The speech we miss: How keyword-based data collection obscures youth participation in online political discourse

Sarah Shugars
they/them/theirs
Assistant Professor, Rutgers University

RUTGERS

Digisurvor workshop
University of Manchester
13 January 2026

Detecting and correcting bias in linked data sources

The speech we miss: How Keyword-Based Data Collection Obscures Youth Participation in Online Political Discourse

Adina Gitomer, Sarah Shugars, Ryan J. Gallagher, Stefan McCabe, Brooke Foucault Welles

Computational Communication Research, 5(1). 2023.




RIP Twitter API 🪦

```
url = 'https://api.twitter.com

response = requests.request(
    "GET", url,
    params= {
    query = 'Trump OR election OR fascism'
    })
```

List of "political" keywords

Capturing "Political" Speech

Keyword search implicitly assumes
political content looks like:

Trump lost the election!

Capturing “Political” Speech

Keyword search implicitly assumes political content looks like:

Trump lost the election!

What political speech actually looks like

Capturing "Political" Speech

Keyword classifier

Trump lost the **election**!

Yes, election-related

Not election-related

The Speech We Miss

Trump lost the **election**!

Keyword classifier

Yes, election-related

Not election-related

RQ1:
How big of a problem is this?

The Speech We Miss

Trump lost the election!

Keyword classifier

Yes, election-related

Not election-related

RQ1:
How big of a problem is this?

RQ2:
Is there variation by **age**?

Data Matching

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records

- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)
  - ➡ Match users to voting records:

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)
  - ➡ Match users to voting records:
    - Name + (city, state) must be unique match

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)
  - ➡ Match users to voting records:
    - Name + (city, state) must be unique match

Match if I'm the **only** "Sarah Shugars" registered to vote in "New Brunswick, NJ"

Hughes et al. (2020)
Shugars et al. (2021)

Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)
  - ➡ Match users to voting records:
    - Name + (city, state) must be unique match
- Demographically representative of Twitter users overall

Match if I'm the **only** "Sarah Shugars" registered to vote in "New Brunswick, NJ"

Hughes et al. (2020)
Shugars et al. (2021)

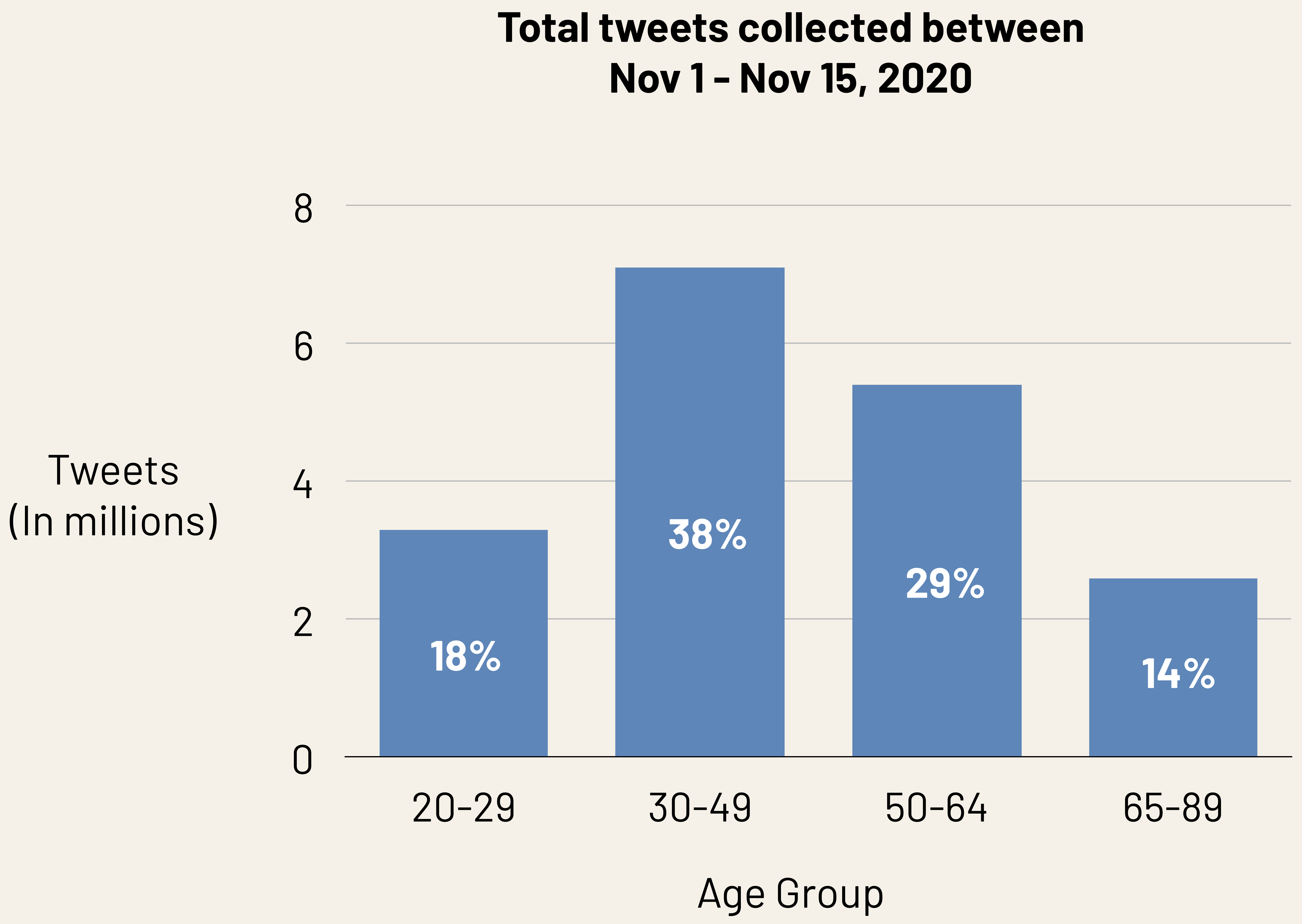
Data Matching

- Panel of 1.6 million Twitter users matched to US voting records
- Matching procedure:
  - ➡ Collect Twitter profiles from 10% sample (2014-2017)
  - ➡ Match users to voting records:
    - Name + (city, state) must be unique match
- Demographically representative of Twitter users overall
  - Youngest users were 17 in 2017

Match if I'm the **only** "Sarah Shugars" registered to vote in "New Brunswick, NJ"

Hughes et al. (2020)
Shugars et al. (2021)

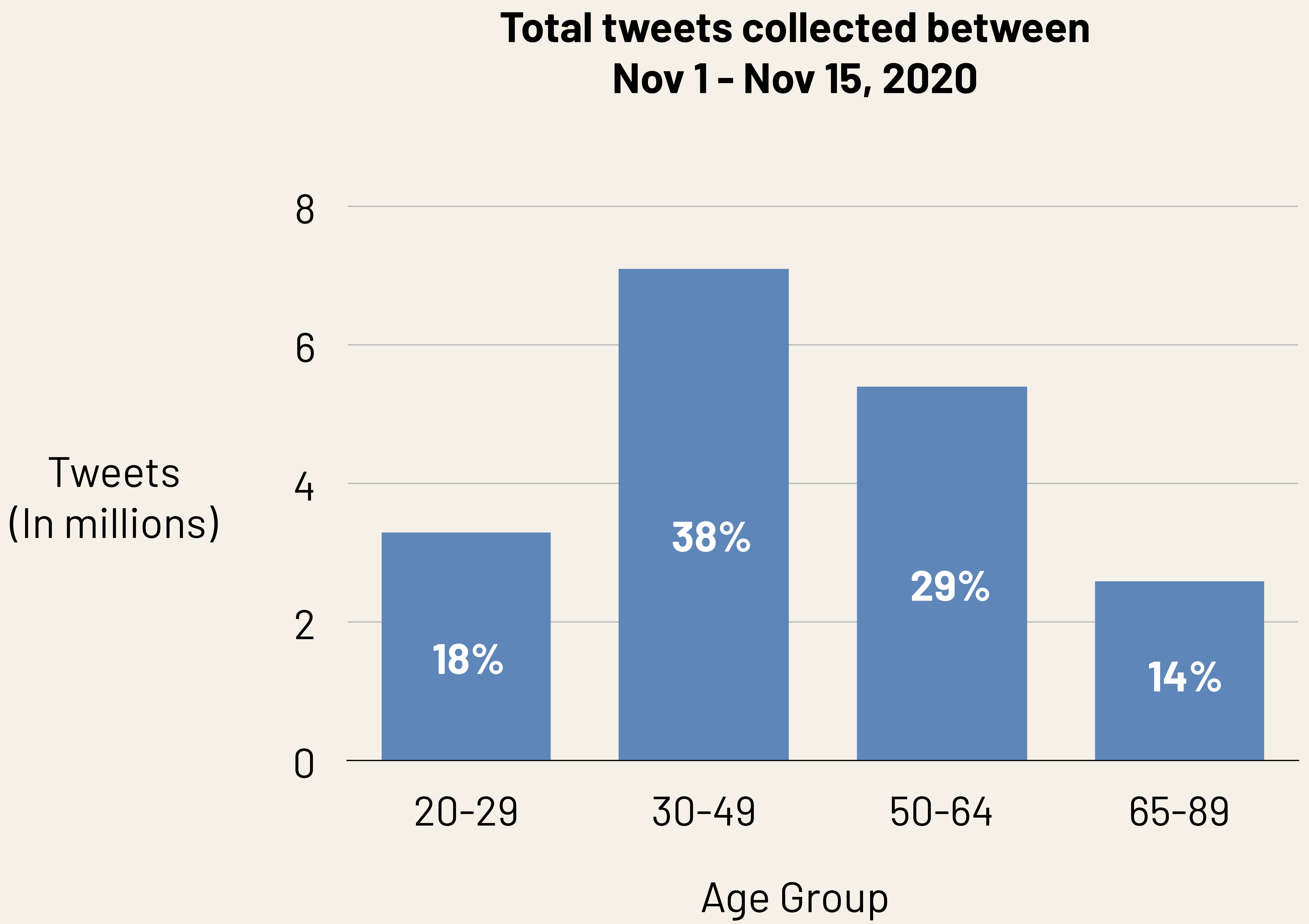
Data



Total tweets collected between
Nov 1 - Nov 15, 2020
8
6
4
2
0
Tweets
(In millions)
18%
38%
29%
14%
20-29
30-49
50-64
65-89
Age Group

Data

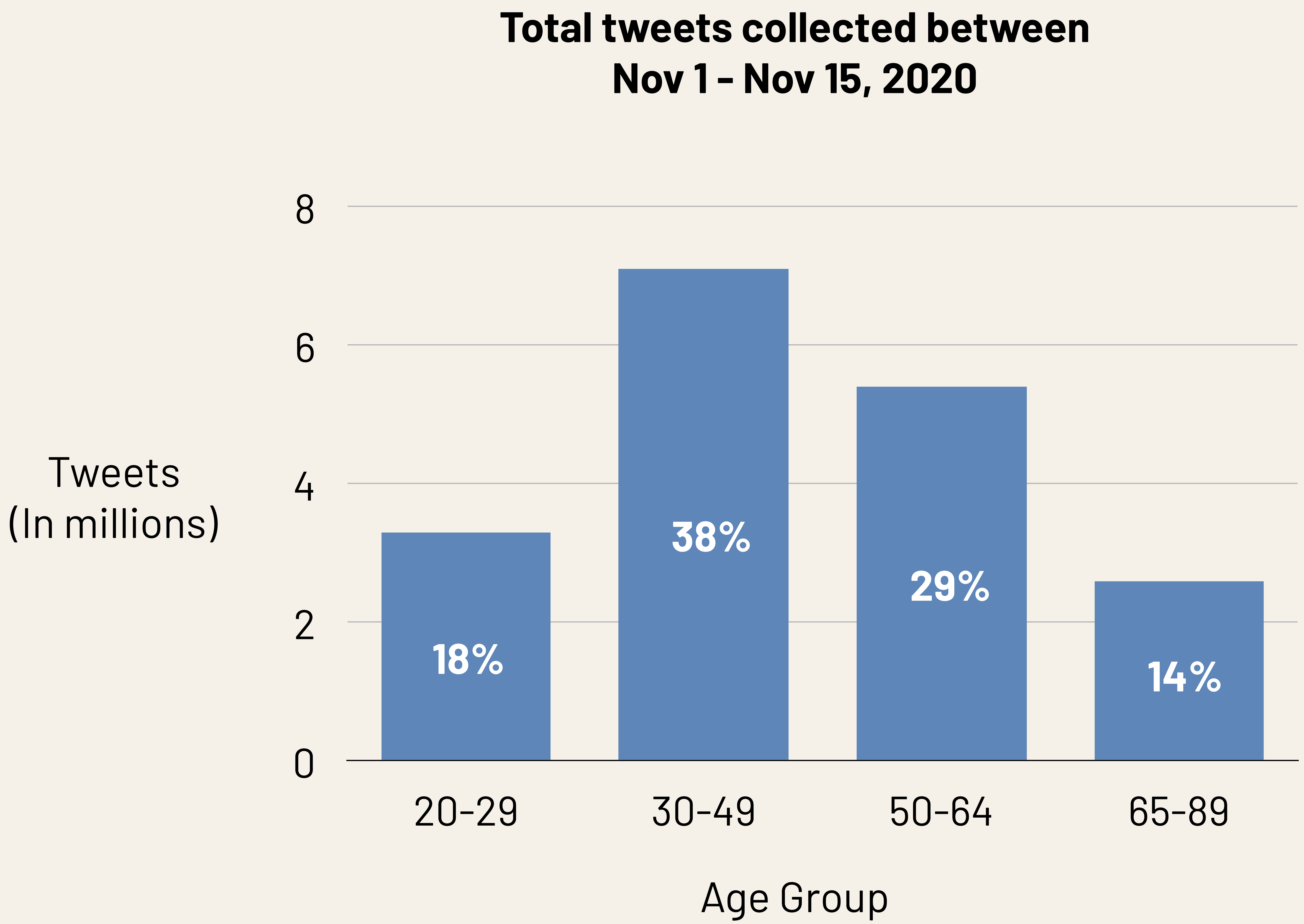
We collect <u>all</u> posts made by panelists between Nov 1 - Nov 15 2020



Data

We collect <u>all</u> posts made by panelists between Nov 1 - Nov 15 2020

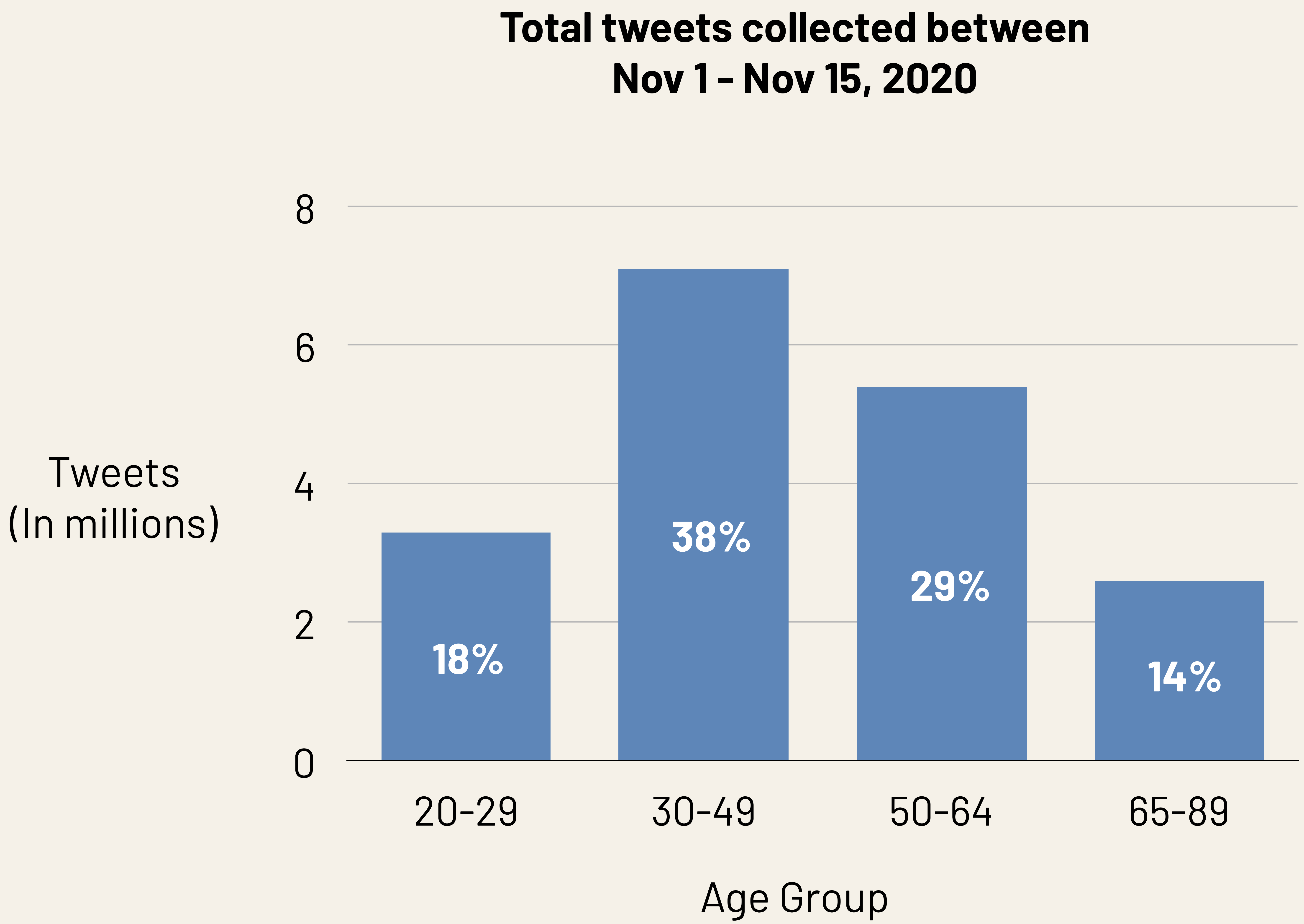
- ➡ Which tweets would be retrieved using keyword-based search?



Data

We collect <u>all</u> posts made by panelists between Nov 1 - Nov 15 2020

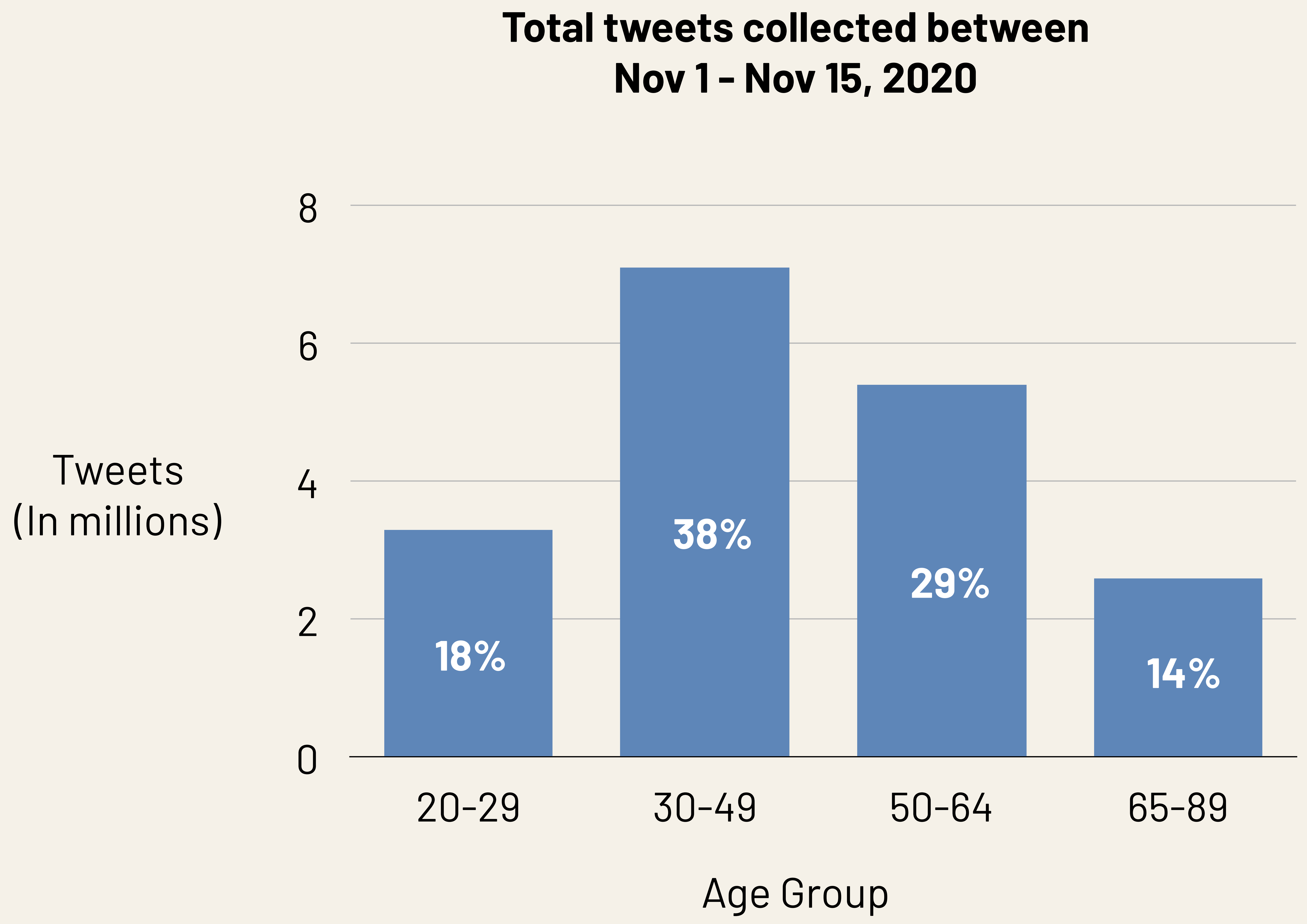
- ➡ Which tweets would be retrieved using keyword-based search?
- ➡ Which tweets are actually election-related? (Via handcoding)



Data

We collect <u>all</u> posts made by panelists between
Nov 1 - Nov 15 2020

- ➡ Which tweets would be retrieved using keyword-based search?
- ➡ Which tweets are actually election-related? (Via handcoding)
- ➡ Stratify sample by age



Method

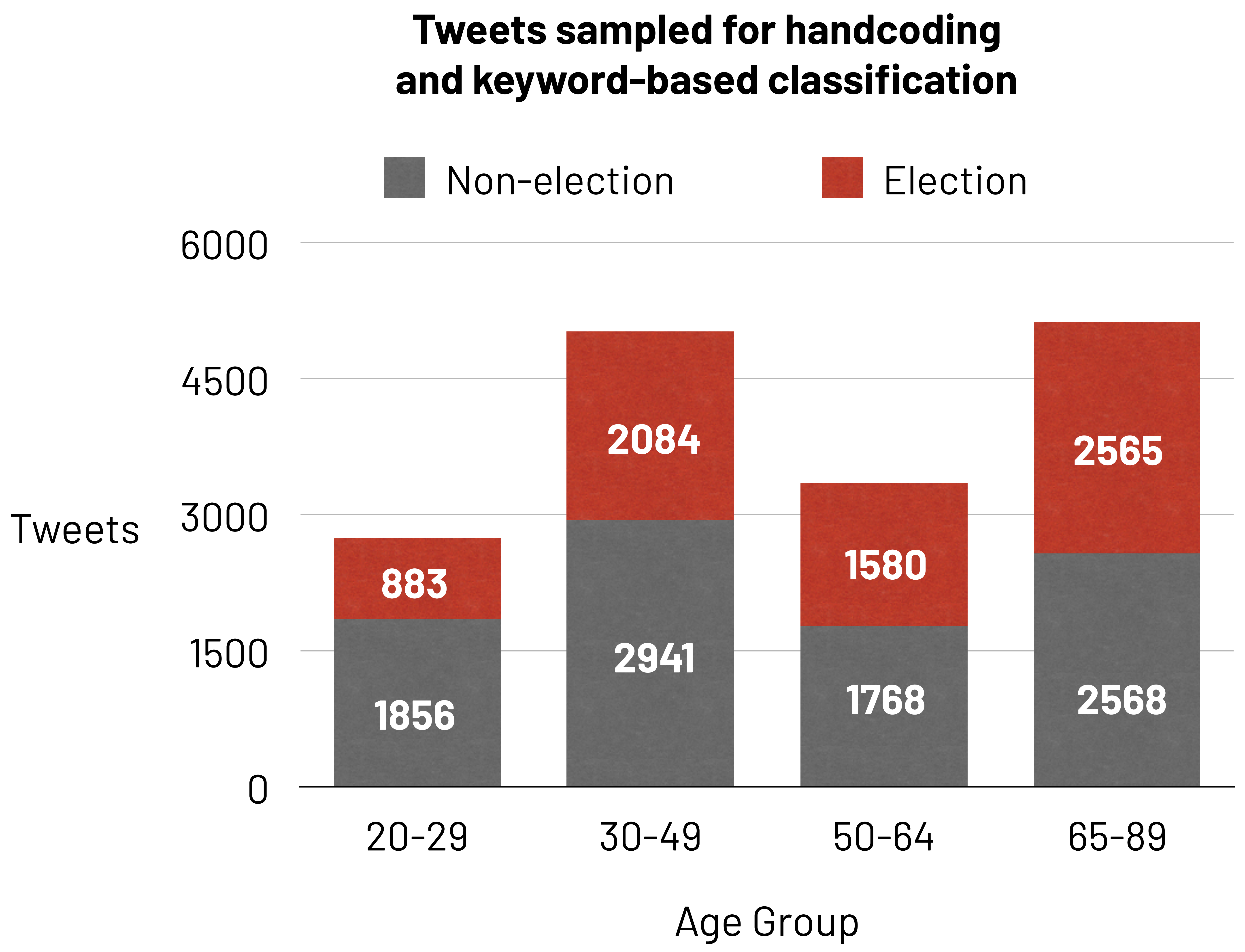
Step 1: Use an initial keyword classifier to take an informed sample for hand coding

Step 2: Hand code 16,245 tweets for election relevance

Method

Step 1: Use an initial keyword classifier to take an informed sample for hand coding

Step 2: Hand code 16,245 tweets for election relevance



Keyword Classifier Accuracy

How frequently did the keyword classifier disagree with our handcoding?

Keyword Classifier Accuracy

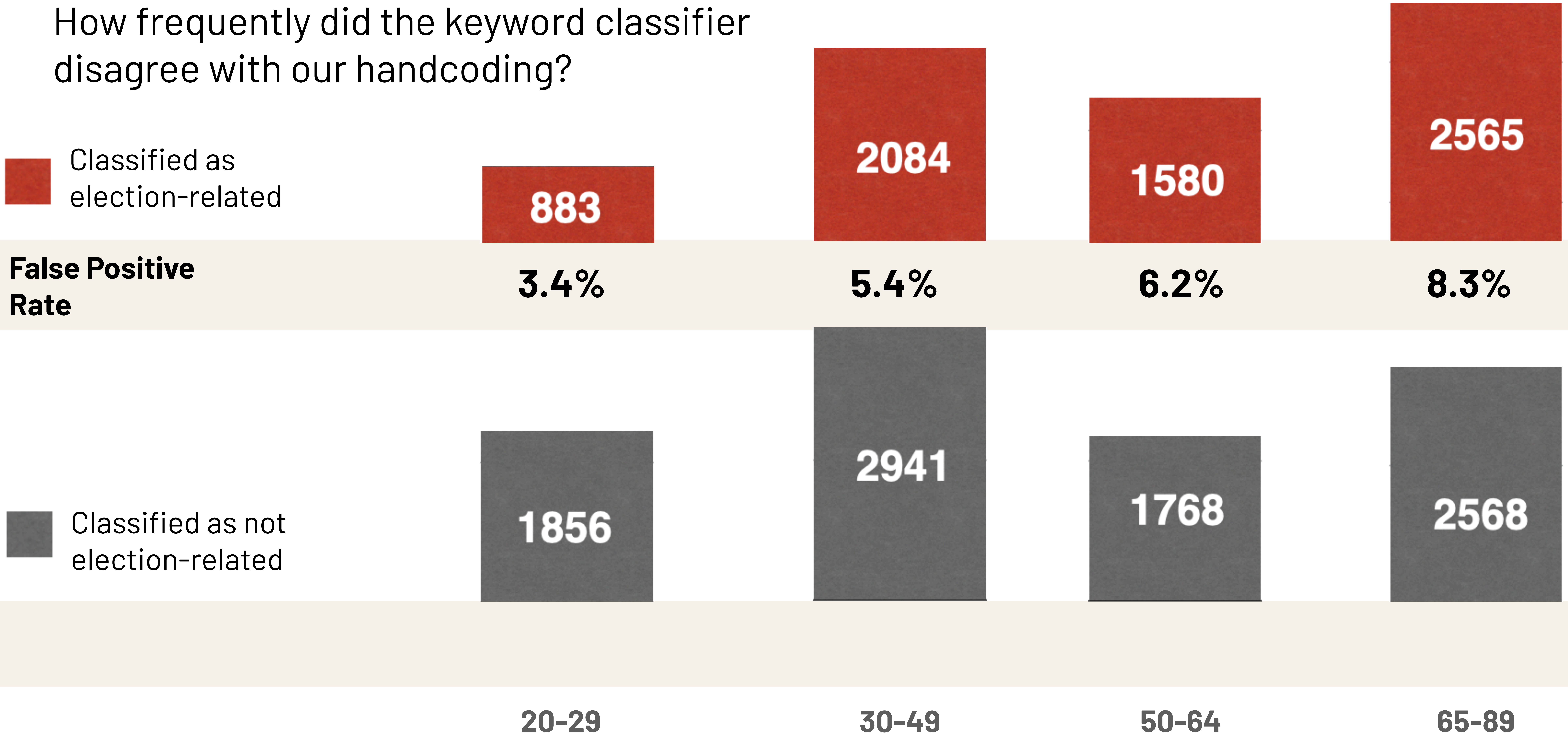
How frequently did the keyword classifier disagree with our handcoding?

Keyword Classifier Accuracy

How frequently did the keyword classifier disagree with our handcoding?

Keyword Classifier Accuracy

How frequently did the keyword classifier disagree with our handcoding?



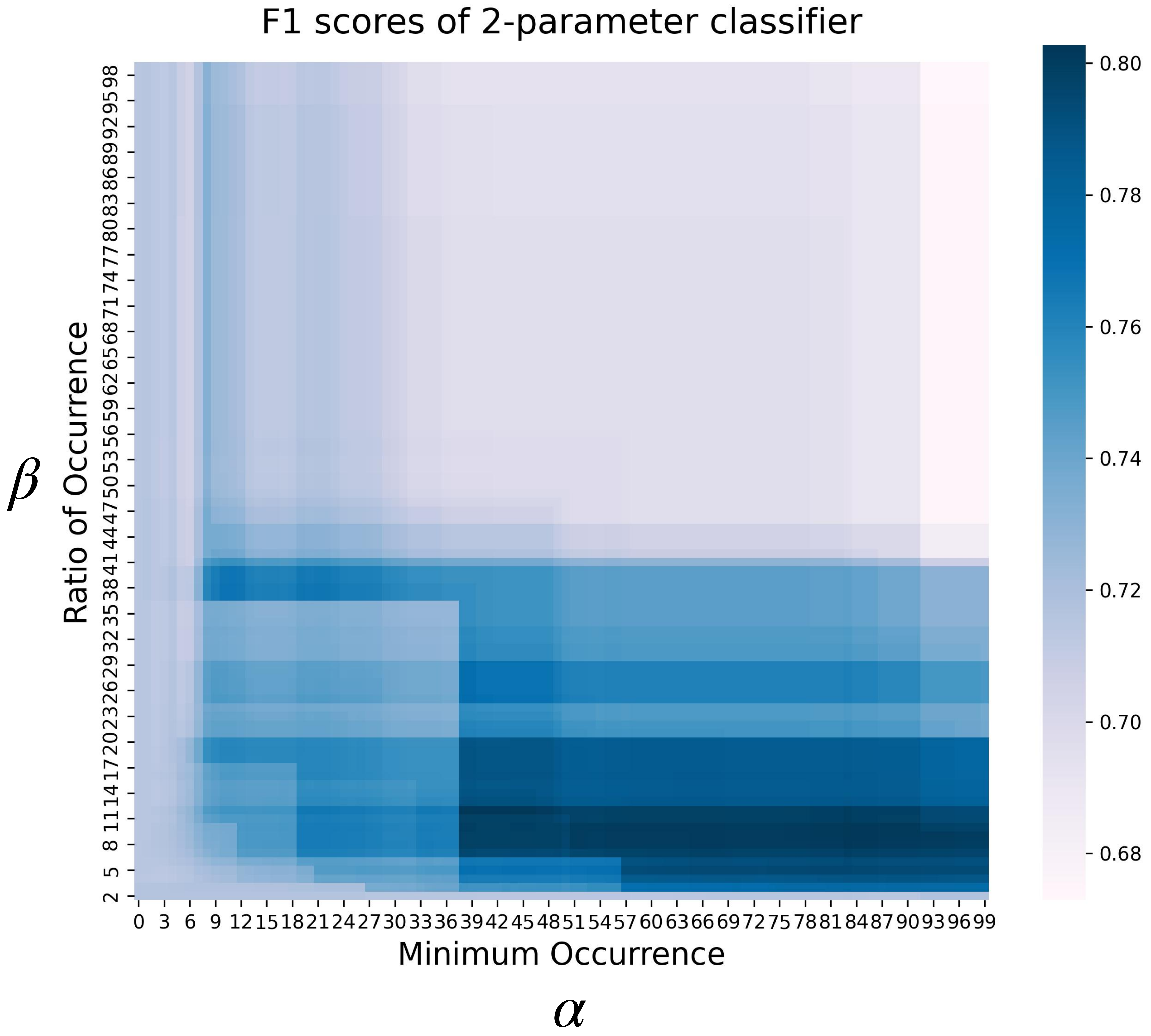
Keyword Classifier Accuracy

How frequently did the keyword classifier disagree with our handcoding?

Keyword Classifier Accuracy

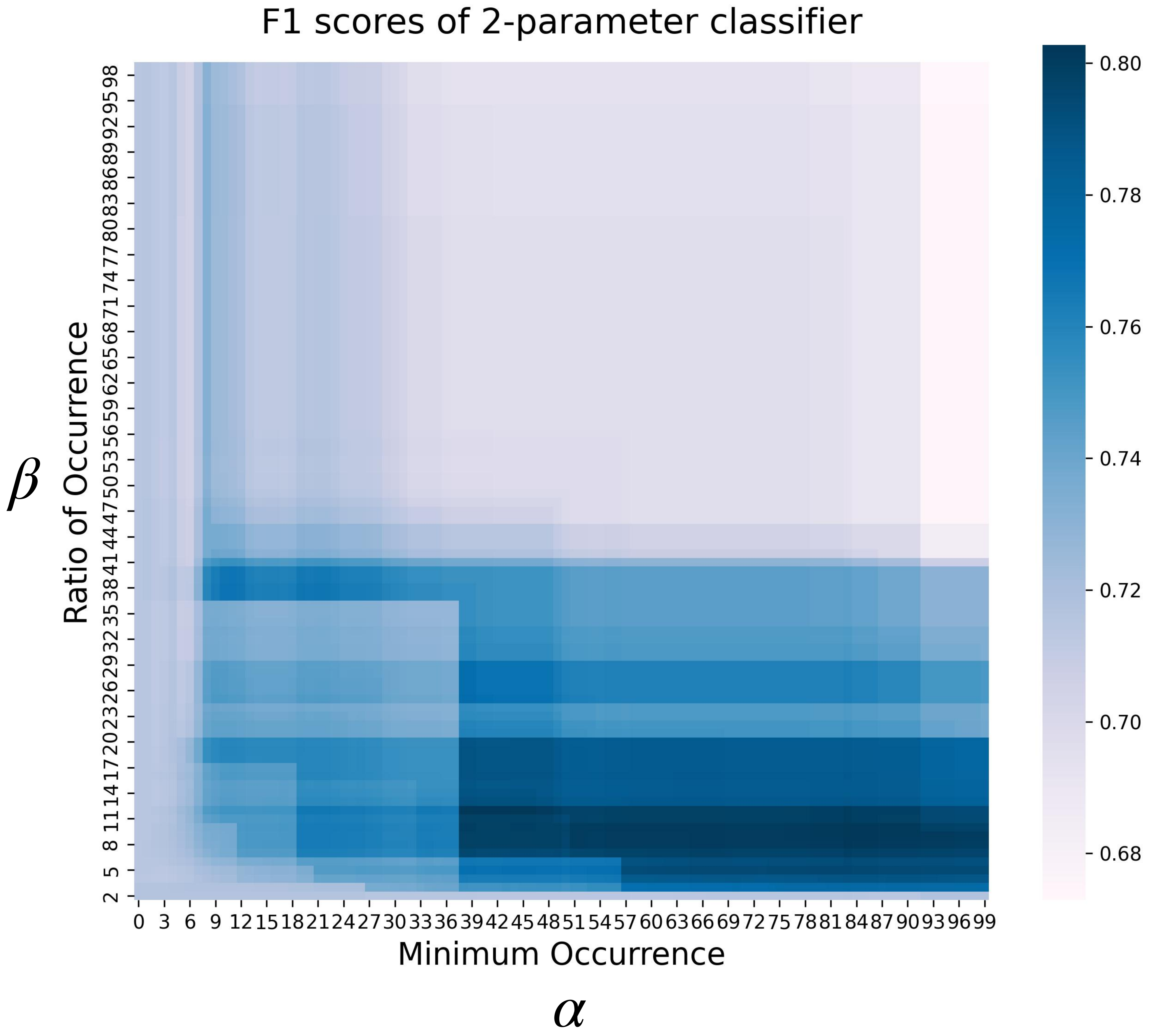
How frequently did the keyword classifier disagree with our handcoding?

Could we build a better keyword classifier?



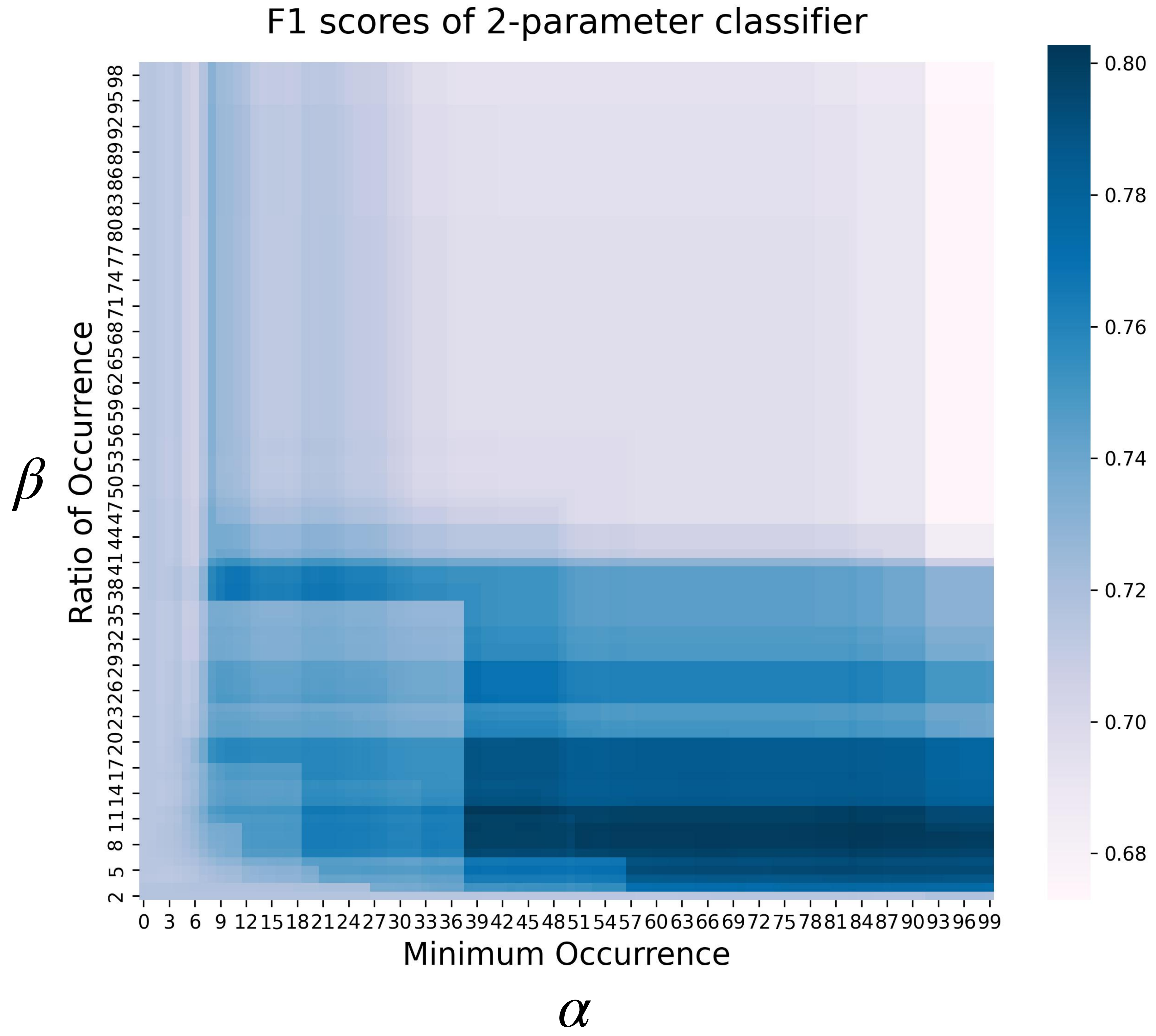
Could we build a better keyword classifier?

- Test 10,000 potential classifiers through a two-parameter model



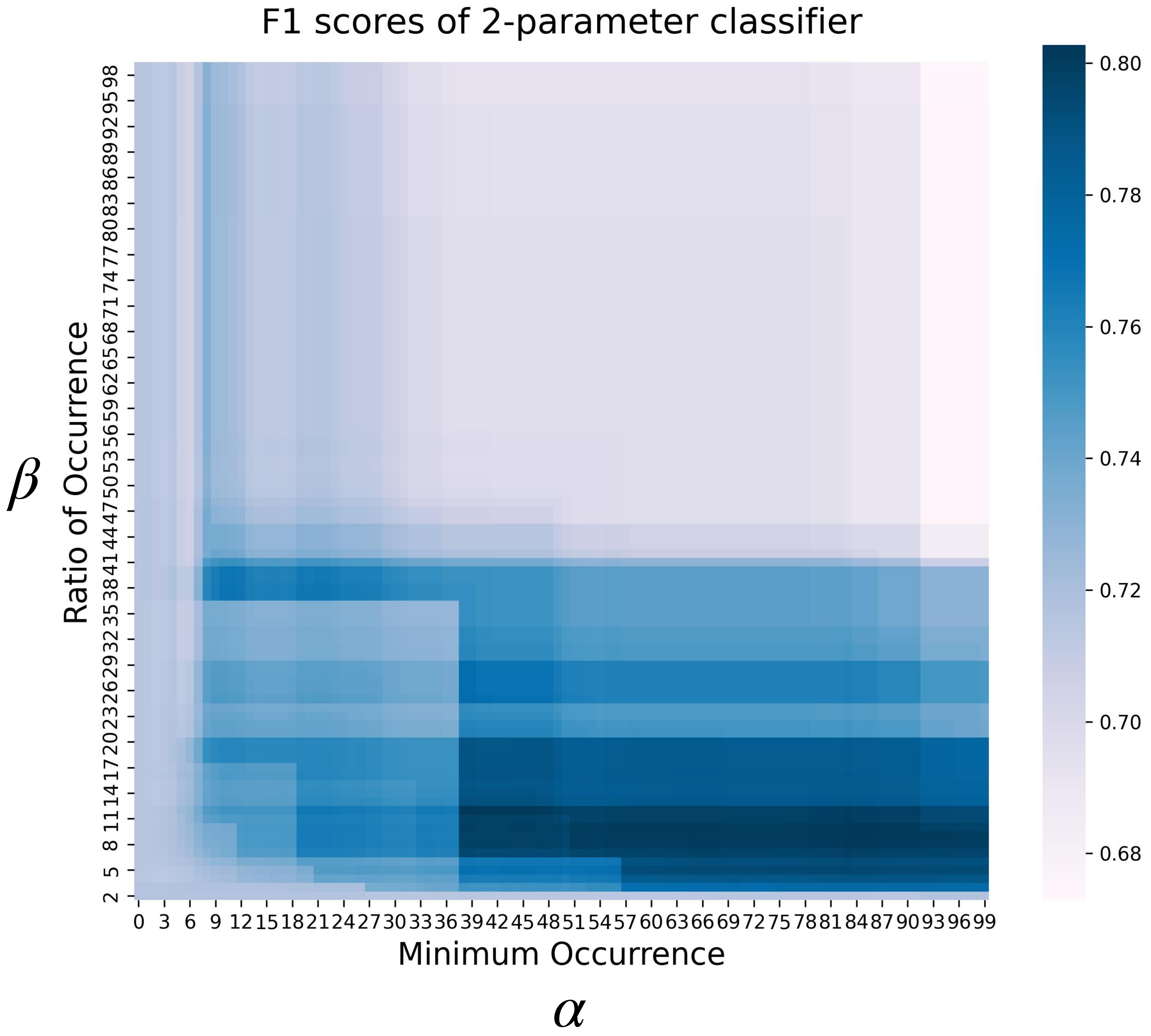
Could we build a better keyword classifier?

- Test 10,000 potential classifiers through a two-parameter model
- Terms included in keyword list if:



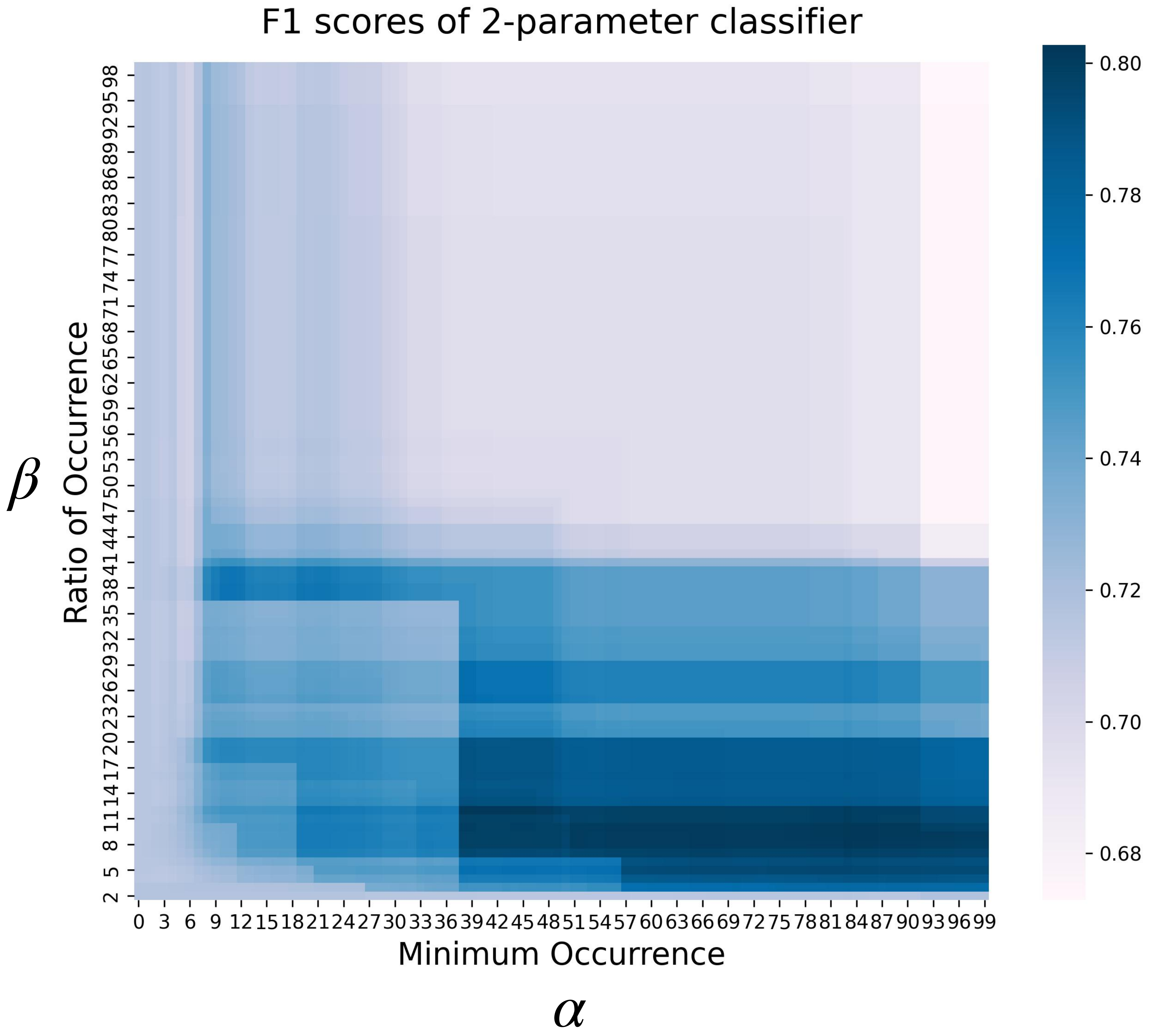
Could we build a better keyword classifier?

- Test 10,000 potential classifiers through a two-parameter model
- Terms included in keyword list if:
  - ➡ Occur at least α times



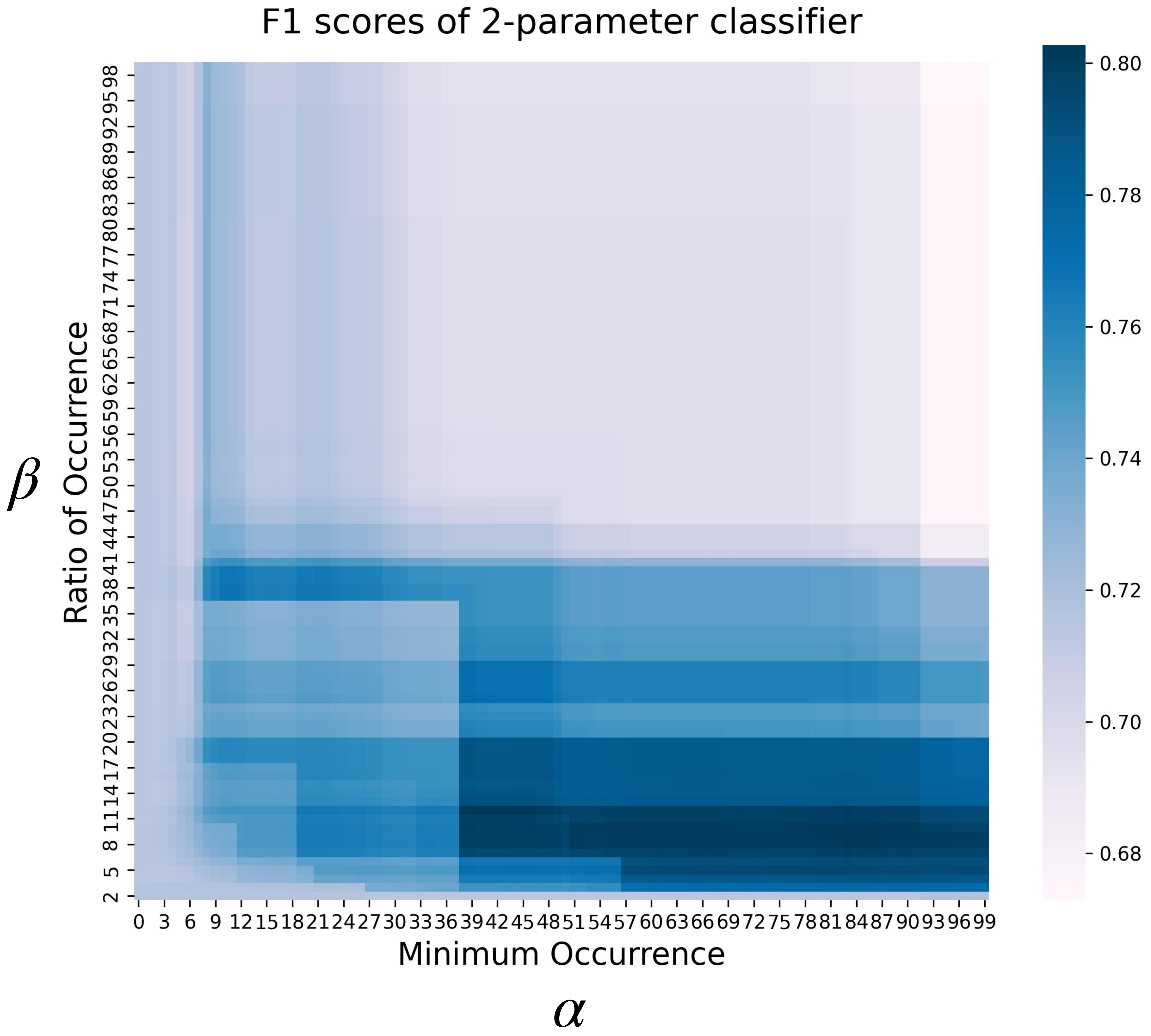
Could we build a better keyword classifier?

- Test 10,000 potential classifiers through a two-parameter model
- Terms included in keyword list if:
  - ➡ Occur at least α times
  - ➡ Occur β times more in election tweets (defined from handcoding)



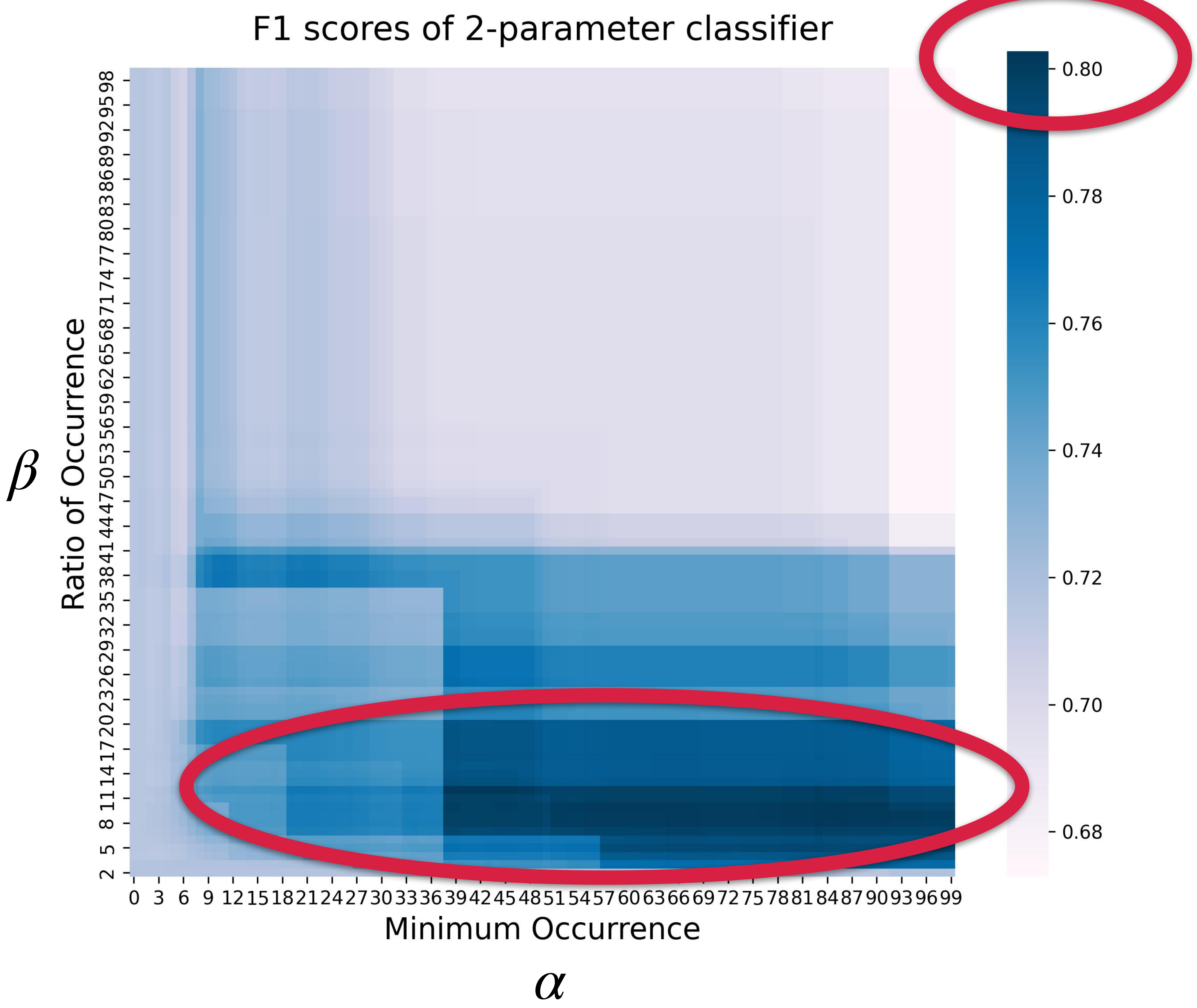
Could we build a better keyword classifier?

- Test 10,000 potential classifiers through a two-parameter model
- Terms included in keyword list if:
  - ➡ Occur at least α times
  - ➡ Occur β times more in election tweets (defined from handcoding)
- Calculate accuracy for each resulting keyword list

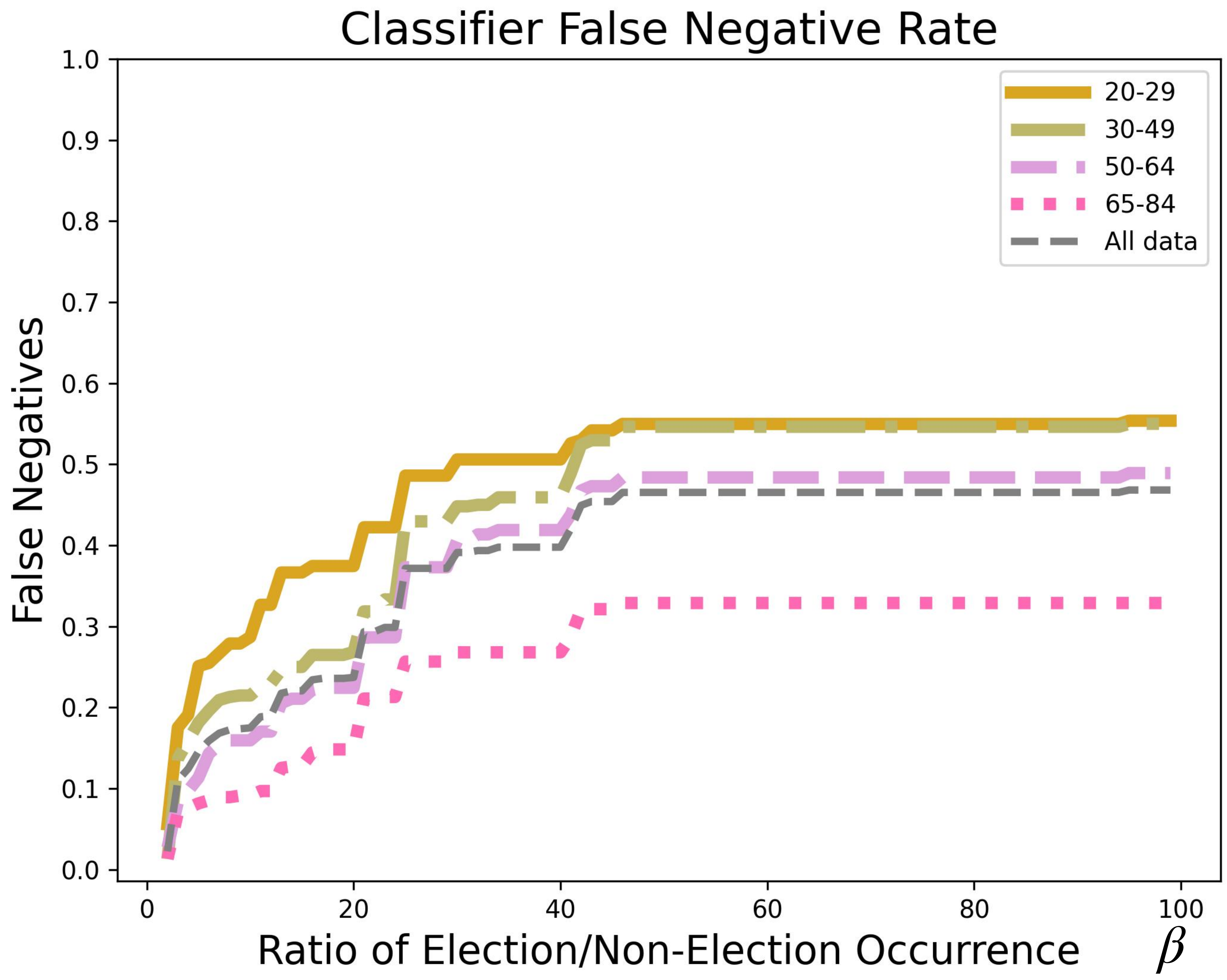


Could we build a better keyword classifier?

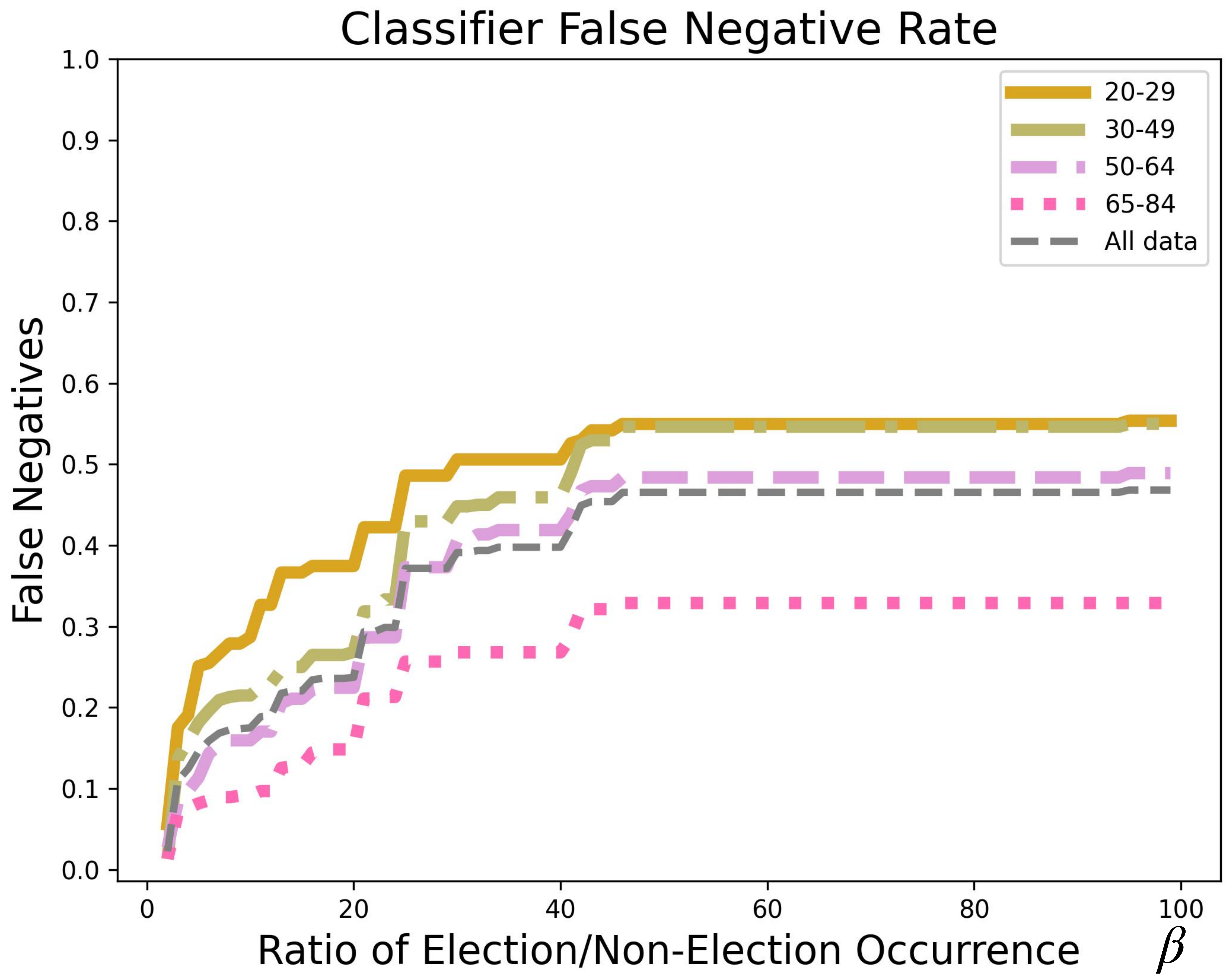
- Test 10,000 potential classifiers through a two-parameter model
- Terms included in keyword list if:
  - ➡ Occur at least α times
  - ➡ Occur β times more in election tweets (defined from handcoding)
- Calculate accuracy for each resulting keyword list
- Best models achieved (only) an F1-score of ~ 0.8



Could we build a better keyword classifier?

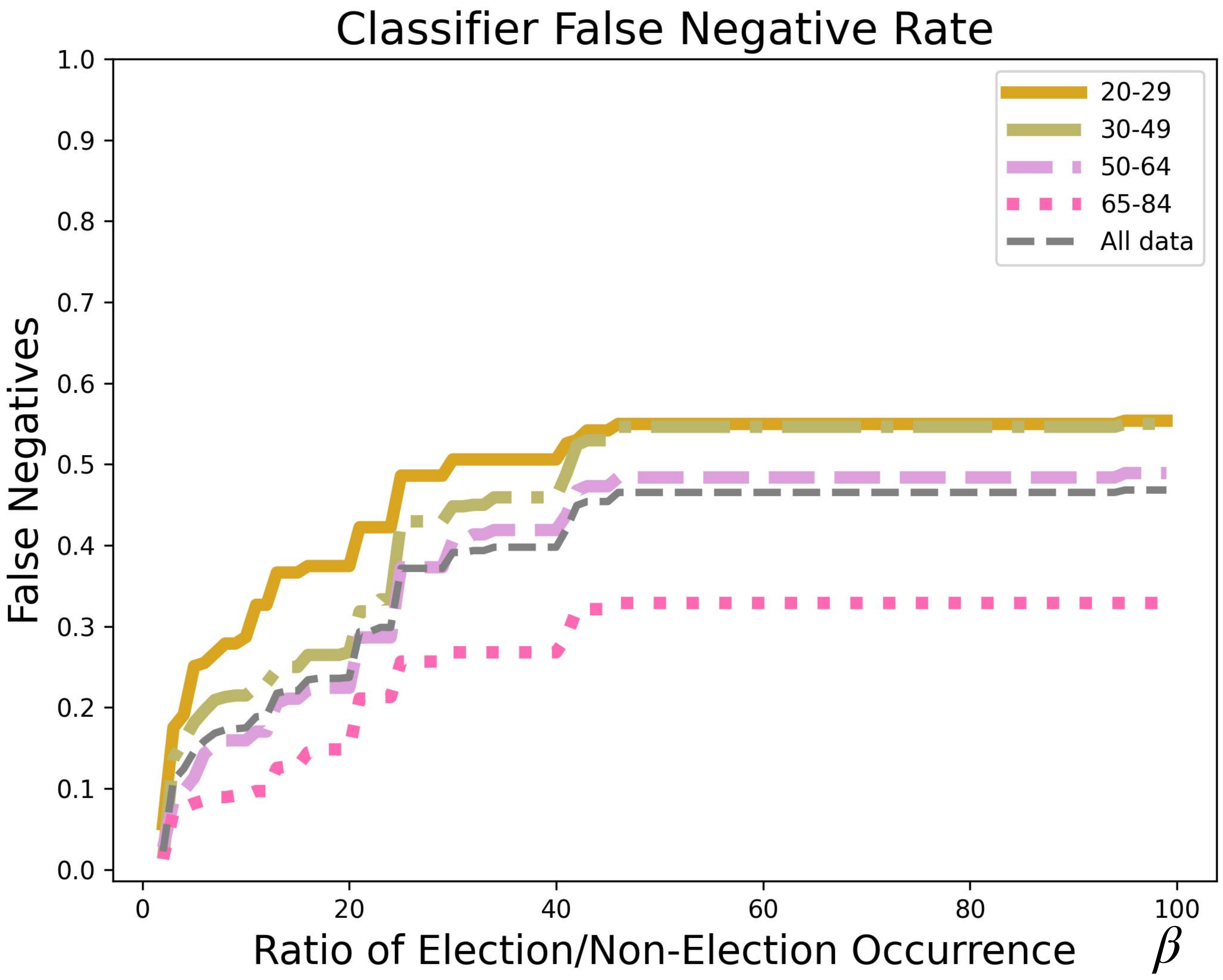


Could we build a better keyword classifier?



Older users consistently have fewer of their election tweets misclassified

Could we build a better keyword classifier?

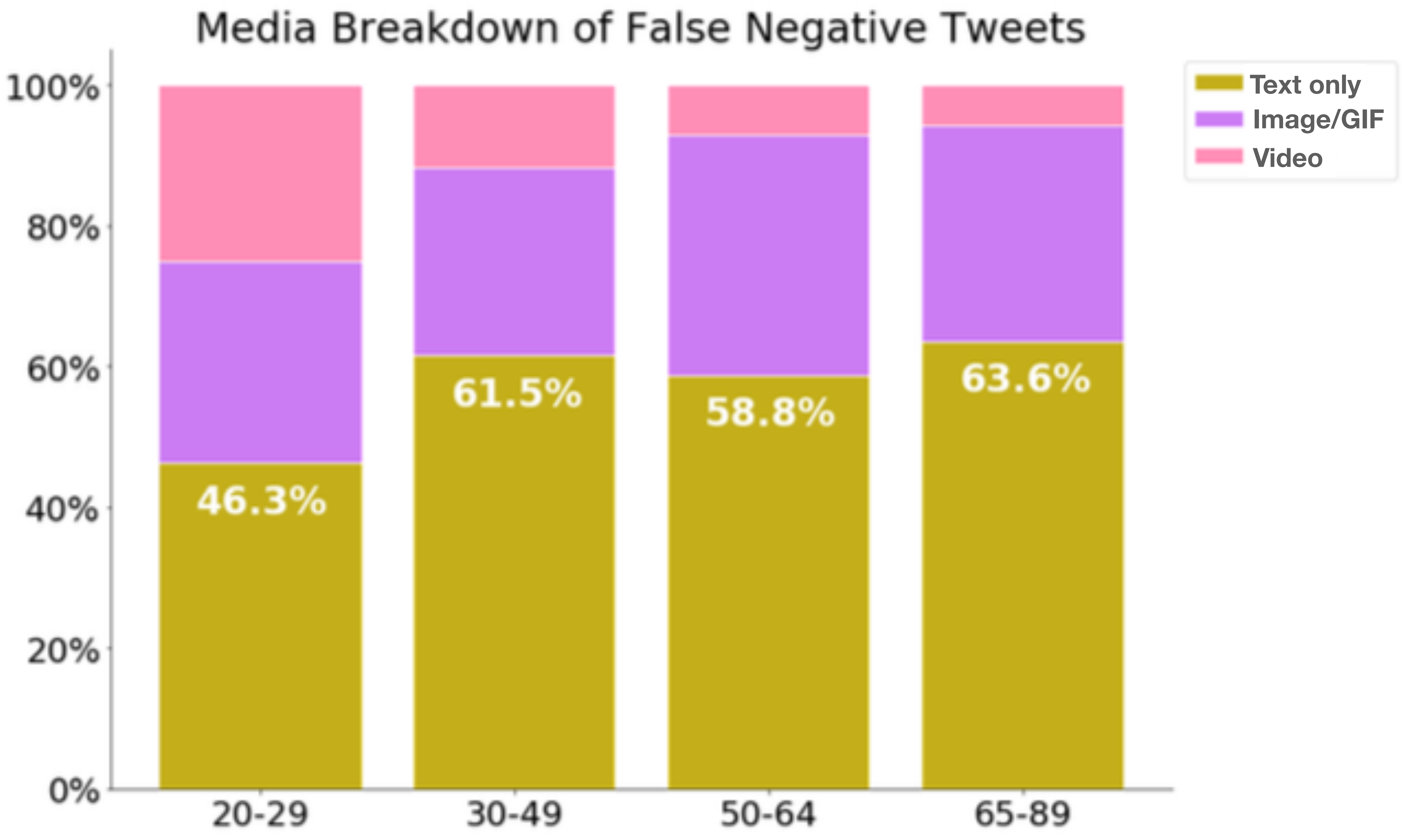


Younger users consistently have more of their election tweets misclassified

Older users consistently have fewer of their election tweets misclassified

Generational Trends?

Younger users frequently communicate using non-textual media, so keyword classifiers may underrepresent their political speech



Final Thoughts

Final Thoughts

- How we measure “speech” has implications for whose speech we capture

Final Thoughts

- How we measure “speech” has implications for whose speech we capture
- Matched panel data can help us interrogate these measurement biases

Final Thoughts

- How we measure “speech” has implications for whose speech we capture

- Matched panel data can help us interrogate these measurement biases
  - ➡ Have meaningful “negative” samples (eg, all users’ posts)

Final Thoughts

- How we measure "speech" has implications for whose speech we capture

- Matched panel data can help us interrogate these measurement biases
    - Have meaningful "negative" samples (eg, all users' posts)
    - Have age and other demographic info, can compare impact across categories

Final Thoughts

- How we measure "speech" has implications for whose speech we capture
- Matched panel data can help us interrogate these measurement biases
  - Have meaningful "negative" samples (eg, all users' posts)
  - Have age and other demographic info, can compare impact across categories

Thank you!

Sarah Shugars
they/them/theirs

Assistant Professor
Rutgers University

sarah.shugars@rutgers.edu
BlueSky: @Shugars