Factoring Fact-Checks:

Structured Information Extraction from Fact-Checking Articles

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, Cong Yu

Northeastern University

Google AI

Background: fact-checks

Factoring Fact-Checks

What is it?

Background: fact-checks

Factoring Fact-Checks

WIKIPEDIA
The Free Encyclopedia

Fact-checking is the act of checking factual information in non-fictional text in order to determine the veracity and correctness of the factual statements in the text.

Background: fact-checks

Factoring Fact-Checks

Fact-checking = **the act of checking facts.**

An article that does fact-checking is called **“fact-checks”**.

Background: fact-checks

POLITIFACT **No, ‘newspaper’ isn’t an acronym for ‘north, east, west, south, past and present event report’**

Has the word “newspaper” really been an acronym all this time?
That’s what one viral Facebook post claims.
According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”
……
The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros”.
This claim is a repurposed hoax. We rate it Pants on Fire!

Background: **factors**

POLITIFACT

No, ‘newspaper’ isn’t an acronym for ‘north, east, west, south, past and present event report’

Has the word “newspaper” really been an acronym all this time?
That’s what one viral Facebook post claims.
According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.” Claim
……
The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros”.
This claim is a repurposed hoax. We rate it Pants on Fire!

Background: **factors**

POLITIFACT **No, 'newspaper' isn't an acronym for 'north, east, west, south, past and present event report'**

Has the word "newspaper" really been an acronym all this time?
That's what one viral Facebook post claims. Claimant
According to the post, which has gotten over 2,400 shares in 24 hours, "newspaper" is an acronym for "North, East, West, South, Past and Present Report." Claim
……
The word paper alone has origins in the Latin word "papyrus," the stalks used to make paper, and the Greek word "papyros".
This claim is a repurposed hoax. We rate it Pants on Fire!

Background: **factors**

POLITIFACT **No, 'newspaper' isn't an acronym for 'north, east, west, south, past and present event report'**

Has the word "newspaper" really been an acronym all this time?
That's what one viral Facebook post claims. Claimant
According to the post, which has gotten over 2,400 shares in 24 hours, "newspaper" is an acronym for "North, East, West, South, Past and Present Report." Claim
……
The word paper alone has origins in the Latin word "papyrus," the stalks used to make paper, and the Greek word "papyros".
This claim is a repurposed hoax. We rate it Pants on Fire! Verdict

Background: factors

POLITIFACT No, 'newspaper' isn't an acronym for 'north, east, west, south, past and present event report'

Claim: "newspaper" is an acronym for "North, East, West, South, Past and Present Report."

Claimant: viral Facebook post

Verdict: Pants on Fire!

Background: application

newspaper acronym

No, 'newspaper' isn't an acronym for 'north, east, west, south ...

https://www.politifact.com › statements › sep › facebook-posts › no-newsp...

Claim: Says the word newspaper stands for "north, east, west, south, past and present event report."

Claimed by: Facebook posts

Fact check by PolitiFact: Pants on Fire

Background: application

bing

newspaper acronym

No, 'newspaper' is not an acronym of 'North, East, West ...

https://africacheck.org/fbcheck/no-newspaper-is-not-an-acronym-of-north-east-west...

Claim: 'Newspaper' is an acronym of 'North, East, West, South, Past and Present Events Report'

False · Fact checked by ResultPartUpdater

Background: motivation

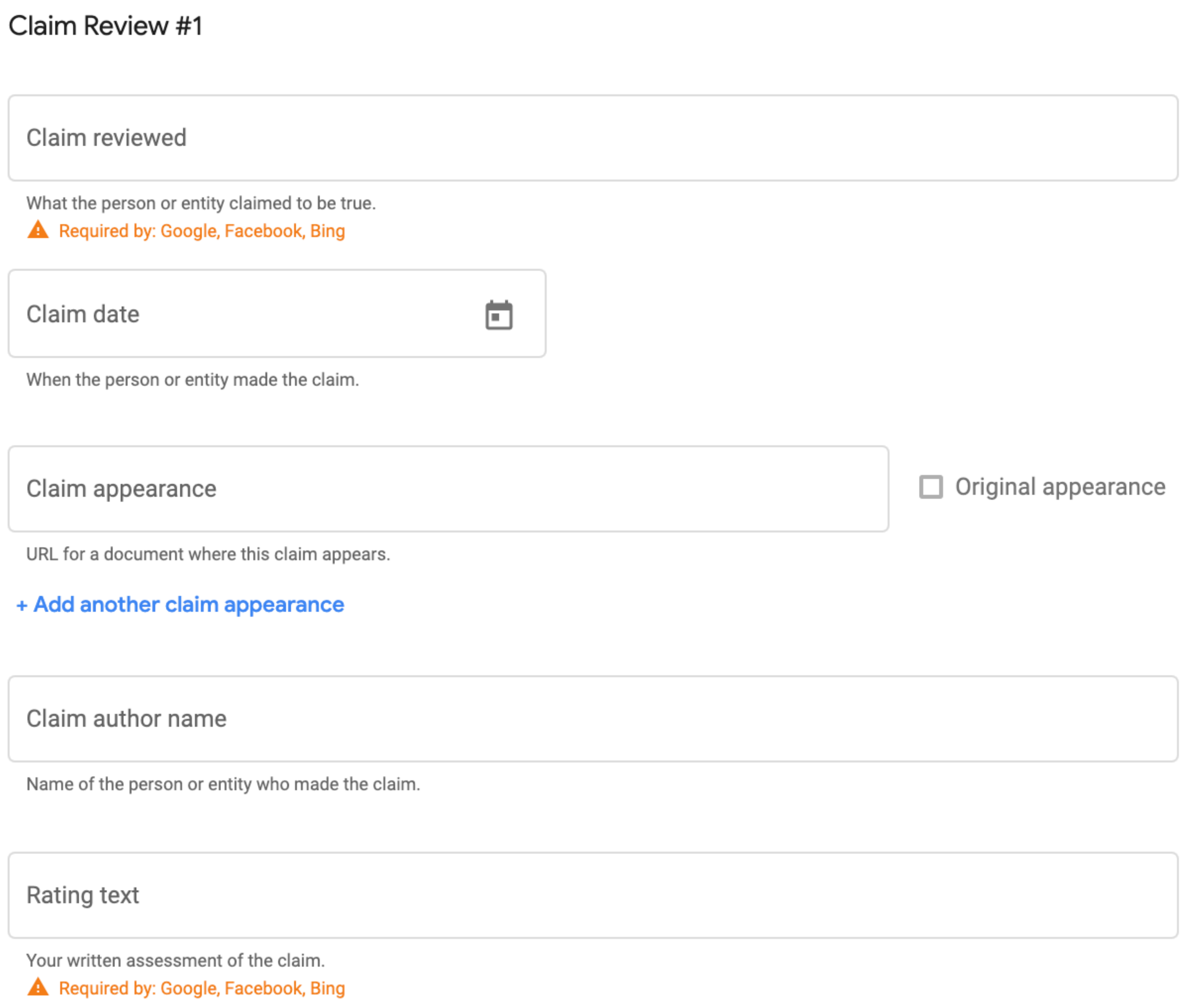
Question:

How to get these factors?

Background: motivation

Fact-check markup tool:

https://toolbox.google.com/factcheck


Claim Review #1

Claim reviewed

What the person or entity claimed to be true.

Required by: Google, Facebook, Bing

Claim date

When the person or entity made the claim.

Claim appearance

Original appearance

URL for a document where this claim appears.

+ Add another claim appearance

Claim author name

Name of the person or entity who made the claim.

Rating text

Your written assessment of the claim.

Required by: Google, Facebook, Bing

Background: motivation

ClaimReview markup:

https://schema.org/ClaimReview

```
<head>
<title>The world is flat</title>
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "ClaimReview",
  "datePublished": "2016-06-22",
  "url": "http://example.com/news/science/worldisflat.html",
  "claimReviewed": "The world is flat",
  "itemReviewed": {
    "@type": "Claim",
    "author": {
      "@type": "Organization",
      "name": "Square World Society"
    }
  },
  "reviewRating": {
    "@type": "Rating",
    "alternateName": "False"
  }
}
</script>
</head>
```

Background: problem

Problem:

It takes time!

As of July 2019, < 50% fact-checkers use it. [1]

[1] Joel Luther. 2019. Reporters' Lab Launches Global Effort to Expand the Use of ClaimReview.
https://reporterslab.org/lab-launches-global-effort-to-expand-claimreview

Background: proposal

Proposal:

Automatically extracting factors from fact-checks.

(factoring fact-checks)

Background: steps

Steps:

- Explore fact-check data for patterns of factors.
- Experiment with information extraction models.

Steps: data

Steps:

- Explore fact-check data for patterns of factors.
- Experiment with information extraction models.

Data: source

- Fact-check dataset from DataCommons. [2]

[2] DataCommons. 2019. Fact-Check Dataset. https://datacommons.org/factcheck

Data: source

- Fact-check dataset from DataCommons. [2]
- 6,216 fact-checks (English).

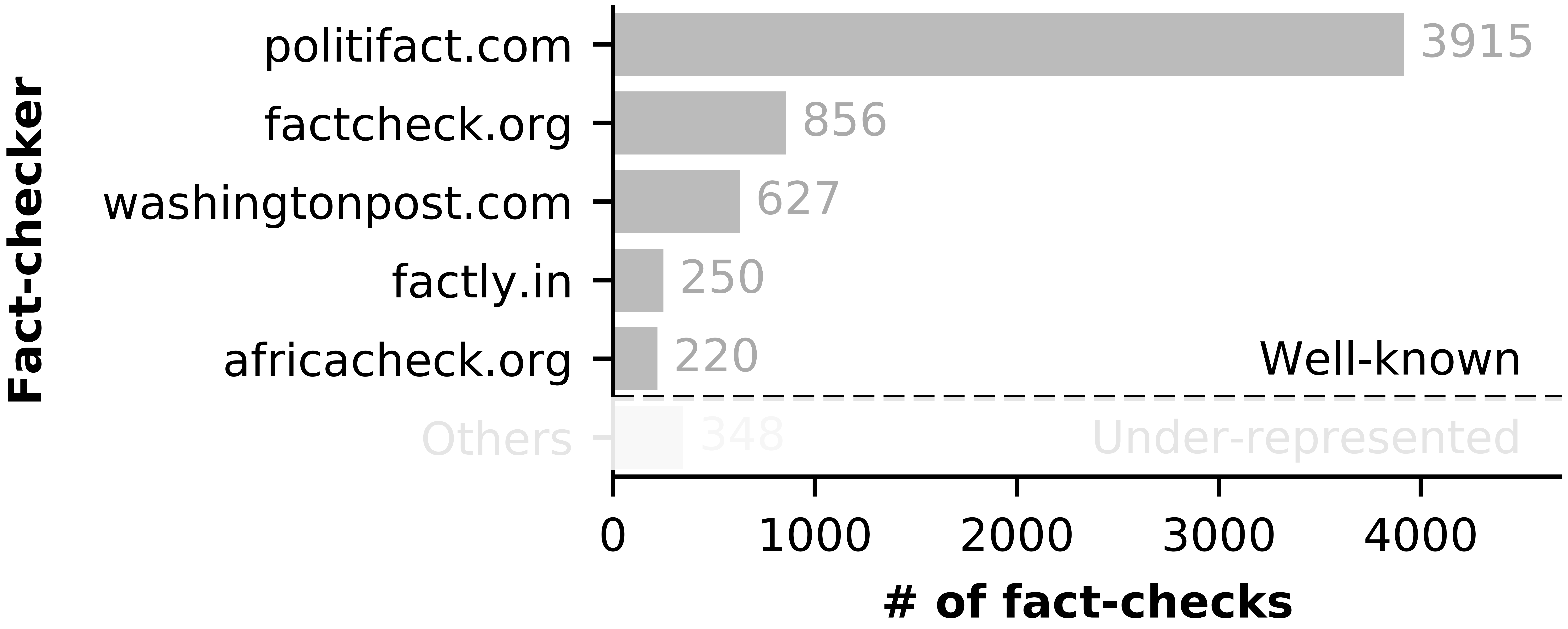
[2] DataCommons. 2019. Fact-Check Dataset. https://datacommons.org/factcheck

Data: source

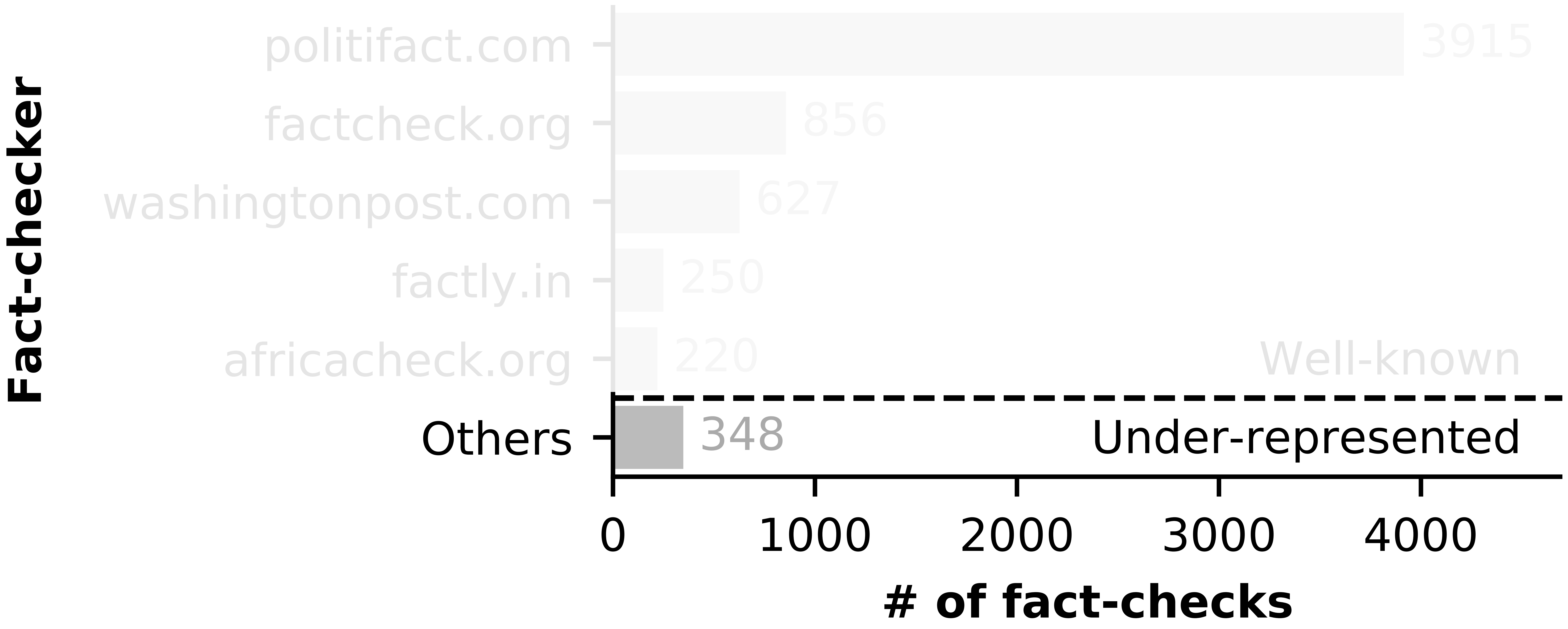
- Fact-check dataset from DataCommons. [2]
- 6,216 fact-checks (English).
- Reported factors (claim, claimant, verdict).

[2] DataCommons. 2019. Fact-Check Dataset. https://datacommons.org/factcheck

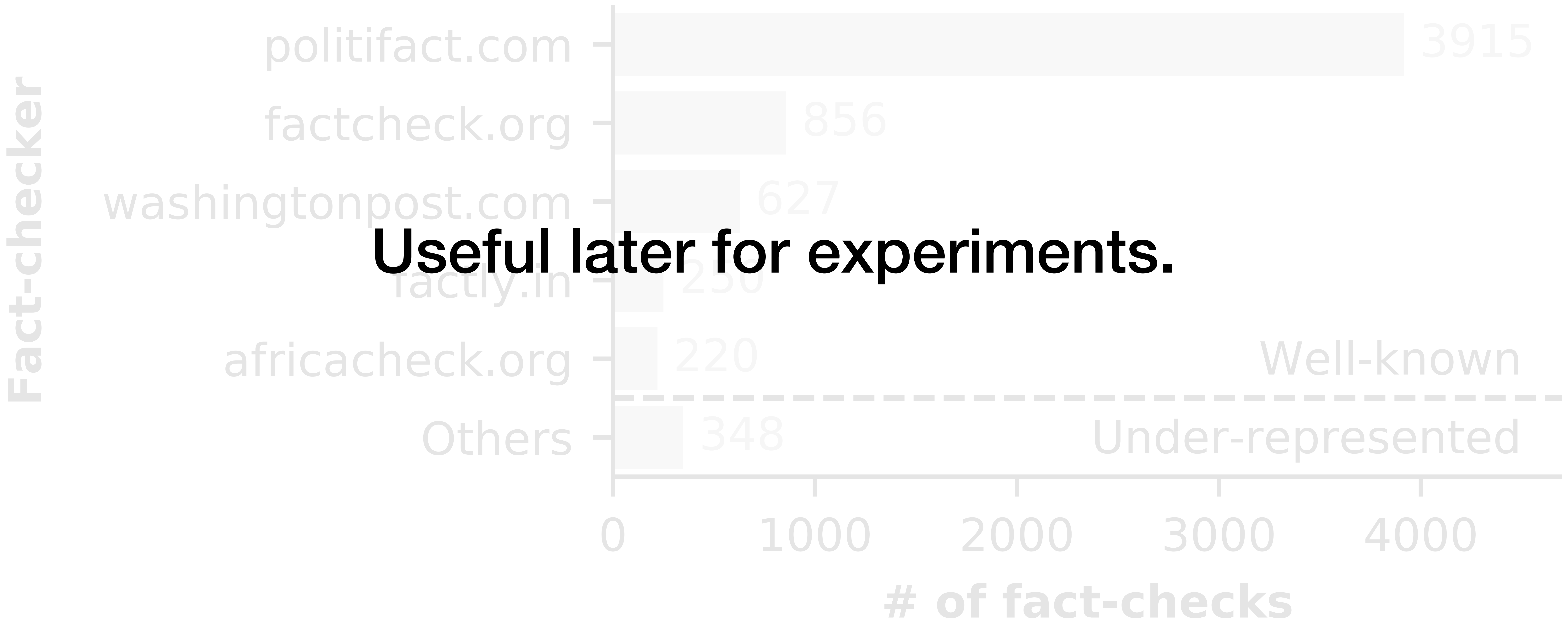
Data: who are the fact-checkers?



Data: who are the fact-checkers?



Data: who are the fact-checkers?



Useful later for experiments.

Data: can factors be found in the fact-checks?

- Exact string matching.
- Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.

Data: can factors be found in the fact-checks?

- Exact string matching.
- Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.

Paraphrasing.

Claim in article: “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

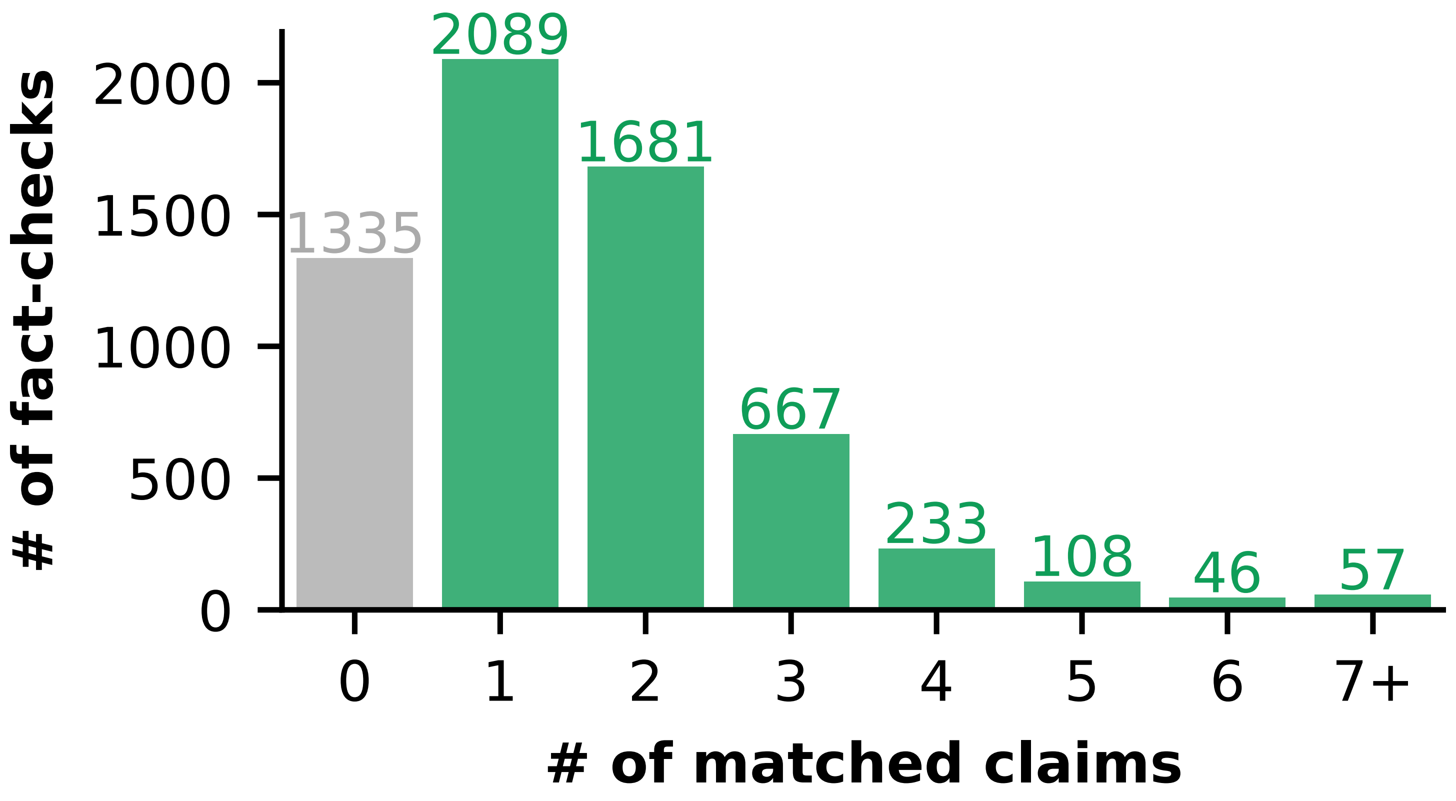
Reported claim: says the word newspaper stands for “north, east, west, south, past and present report.”

Data: can factors be found in the fact-checks?

- ~~Exact string matching.~~
- ~~Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.~~
- At least 2/3 of overlap.
- Minimum window substring matching. [3]

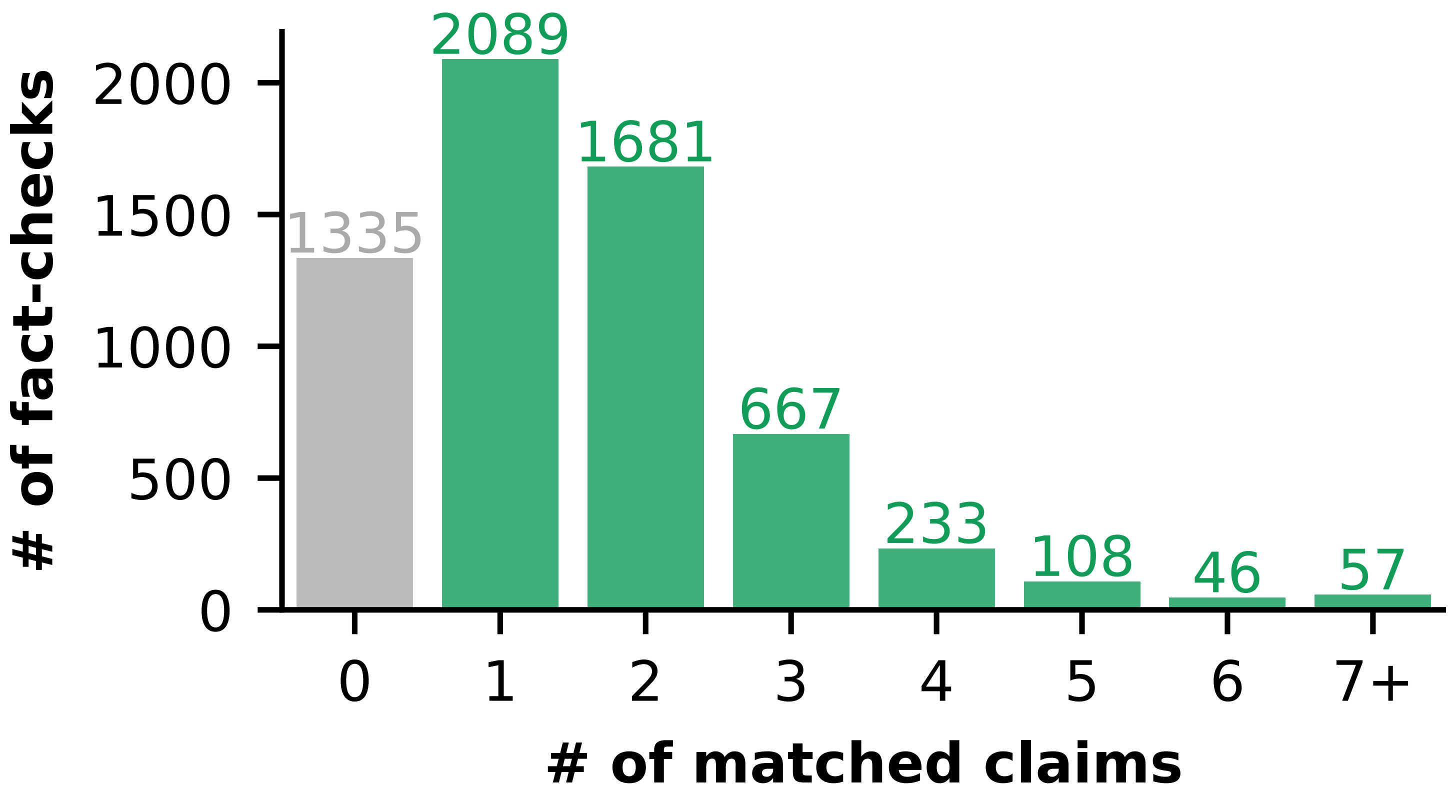
[3] LeetCode. 2014. Minimum Window Substring. https://leetcode.com/problems/minimum-window-substring

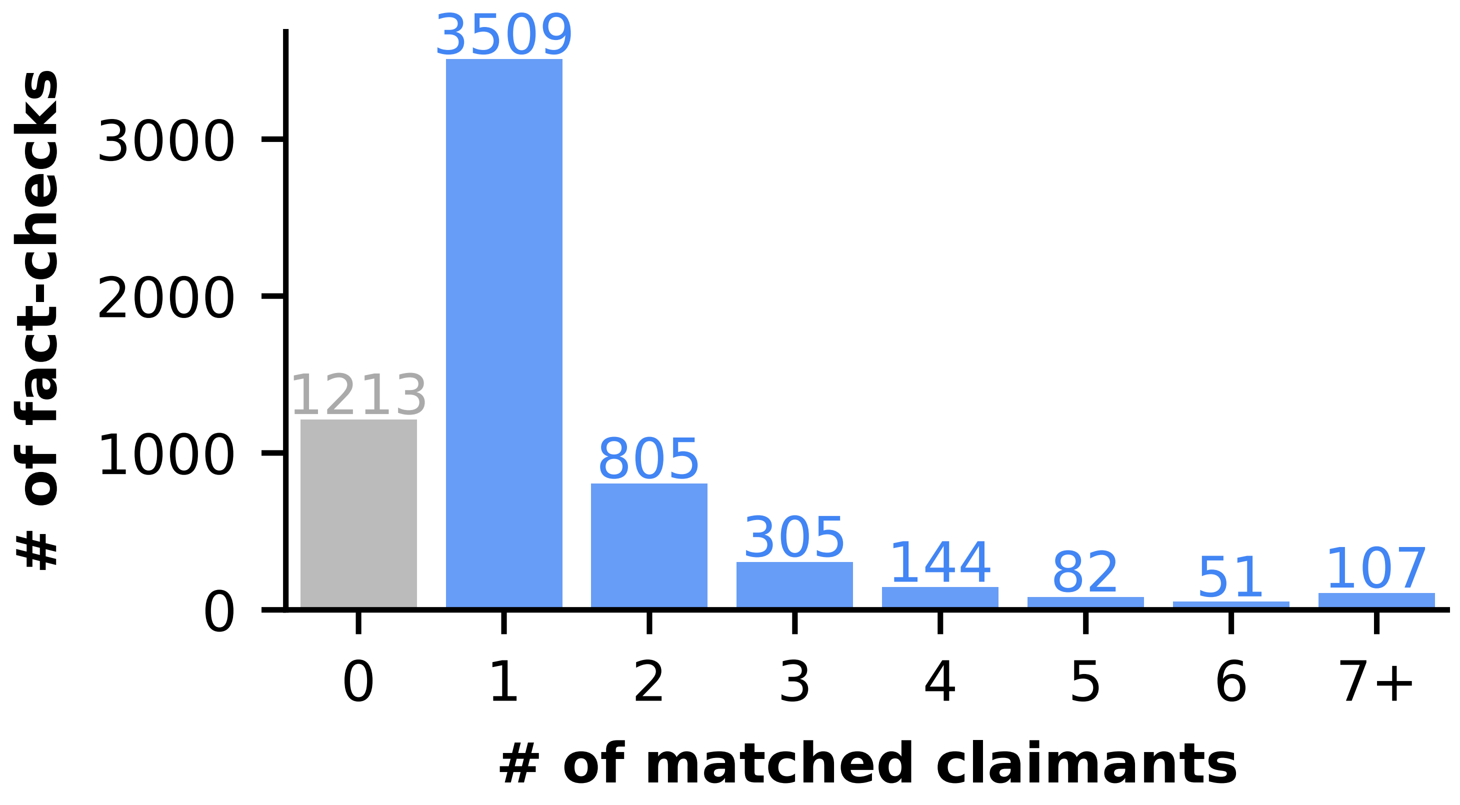
Data: can factors be found in the fact-checks?



79% of claims.

Data: can factors be found in the fact-checks?

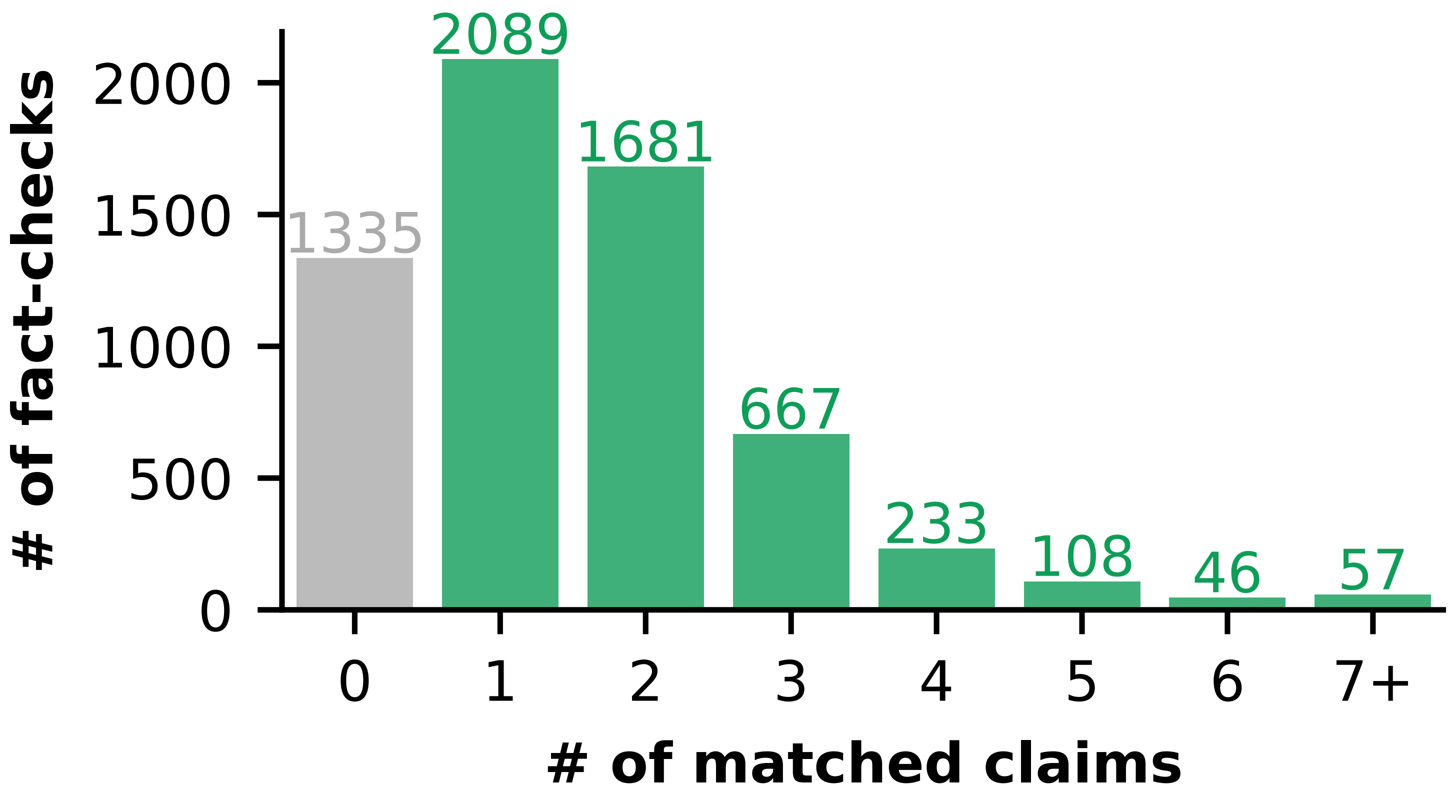


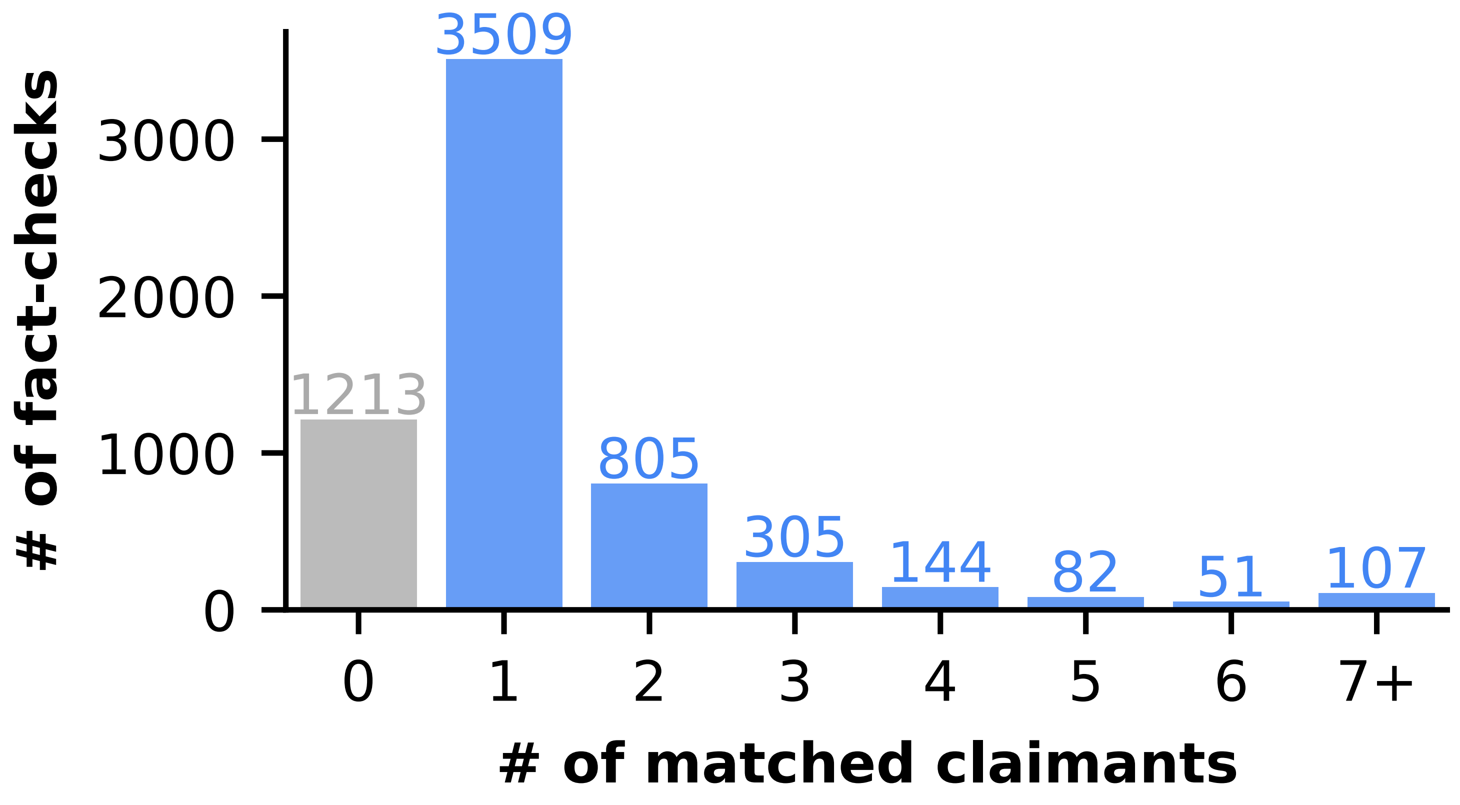


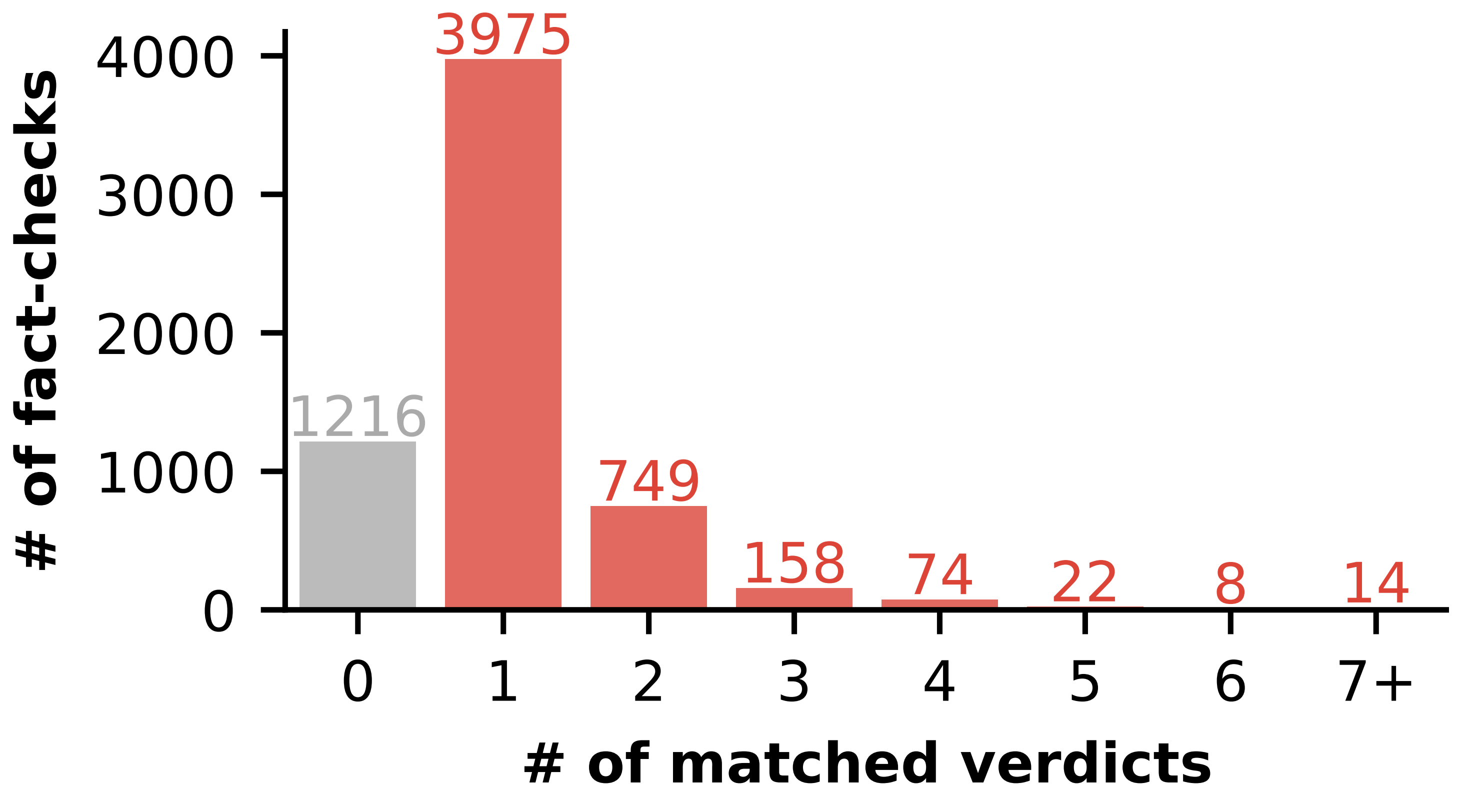
79% of claims.

80% of claimants.

Data: can factors be found in the fact-checks?







79% of claims. 80% of claimants. 80% of verdicts.

Data: where are the factors in the fact-check?

- Relative position.
- Position / length of the fact-check.

  0 = the head of the fact-check.

  1 = the tail of the fact-check.

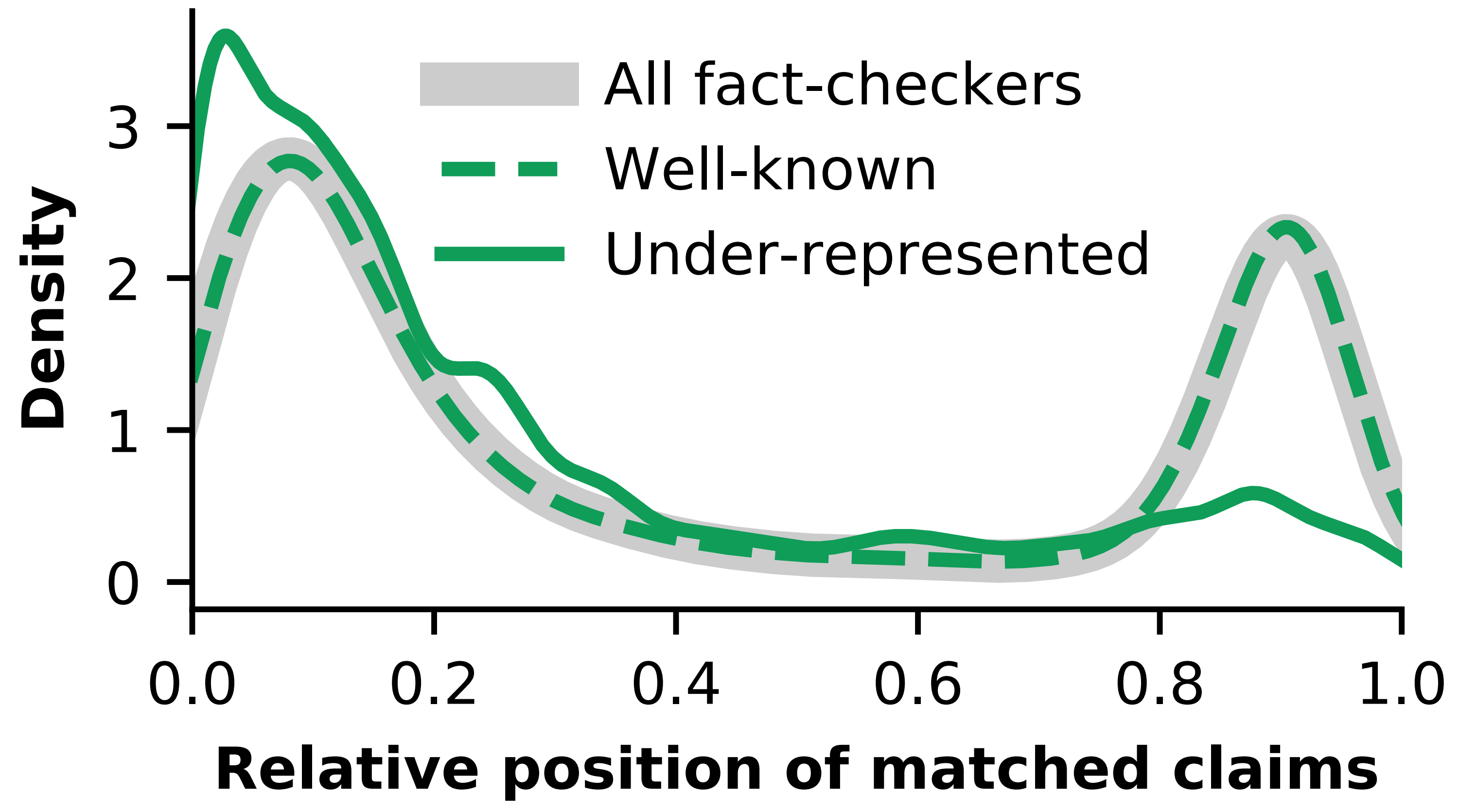
Data: where are the factors in the fact-check?

- Relative position.
- Position / length of the fact-check.

  0 = the head of the fact-check.

  1 = the tail of the fact-check.
- Separate well-known and under-represented fact-checkers

Data: where are the factors in the fact-check?

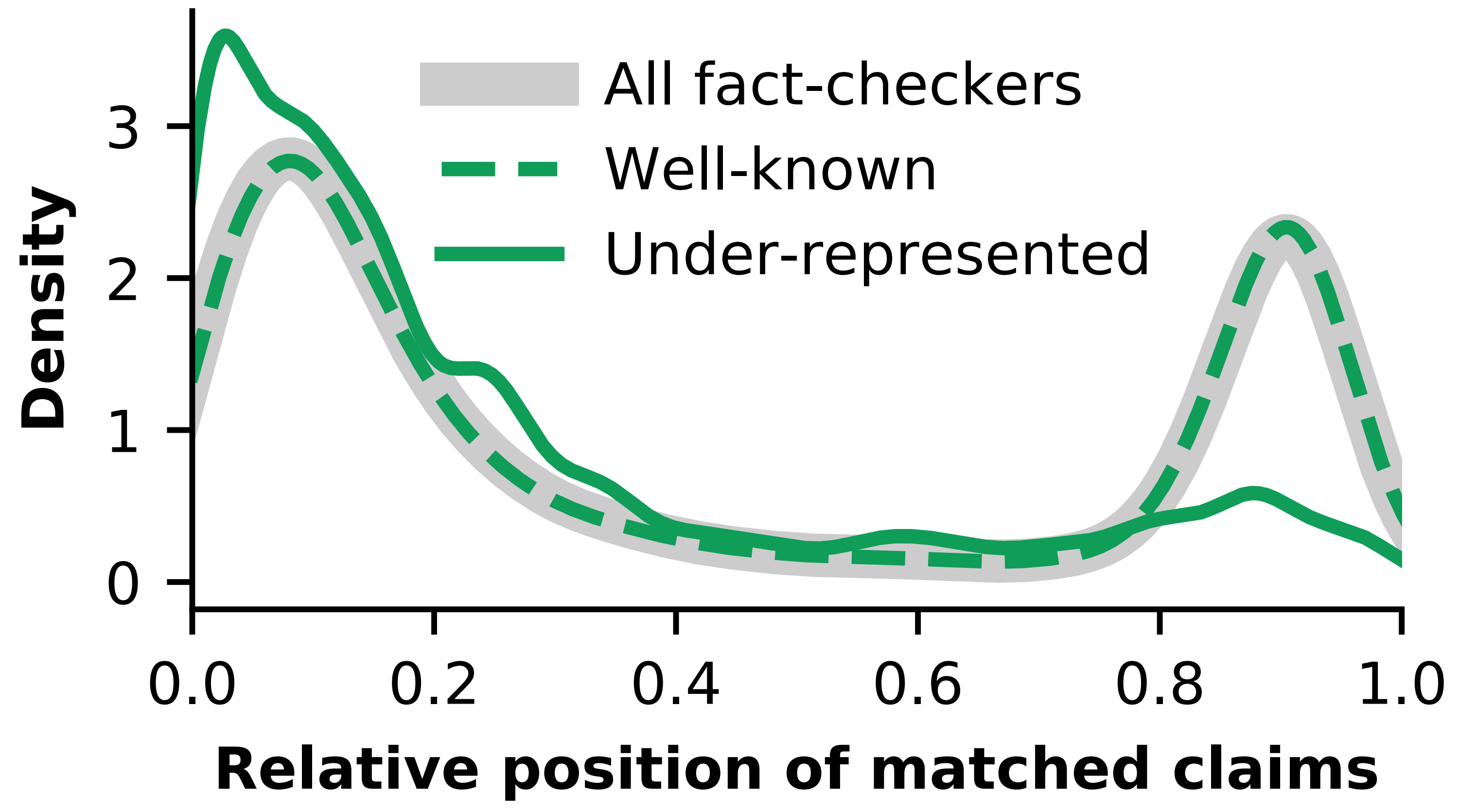


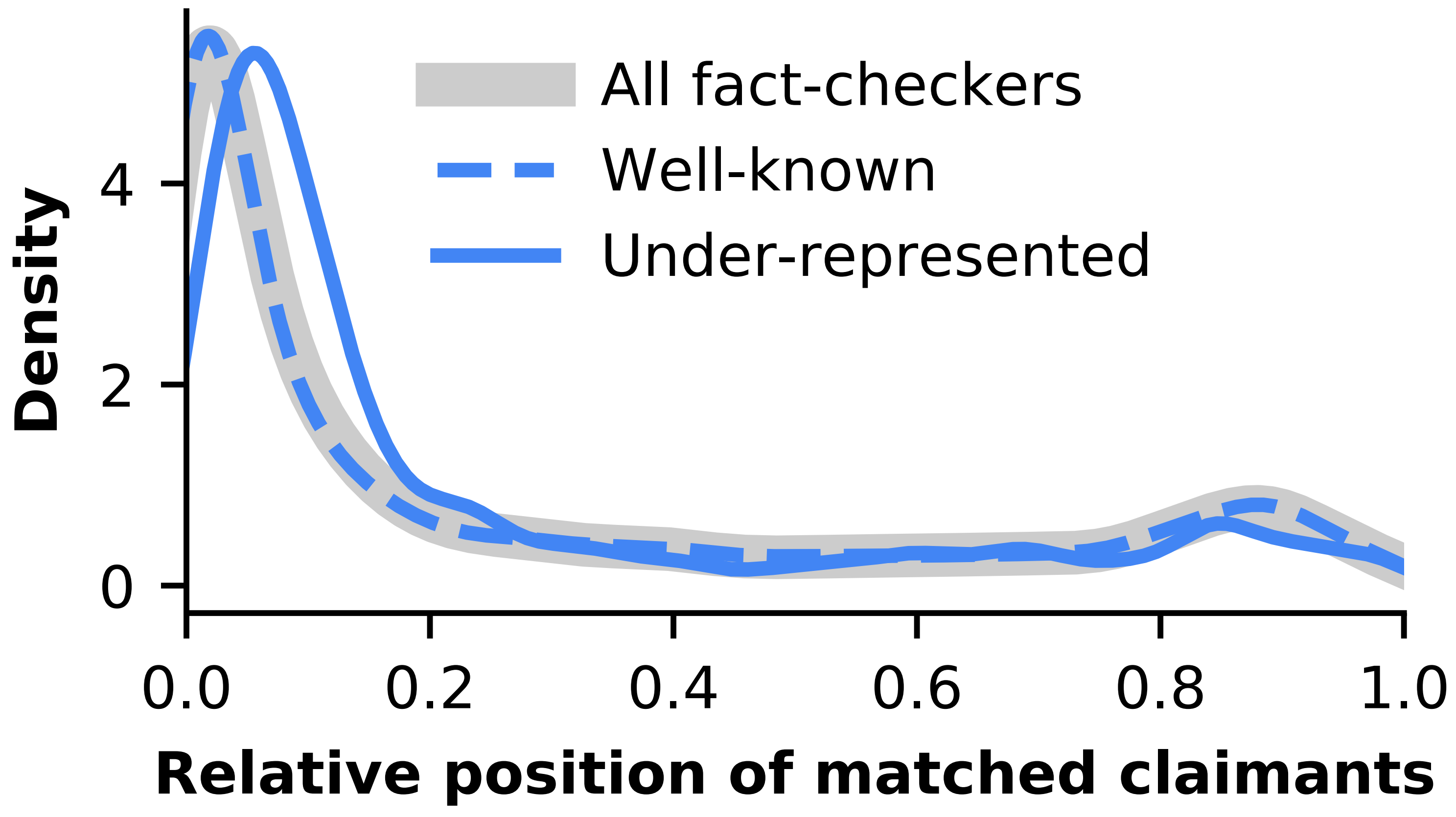
Claims:

Well-known:
head and tail

Under-represented:
head only

Data: where are the factors in the fact-check?





Claims:

Well-known:
head and tail

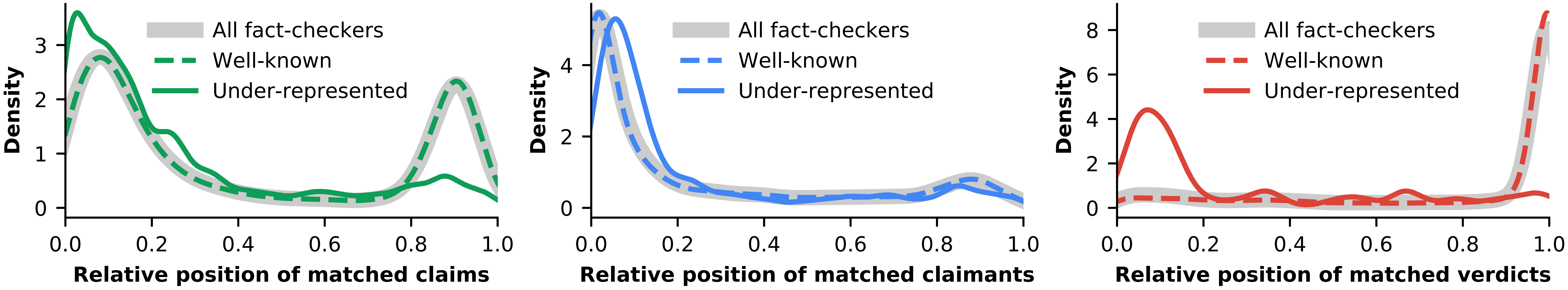
Under-represented:
head only

Claimants:

Well-known:
head only

Under-represented:
head only

Data: where are the factors in the fact-check?



Claims:

Well-known:
head and tail

Under-represented:
head only

Claimants:

Well-known:
head only

Under-represented:
head only

Verdicts:

Well-known:
tail only

Under-represented:
head only

Data: where are the factors in the fact-check?

Useful later for experiments.

Claims:	Claimants:	Verdicts:
Well-known: head and tail	Well-known: head only	Well-known: tail only
Under-represented: head only	Under-represented: head only	Under-represented: head only

Steps: experiments

Steps:

- ~~Explore fact-check data for patterns of factors.~~
- Experiment with information extraction models.

Task: intuition

- The factor *per se*.

Claim: factual statement, numbers, statistics, etc.
Claimant: person, organization, etc.
Verdict: true, false, pants on fire, Pinocchio, etc.

Task: intuition

- The factor *per se*.

Claim: factual statement, numbers, statistics, etc.
Claimant: person, organization, etc.
Verdict: true, false, pants on fire, Pinocchio, etc.

- Surrounding context of the factor.

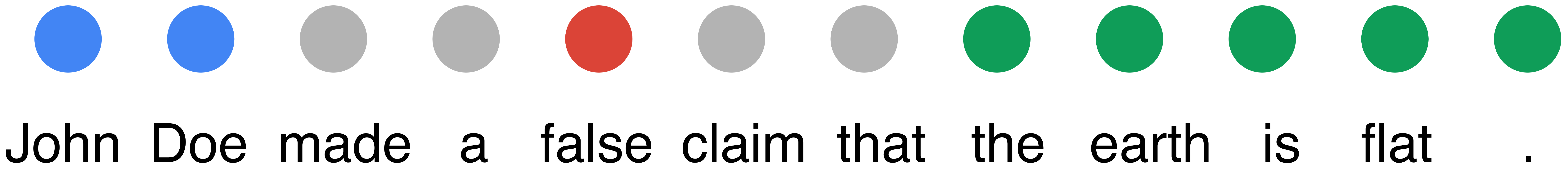
Claim: someone said/claimed (…)
Claimant: (someone) said/claimed …
Verdict: we rate it (…), a (false) rumor claims …

Task: formulation

- Sequence tagging task.
- Input: fact-check (sequence of tokens).
- Output: equal-length sequence of labels.

Task: formulation

- Sequence tagging task.
- Input: fact-check (sequence of tokens).
- Output: equal-length sequence of labels.



Task: problem

Problem:

Factors can be paraphrased.

Need to generate ground-truth token-level labels.

Task: ground-truth

Generating ground-truth labels w/ rule-based taggers.

Claim: The earth is flat . **Claimant:** John Doe **Verdict:** False

Fluent tagger:

John M Doe made a false claim that the earth is actually 100% flat .

Concise tagger:

John M Doe made a false claim that the earth is actually 100% flat .

Experiments: baseline

- Claim: ClaimBuster, top "check-worthiness". [4]

[4] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In KDD.

Experiments: baseline

- Claim: ClaimBuster, top "check-worthiness". [4]
- Claimant: entity tagging + majority.

[4] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In KDD.

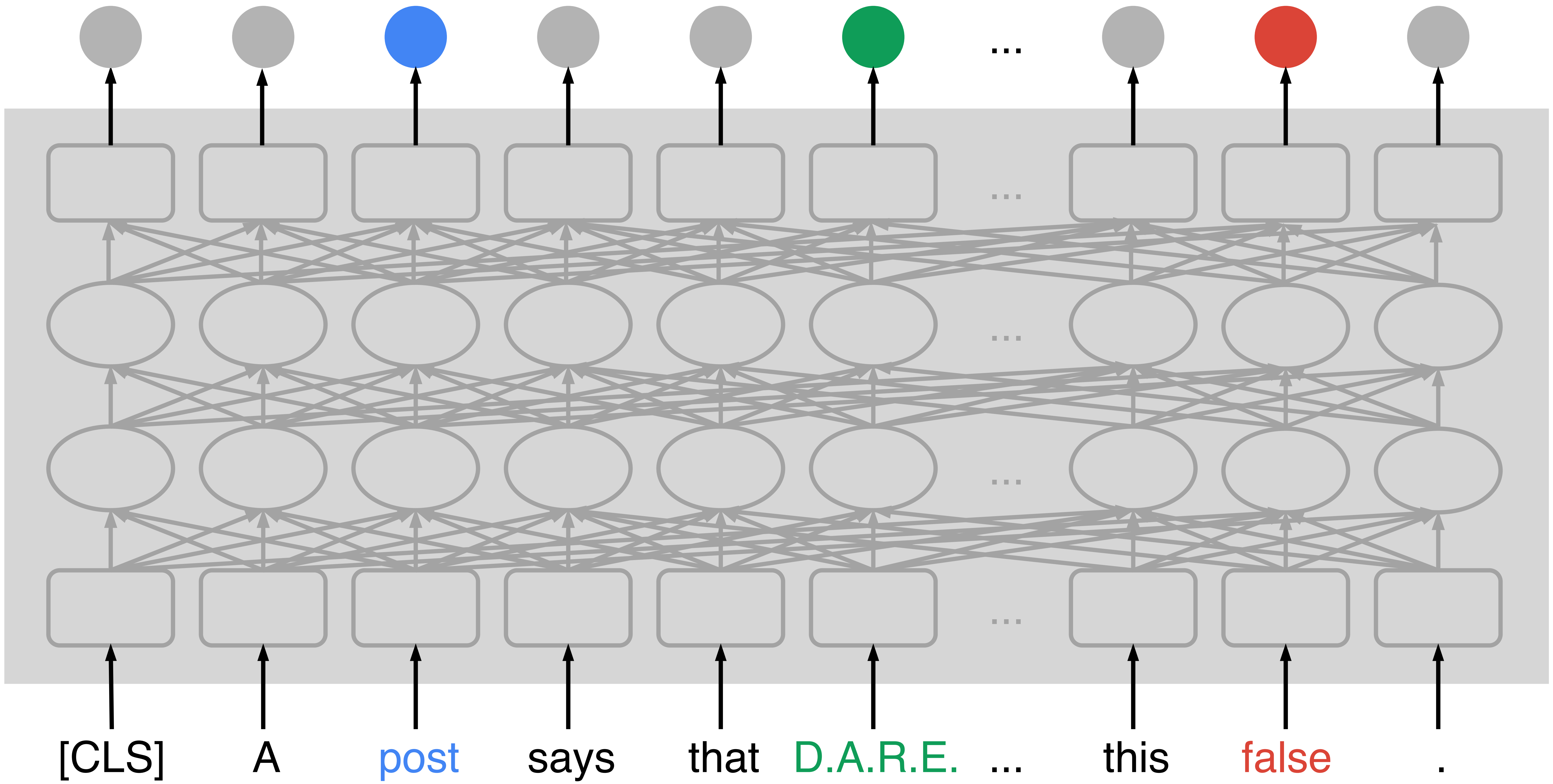
Experiments: baseline

- Claim: ClaimBuster, top "check-worthiness". [4]
- Claimant: entity tagging + majority.
- Verdict: majority.

[4] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In KDD.

Experiments: **BERT**

Replace last layer w/ tagging + cross entropy loss.



Experiments: problem

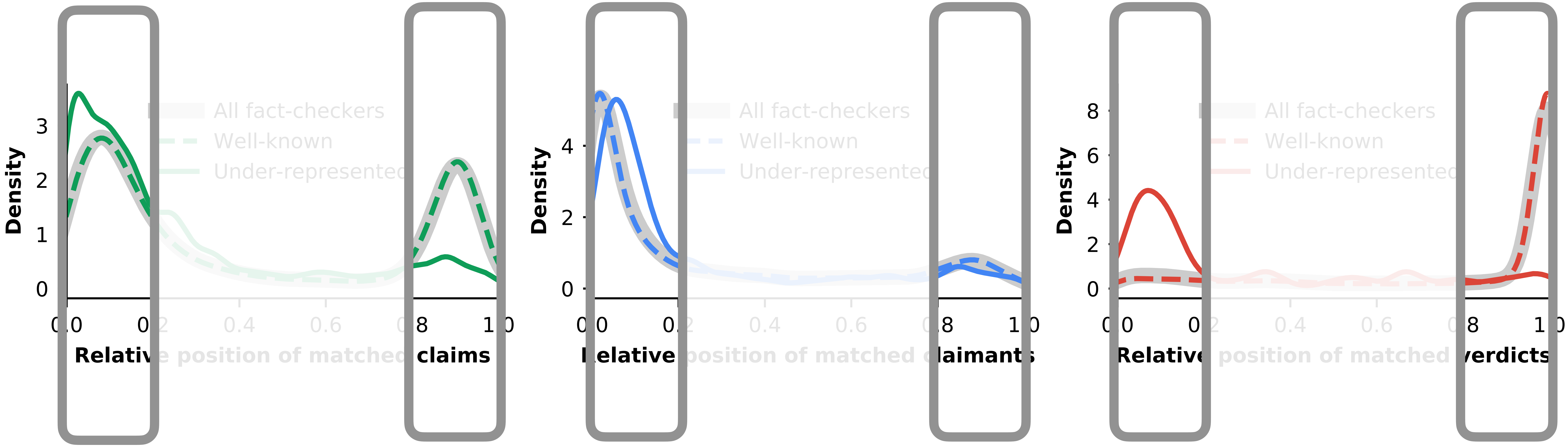
- BERT has default maximum sequence length: 512.
- Feed to it paragraph by paragraph.

Experiments: **problem**

- BERT has default maximum sequence length: 512.
- Feed to it paragraph by paragraph.

- Model only uses information of the input *per se*.
- Add external information.

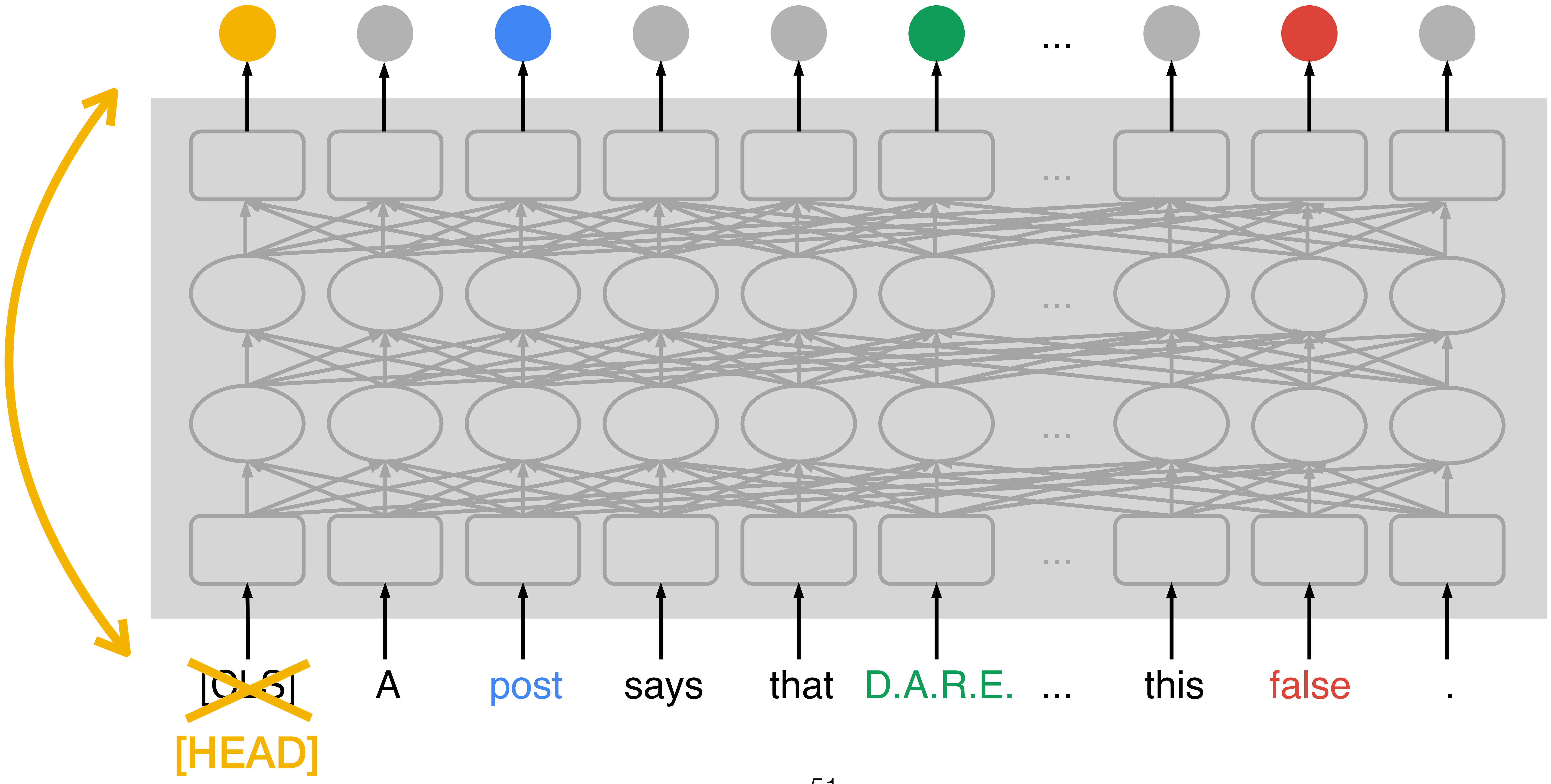
Experiments: previous observation



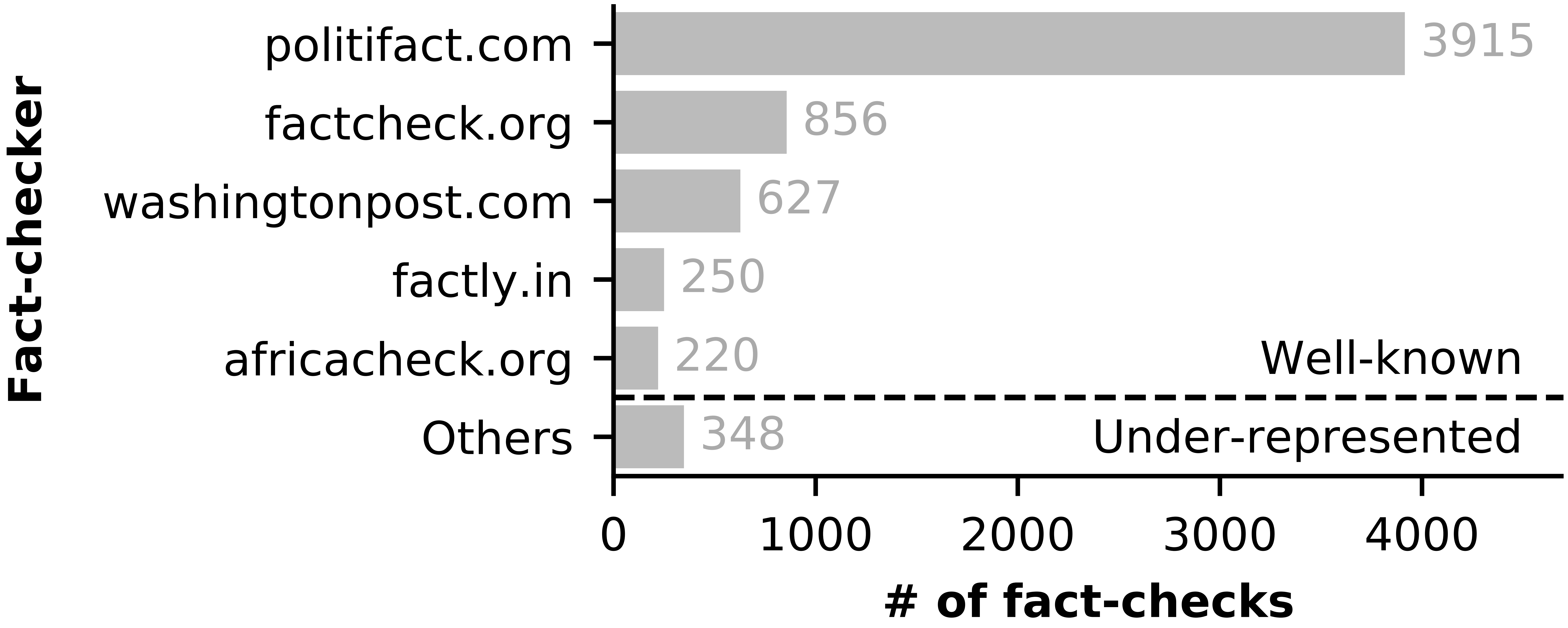
Most factors are in heads and tails of fact-checks.

Experiments: modification

Replace [CLS] w/ paragraph position [HEAD]/[BODY]/[TAIL].



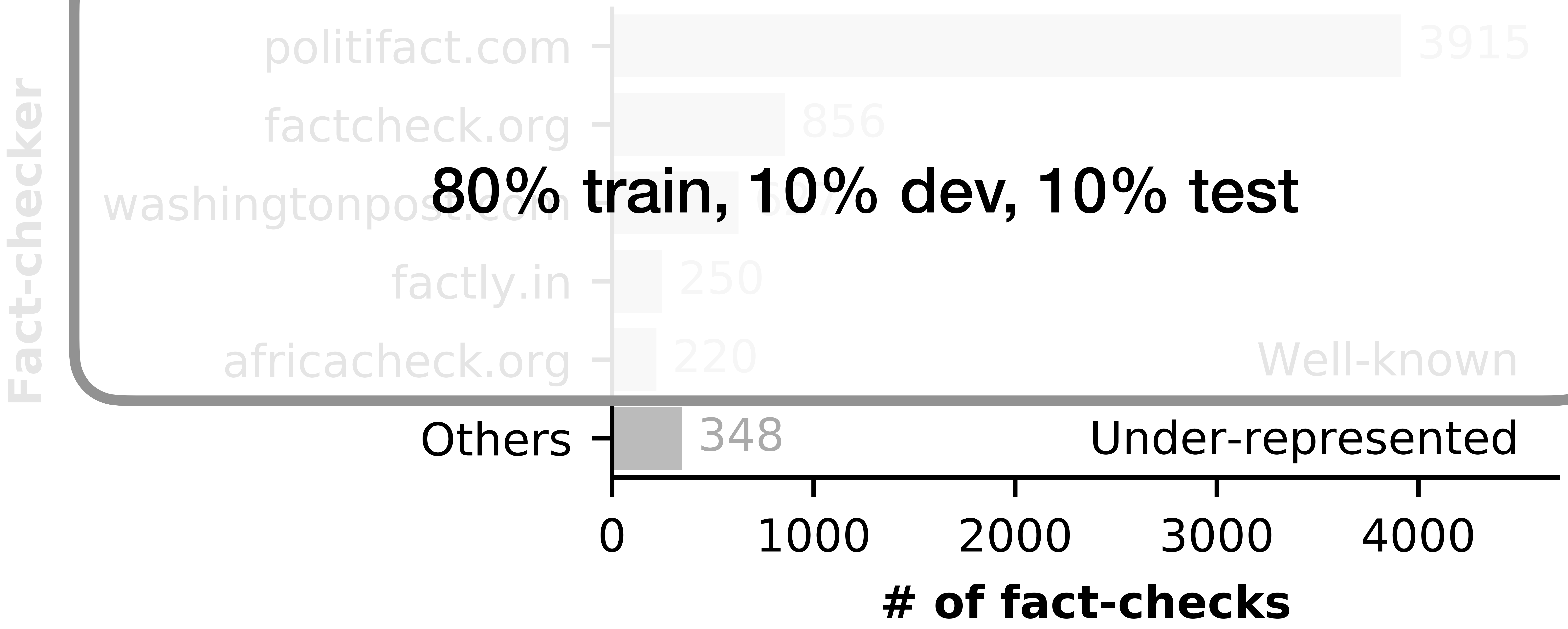
Experiments: previous observation



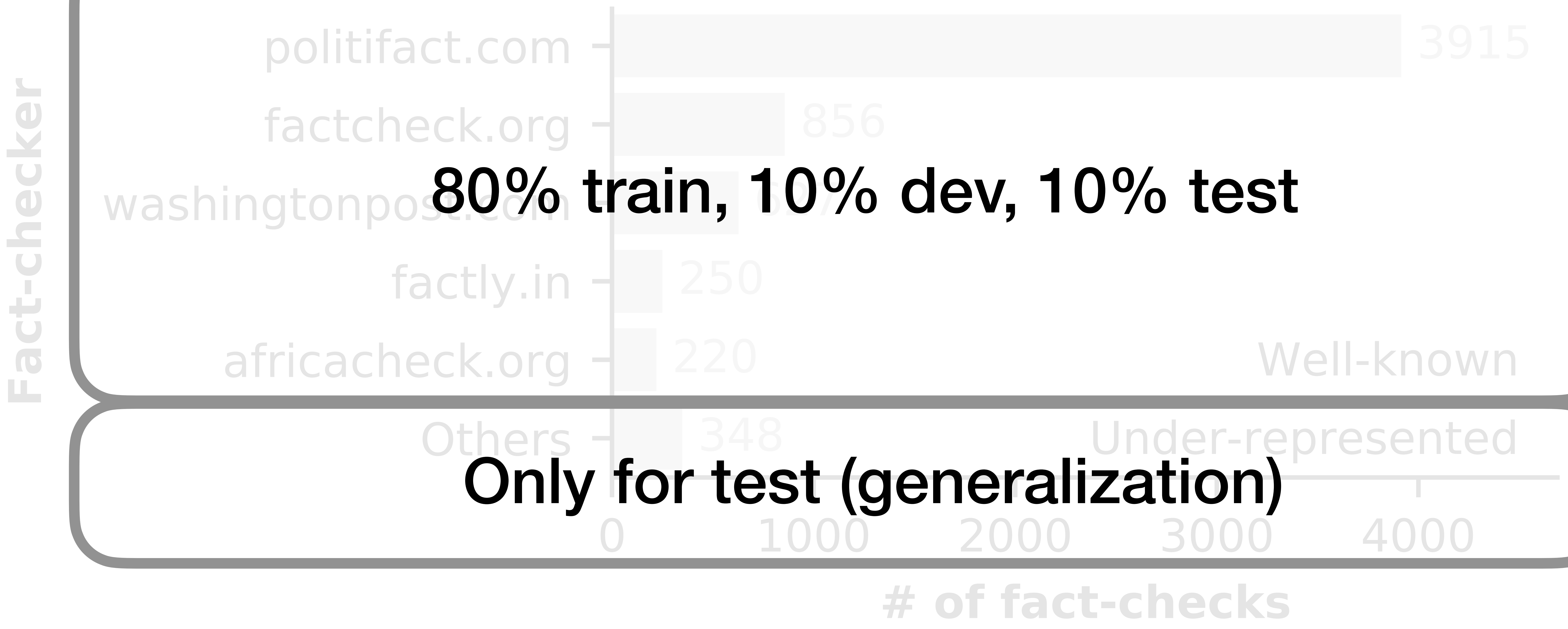
Power-law distribution of fact-checkers.

Experiments: data splitting

80% train, 10% dev, 10% test



Experiments: data splitting



80% train, 10% dev, 10% test

Only for test (generalization)

Experiments: evaluation

- ROUGE (F1, precision, recall)

- Tight score: if not tagged, ROUGE = 0.
- Loose score: only count if tagged.
- In a cell: tight score (loose score)

Results: overall performance

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

Results: overall performance

Lead token	Tagger	Claim ROUGE-1 F1	Claim ROUGE-1 Precision	Claim ROUGE-1 Recall	Claimant ROUGE-1 F1	Claimant ROUGE-1 Precision	Claimant ROUGE-1 Recall	Verdict ROUGE-1 F1	Verdict ROUGE-1 Precision	Verdict ROUGE-1 Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Poor performance of baseline methods.

Results: overall performance

Lead token	Tagger	**Claim** ROUGE-1			**Claimant** ROUGE-1			**Verdict** ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Poor performance of baseline methods.
- Improved performance w/ vanilla BERT.

Results: overall performance

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Poor performance of baseline methods.
- Improved performance w/ vanilla BERT.
- Further improved performance w/ paragraph tokens.

Results: overall performance

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Claim: ~0.65 (~0.85).

Results: overall performance

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Claim: ~0.65 (~0.85).
- Claimant: ~0.8 (~0.9).

Results: overall performance

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.183 (.183)	.300 (.300)	.141 (.141)	.237 (.237)	.181 (.181)	.352 (.352)	.660 (.660)	.638 (.638)	.702 (.704)
[CLS]	Fluent	.636 (.853)	.669 (.897)	.633 (.850)	.769 (.894)	.803 (.934)	.759 (.883)	.931 (.975)	.934 (.979)	.930 (.974)
	Concise	.592 (.864)	.615 (.897)	.596 (.870)	.784 (.907)	.789 (.913)	.783 (.906)	.938 (.971)	.940 (.973)	.938 (.970)
Paragraph position	Fluent	**.638** (.854)	**.674** (.902)	**.637** (.853)	**.794** (.889)	**.821** (.919)	**.789** (.884)	**.940** (.978)	**.942** (.980)	**.939** (.978)
	Concise	**.646** (.866)	**.664** (.889)	**.652** (.873)	**.839** (.928)	**.852** (.943)	**.834** (.923)	**.941** (.975)	**.944** (.979)	**.940** (.974)

- Claim: ~0.65 (~0.85).
- Claimant: ~0.8 (~0.9).
- Verdict: ~0.94 (~0.97).

Results: generalization

Lead token	Tagger	**Claim** ROUGE-1			**Claimant** ROUGE-1			**Verdict** ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.175 (.175)	.372 (.372)	.122 (.122)	.132 (.132)	.114 (.114)	.204 (.204)	.392 (.392)	.385 (.385)	.409 (.409)
[CLS]	Fluent	.444 (.725)	.483 (.788)	.443 (.724)	.264 (.567)	.364 (.782)	.236 (.506)	**.429** (.806)	**.484** (.910)	**.421** (.792)
	Concise	.386 (.713)	.406 (.748)	.406 (.749)	.323 (.650)	.379 (.764)	.304 (.612)	.451 (.832)	.484 (.892)	.446 (.821)
Paragraph position	Fluent	**.519** (.728)	**.566** (.794)	**.517** (.725)	**.377** (.635)	**.510** (.859)	**.342** (.576)	.367 (.733)	.451 (.902)	.359 (.718)
	Concise	**.527** (.738)	**.532** (.744)	**.559** (.781)	**.462** (.709)	**.549** (.843)	**.436** (.670)	**.473** (.832)	**.520** (.914)	**.467** (.822)

- Deteriorated performance for under-represented fact-checkers.

Results: generalization

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.175 (.175)	.372 (.372)	.122 (.122)	.132 (.132)	.114 (.114)	.204 (.204)	.392 (.392)	.385 (.385)	.409 (.409)
[CLS]	Fluent	.444 (.725)	.483 (.788)	.443 (.724)	.264 (.567)	.364 (.782)	.236 (.506)	**.429** (.806)	**.484** (.910)	**.421** (.792)
	Concise	.386 (.713)	.406 (.748)	.406 (.749)	.323 (.650)	.379 (.764)	.304 (.612)	.451 (.832)	.484 (.892)	.446 (.821)
Paragraph position	Fluent	**.519** (.728)	**.566** (.794)	**.517** (.725)	**.377** (.635)	**.510** (.859)	**.342** (.576)	.367 (.733)	.451 (.902)	.359 (.718)
	Concise	**.527** (.738)	**.532** (.744)	**.559** (.781)	**.462** (.709)	**.549** (.843)	**.436** (.670)	**.473** (.832)	**.520** (.914)	**.467** (.822)

- Deteriorated performance for under-represented fact-checkers.
- BERT w/ paragraph tokens still performs the best.

Results: generalization

Lead token	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline		.175 (.175)	.372 (.372)	.122 (.122)	.132 (.132)	.114 (.114)	.204 (.204)	.392 (.392)	.385 (.385)	.409 (.409)
[CLS]	Fluent	.444 (.725)	.483 (.788)	.443 (.724)	.264 (.567)	.364 (.782)	.236 (.506)	**.429** (.806)	**.484** (.910)	**.421** (.792)
	Concise	.386 (.713)	.406 (.748)	.406 (.749)	.323 (.650)	.379 (.764)	.304 (.612)	.451 (.832)	.484 (.892)	.446 (.821)
Paragraph position	Fluent	**.519** (.728)	**.566** (.794)	**.517** (.725)	**.377** (.635)	**.510** (.859)	**.342** (.576)	.367 (.733)	.451 (.902)	.359 (.718)
	Concise	**.527** (.738)	**.532** (.744)	**.559** (.781)	**.462** (.709)	**.549** (.843)	**.436** (.670)	**.473** (.832)	**.520** (.914)	**.467** (.822)

- Claim: ~0.5 (~0.7) from ~0.65 (~0.85).
- Claimant: ~0.4 (~0.7) from ~0.8 (~0.9).
- Verdict: ~0.4 (~0.8) from ~0.94 (~0.97).

Results: improving generalization

Train set	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Well-known ones only	Fluent	**.519** (.728)	**.566** (.794)	**.517** (.725)	.377 (.635)	.510 (.859)	.342 (.576)	.367 (.733)	.451 (.902)	.359 (.718)
	Concise	**.527** (.738)	.532 (.744)	**.559** (.781)	.462 (.709)	.549 (.843)	.436 (.670)	.473 (.832)	.520 (.914)	**.467** (.822)
under-represented mixed	Fluent	.495 (.761)	.540 (.830)	.489 (.752)	**.550** (.717)	**.639** (.832)	**.528** (.688)	**.475** (.712)	**.573** (.859)	**.469** (.704)
	Concise	.519 (.782)	**.544** (.819)	.536 (.807)	**.575** (.781)	**.599** (.813)	**.581** (.789)	**.482** (.797)	**.562** (.931)	.464 (.768)

- Mix half of under-represented fact-checkers to train.

Results: improving generalization

Train set	Tagger	Claim ROUGE-1			Claimant ROUGE-1			Verdict ROUGE-1		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Well-known ones only	Fluent	**.519** (.728)	**.566** (.794)	**.517** (.725)	.377 (.635)	.510 (.859)	.342 (.576)	.367 (.733)	.451 (.902)	.359 (.718)
	Concise	**.527** (.738)	.532 (.744)	**.559** (.781)	.462 (.709)	.549 (.843)	.436 (.670)	.473 (.832)	.520 (.914)	**.467** (.822)
under-represented mixed	Fluent	.495 (.761)	.540 (.830)	.489 (.752)	**.550** (.717)	**.639** (.832)	**.528** (.688)	**.475** (.712)	**.573** (.859)	**.469** (.704)
	Concise	.519 (.782)	**.544** (.819)	.536 (.807)	**.575** (.781)	**.599** (.813)	**.581** (.789)	**.482** (.797)	**.562** (.931)	.464 (.768)

- Mix half under-represented fact-checkers to train.
- Improved performance for claimant and verdict.
- Similar results for tagging claim.

Results: error analysis

- Not tagging: unseen patterns.

e.g., long and unseen factors with explanations.

Results: error analysis

- Not tagging: unseen patterns.

e.g., long and unseen factors with explanations.

- Wrongly tagging: confusing patterns.

e.g., “(someone) claimed (…) on (date)” in a fact-check has a high likelihood of tagging as claim.

Results: error analysis

- **Not tagging: unseen patterns.**

e.g., long and unseen factors with explanations.

- **Wrongly tagging: confusing patterns.**

e.g., "(someone) claimed (…) on (date)" in a fact-check has a high likelihood of tagging as claim.

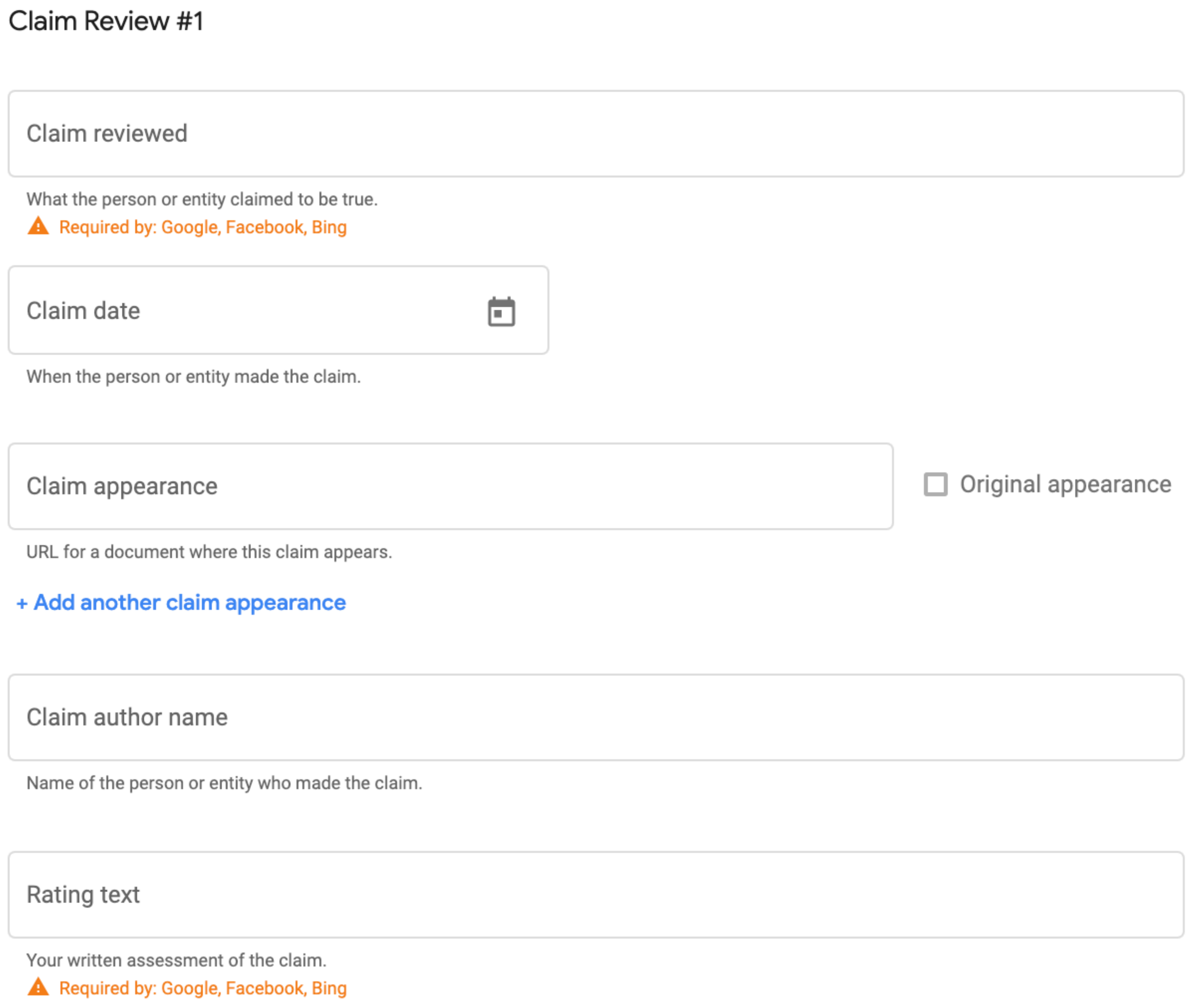
- **Partially tagging: unusual patterns.**

e.g., "the 45th and current president of the United States Donald Trump" as the claimant, our model tend to tag only "Donald Trump".

Application: pre-population

Pre-population the fact-check markup tool:

- Enter article URL.
- Pre-populating factors.
- Check, revise, submit.



Conclusion: takeaways

- Proposed *factoring fact-checks*.

Conclusion: takeaways

- Proposed *factoring fact-checks*.
- Observations from data exploration.

Conclusion: takeaways

- Proposed *factoring fact-checks*.
- Observations from data exploration.
- Applicable performance for well-known fact-checkers.

Conclusion: takeaways

- Proposed *factoring fact-checks*.
- Observations from data exploration.
- Applicable performance for well-known fact-checkers.
- Promising direction for under-represented fact-checkers.

Thank you!

Please send questions to: sjiang@ccs.neu.edu

Northeastern University

Google AI