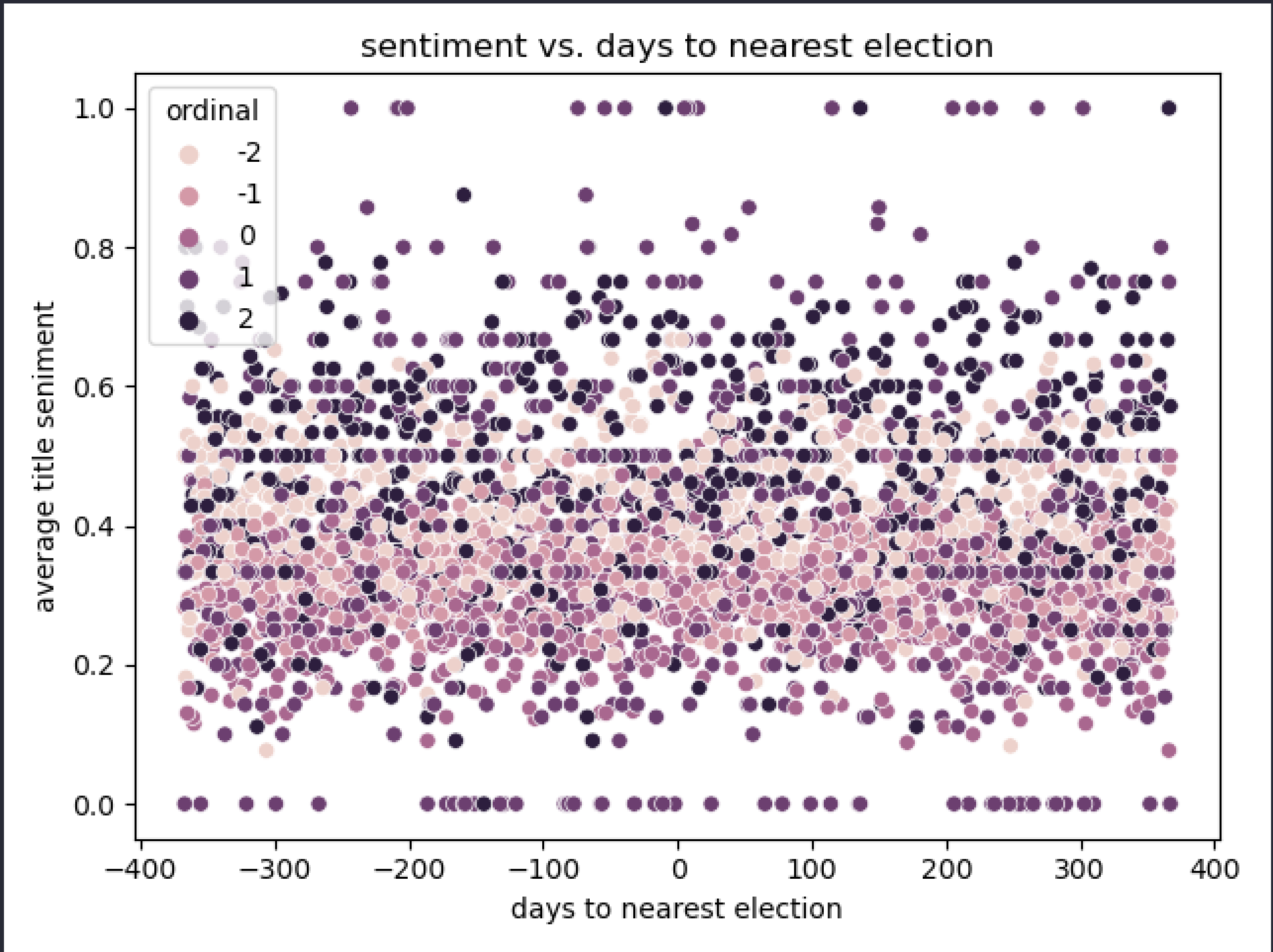
Experiment 4

Results



Experiment 4

Results



Experiment 4

Discussion

Experiment 4

Discussion

- Bump post Obama election for left and center.

Experiment 4

Discussion

- Bump post Obama election for left and center.
- Dip pre Trump election for left and center.

Experiment 4

Discussion

- Bump post Obama election for left and center.
- Dip pre Trump election for left and center.
- Right is all over the place - not enough data?

Experiment 4

Discussion

- Bump post Obama election for left and center.
- Dip pre Trump election for left and center.
- Right is all over the place - not enough data?
- Recency of election not a clear factor.

Experiment 5

regression on title emotional expression.

Experiment 5

Setup

Experiment 5

Setup

- Use pretrained language classifier.

Experiment 5

Setup

- Use pretrained language classifier.
- Previously: Mapped reddit posts to tokens, to embedding, to emotion labels.

Experiment 5

Setup

- Use pretrained language classifier.
- Previously: Mapped reddit posts to tokens, to embedding, to emotion labels.
- Predict: rate of neutral titles decreasing over time.

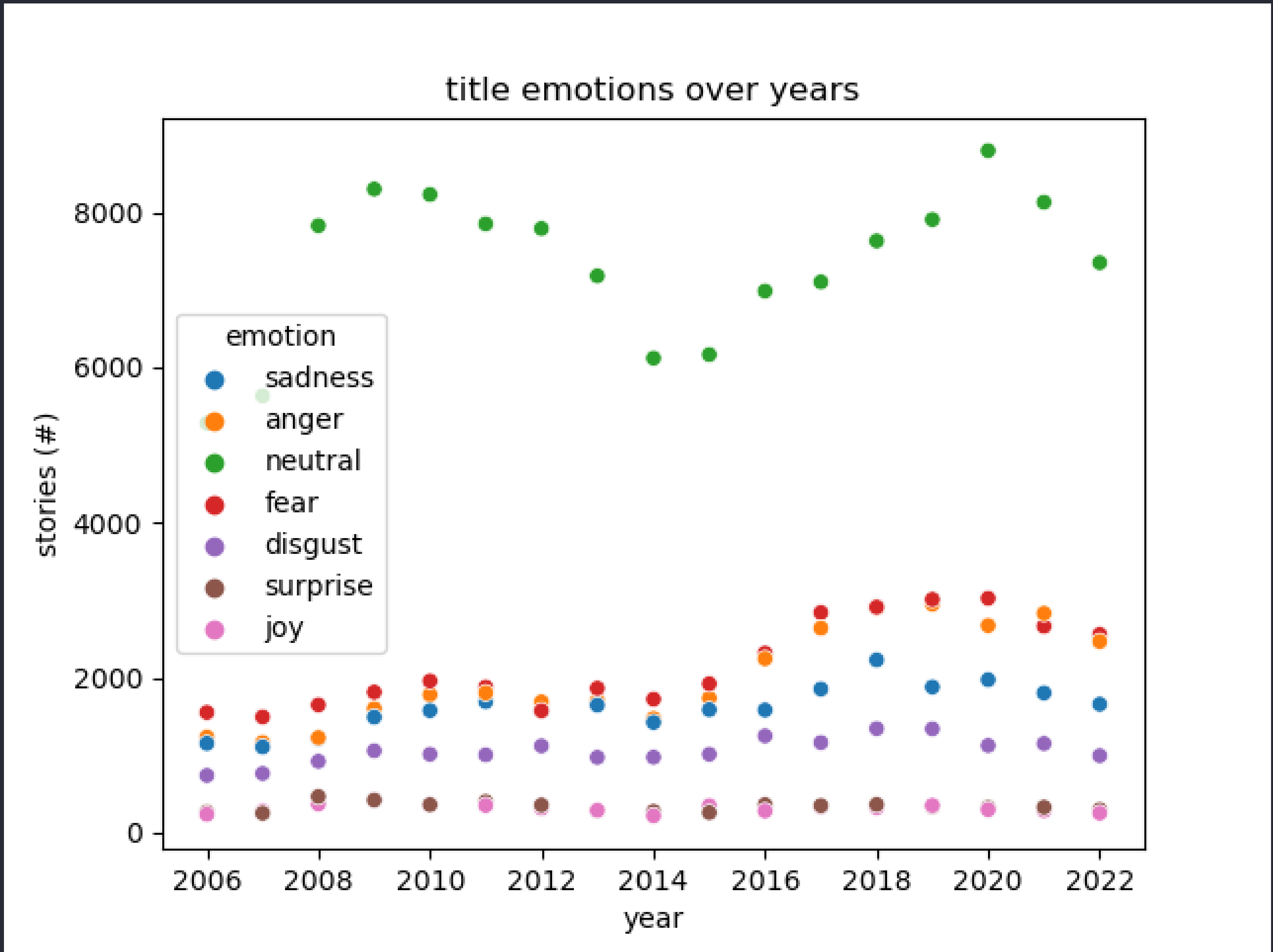
Experiment 5

Setup

- Use pretrained language classifier.
- Previously: Mapped reddit posts to tokens, to embedding, to emotion labels.
- Predict: rate of neutral titles decreasing over time.
- Classify:
  - features: emotional labels
  - labels: bias

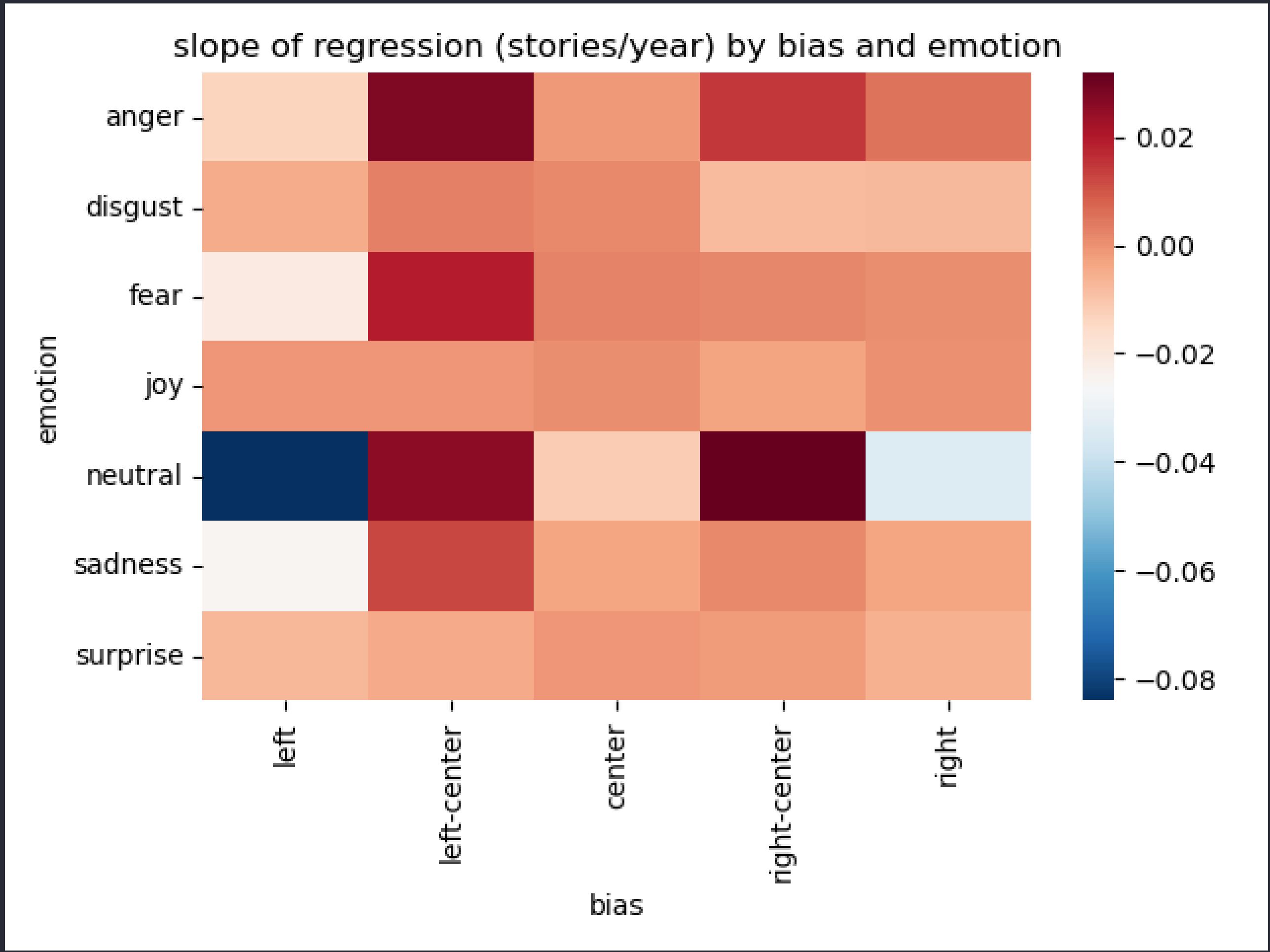
Experiment 5

Results



Experiment 5

Results



Experiment 5

Discussion

Experiment 5

Discussion

- Neutral story titles dominate the dataset.

Experiment 5

Discussion

- Neutral story titles dominate the dataset.
- Increase in stories published might explain most of the trend.

Experiment 5

Discussion

- Neutral story titles dominate the dataset.
- Increase in stories published might explain most of the trend.
- Far-right and far-left both became less neutral.

Experiment 5

Discussion

- Neutral story titles dominate the dataset.
- Increase in stories published might explain most of the trend.
- Far-right and far-left both became less neutral.
- Left-Center and right-center became more emotional, but also neutral.

Experiment 5

Discussion

- Neutral story titles dominate the dataset.
- Increase in stories published might explain most of the trend.
- Far-right and far-left both became less neutral.
- Left-Center and right-center became more emotional, but also neutral.
- Not a lot of movement overall.

Conclusion

Hypothesis

- The polarization is not evenly distributed across publishers. **unproven**
- The polarization is not evenly distributed across political specturm. **unproven**
- The polarization increases near elections. **false**
- Similarly polarized publishers link to each other. **sorta**
- 'Mainstream' media uses more neutral titles. **true**
- Highly polarized publications don't last as long. **untested**

Conclusion

Conclusion

- Article titles do not have a lot of predictive power.

Conclusion

- Article titles do not have a lot of predictive power.
- Mainstream, neutral publications dominate the dataset.

Conclusion

- Article titles do not have a lot of predictive power.
- Mainstream, neutral publications dominate the dataset.
- Link frequency, sentence embeddings, and sentiments are useful features.

Conclusion

- Article titles do not have a lot of predictive power.
- Mainstream, neutral publications dominate the dataset.
- Link frequency, sentence embeddings, and sentiments are useful features.
- A few questions remain.

Questions

References

[1]: Stewart, A.J. et al. 2020. Polarization under rising inequality and economic decline. Science Advances. 6, 50 (Dec. 2020), eabd4201. DOI:https://doi.org/10.1126/sciadv.abd4201.