NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge

(Reddy et al. 2022)

Presenter: Chong Shen

Argument Mining, Summer 2023

Eva Maria Vecchi
Institute for Natural Language Processing
University of Stuttgart

Table of Contents

Background and Motivation

The large amount of information emerging in news boosts the need to mitigate misinformation and disinformation.

- Fake news about newly emerged topics, e.g., COVID-19.
- *Echo Chamber:* People are exposed only to opinions that they agree with. (Orbach et al. 2020)
- Impact on decision making.

How to achieve this goal?

- Detect claims relevant to topics of interest;
- Select claims that are worth fact-checking;
- Identify factual / non-factual claims.

Limitations of previous work

1. Lacks knowledge of additional attributes (Jaradat et al. 2018)
   - News articles have complex arguments, requiring deeper understanding of
     - what *topic* the claim covers
     - which object the claim targets to (i.e. *claim object*)
     - where the claim comes from (i.e. *claimer* / *actor*)
     - what is the claimer's position (i.e. *stance*)
   - Crucial to assess whether a news article is worth fact-checking.

Limitations of previous work

1. Lacks knowledge of additional attributes (Jaradat et al. 2018)
2. Attributes (e.g., claimer) are organized in the dataset in a structured manner (Gencheva et al. 2017; Sundriyal et al. 2022)
   - News articles: Attributes are embedded in an unstructured way.
     - Not neccesarily before the claim sentence;
     - Mentioned in another sentence.

   $\Rightarrow$ Harder to identify.
   (See the examples on the next slide.)

Limitations of previous work: Examples

Claims in a debate (Gencheva et al. 2017):

Trump: **Hillary Clinton** attacked those same women and attacked them viciously.
Clinton: They're doing it to try to influence the election for **Donald Trump**.

Claims in a tweet (Sundriyal et al. 2022):

RT @PirateAtLaw: No no no. **Corona beer is the cure not the disease.**

We don't have evidence but we are positive **our wine keeps you from getting #COVID19** if you drink enough of it. **Better alternative to #DisinfectantInjection** don't you think? #winecures.

@adamseconomics **Vaccine is probably made from Chinese ingredients sourced in Wuhan.**

RT @angeliicamdc: **Mexicans are immune to the coronavirus** because we have sana sana colita de rana

Claims in a news article (Reddy et al. 2022):

News Text: With the coronavirus pandemic continuing to spread around the globe, people are panicked, and they're looking for answers and explanations. *One wild theory that has made its way around the web is that the virus came from space*. Recently, Chandra Wickramasinghe, known for his work in astronomy and astrobiology, spread the idea that the virus was living on a comet and a piece of that space rock may have fallen to Earth.

Topic: Origin of the virus
Stance: Affirm
Claim Object: space
Claimer: Chandra Wickramasinghe

Limitations of previous work

1. Lacks knowledge of additional attributes (Jaradat et al. 2018)
2. Attributes (e.g., claimer) are organized in the dataset in a structured manner (Gencheva et al. 2017; Sundriyal et al. 2022)
3. Built on fixed-size, pre-defined claim ontology (Padó et al. 2019)
   - Limited number of claim categories
     $\Rightarrow$ Limited capability of generalizing to newly emerged, previously unseen topics.

Hypotheses about the benefits of claim attributes

- The claim attributes enable comparing claims at a more fine-grained level (i.e. *claim-claim relations*)
  - Claims with the same topic, object and stance can be considered equivalent.
  - Claims with similar claim objects but opposing stance could be contradicting.
  - The claimer is useful to examine how current claims compare to previous ones by the same person/organization (*Claim Change Detection*).

Contributions of this work

1. **NewsClaims:** A benchmark for attribute-aware claim detection in the news.
2. Extended task: traditional claim detection + claim attributes extraction.
   - Lack of knowledge of claim attributes
     - Solution: Extract claim object, claimer, stance, claim span.
   - Unstructured embedding of claimer in news articles
     - Solution: Cross-sentence reasoning for claimer identification.
   - Bad generalization due to limited number of claim categories
     - Solution: zero-/few-shot claim detection.
3. Dataset: 889 annotated claims from 143 news articles.

Task Definition

- Given a news article, extract all claim sentences related to a set of topics, along with the corresponding attributes, including
  1. **claim span:** the exact claim boundries in the text
  2. **claim object:** an entity that identifies what is being claimed w.r.t. the topic.
  3. **claimer:** the source of the claim, such as
     - an entity (e.g., person, organization), or
     - a publication (e.g., study, report, investigation).
  4. **stance of the claimer:** asserting (*affirm*) or refuting (*refute*) the claim

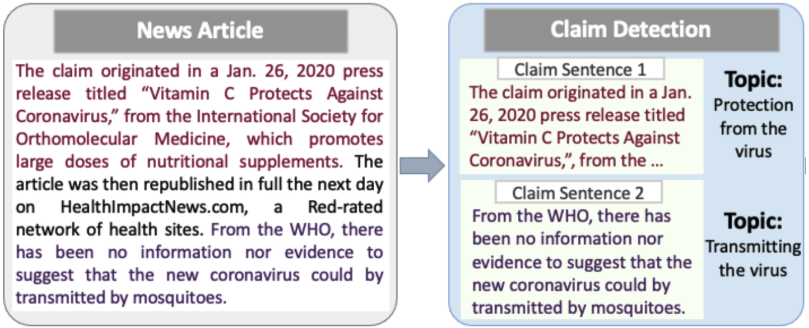
Subtasks

1. Claim Sentence Detection
2. Claim Object Detection
3. Stance Detection
4. Claim Span Detection
5. Claimer Detection

Methodology

Subtask 1: Claim Sentence Detection

- Input: a news article (i.e. document) + a set of pre-defined topics
- Output: claim sentences and the corresponding topics



Subtask 1: Claim Sentence Detection

- Two steps:
  - Step 1: *Check-worthiness Estimation:*
    Identify sentences containing factually verifiable claims.
    - Method: Apply ClaimBuster (Hassan et al. 2017)
    - Input: a news article
    - Output: Worth-checking claim sentences
  - Step 2: *Zero-shot topic classification:*
    Predict a topic label for the claim sentence.
    - Method: Fine-tune ClaimBuster + BART-large (Lewis et al. 2019) on *dev* set
    - Input: A check-worthy claim sentence as **premise** + a **hypothesis** constructed from each candidate topic
    - Output: predicted topic $\hat{t} = \arg\max_t score$

Subtask 1: Claim Sentence Detection

Template

Premise : < Claim Sentence >

Hypothesis : This is about < topic >

Example

Premise : There is no evidence that eating garlic prevents you from being infected with the coronavirus

Hypothesis 1 *(protection)* : This is about protection from the virus

Hypothesis 2 *(origin)* : This is about origin of the virus

Hypothesis 3 *(cure)* : This is about cure for the virus

Hypothesis 4 *(transmission)* : This is about transmission of the virus

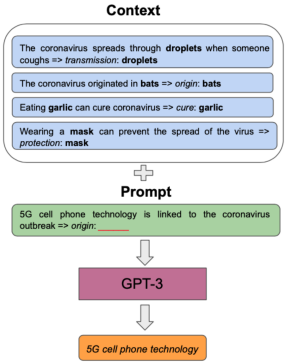
Subtask 2: Claim Object Detection

- Input: a claim sentence + its topic
- Output: claim object, i.e. a span in the claim sentence which refers to what is being claimed by the sentence w.r.t. the topic
- Three baselines:
  1. B1: few-shot prompt-based in-context learning
     - Method: Apply GPT-3 (Ada) (Brown et al. 2020)
  2. B2: few-shot prompt-based fine-tuning
     - Method: Fine-tune T5-base (Raffel et al. 2020) with Masked Language Modeling objective
  3. B3: zero-shot prompting
     - Method: GPT-3 (Ada) w/o. labeled examples, T5-base w/o. fine-tuning

Subtask 2: Claim Object Detection

Template for B1:

- Context: 4 few-shot examples, each from one of the 4 topics.
- Prompt: target claim sentence, with masked claim object to be generated.

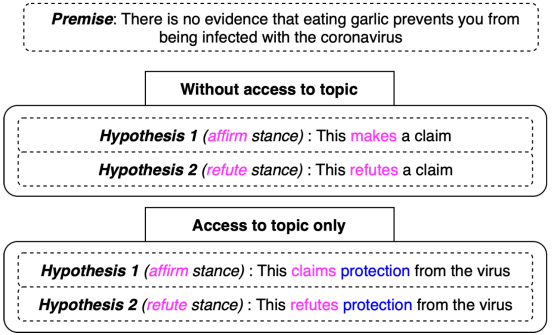


Subtask 3: Stance Detection

- Binary classification problem
- Input: a claim sentence (+ topic, optional)
- Output: a stance label from $\{\textit{affirm}, \textit{refute}\}$
- Two baselines:
  1. B1: Majority class baseline
     - Always predict *affirm*
  2. B2: zero-shot Natural Language Inference (NLI)
     - Method: Apply BART-large

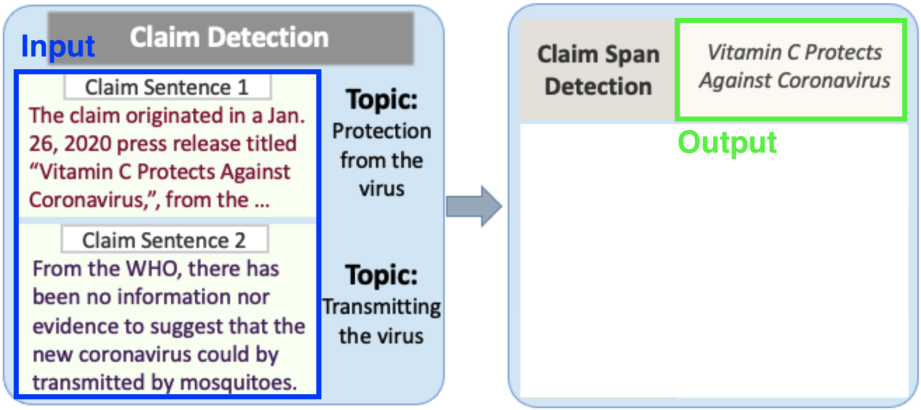
Subtask 3: Stance Detection

Template for B2: <Premise>+<Hypothesis>



Subtask 4: Claim Span Detection

- Input: a claim sentence
- Output: the exact claim boundaries within the claim sentence



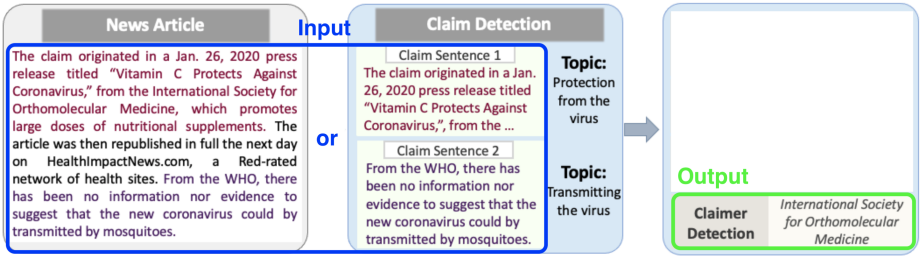
Subtask 4: Claim Span Detection

- Input: a claim sentence
- Output: the exact claim boundaries within the claim sentence
- Two baselines:
  1. B1: Debater Boundary Detection
     - Method: Apply Claim Boundry Detection function of Project Debater (BERT-large) (Slonim et al. 2021)[1]
  2. B2: PolNeAR-Context (Newell, Margolin, and Ruths 2018)
     - Method: Fine-tune BERT-large (Devlin et al. 2018) with a *start classifier* and an *end classifier*.

[1] **https://early-access-program.debater.res.ibm.com/terms?**.

Subtask 5: Claimer Detection

- Input: a news article or a claim sentence
- Output: a span representing the claimer + {*Journalist*, *Reported Source*}
  (The input and output formats depend on the baseline. See the next slide.)

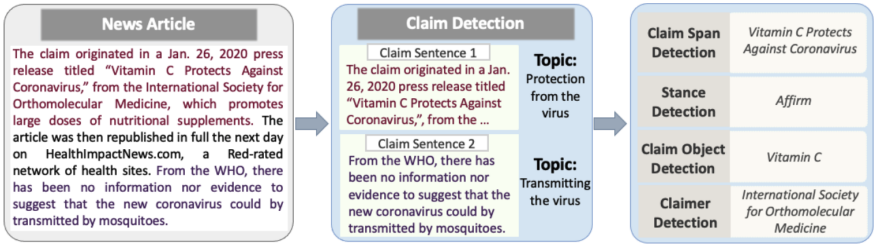


Subtask 5: Claimer Detection

- Two baselines:
  1. B1: PolNeAR-Source (Slonim et al. 2021)
     - Method: Fine-tune BERT-large with a *start classifier* and an *end classifier*
     - Input: a news article with claim span marked by special tokens
     - Output: a claimer span + *Journalist* / *Reported Source*
  2. B2: Semantic Role Labeling (SRL)
     - Method: Apply AllenNLP SRL Parser (Gardner et al. 2018)
     - Input: a claim sentence
     - SRL parser: (*ARG* 0, *VERB*, *ARG* 1)
     - Output: span corresponding to *ARG* 0, or *Journalist*

Illustration of the entire task

- Input: a news article
- Output: claim sentence, claim span, claim stance, claim object, claimer.



Dataset

News articles about the COVID-19 pandemic.

Four pre-defined topics (with *topic label*):

1. origin of the virus (*Origin*)
2. possible cure for the virus (*Cure*)
3. transmission of the virus (*Transmission*)
4. protecting against the virus (*Protection*)

No training set due to the zero-/few-shot setting.

dev set: 18 news articles with 103 annotated claims.

test set: 125 news articles with 786 annotated claims.

In total: 143 news articles with 889 annotated claims.

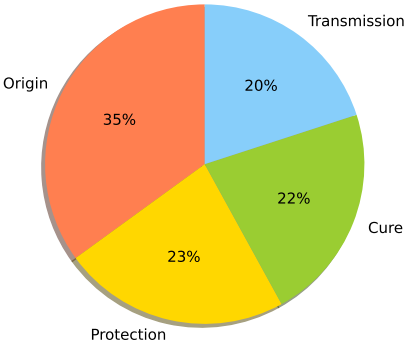
Dataset Annotation

For a news article,

1. identify claim sentences with corresponding topics
   - given a target sentence highlighted in red
   - whether this sentence contain a claim associated with the four topics
   - 3 annotators per example
   - sentences with unanimous support as valid claims
2. identify attributes for the claim sentences
   - given the entire article with a claim sentence highlighted
   - identify the attributes (i.e. claim span, claim object, claimer, stance)
   - no claimer mentioned in the article $\Rightarrow$ *journalist*
   - 1 annotator per example

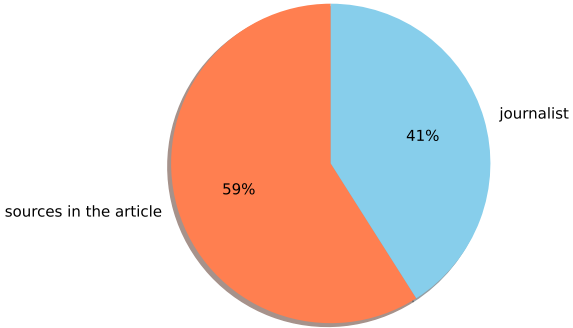
Dataset Analysis

- Topic distribution



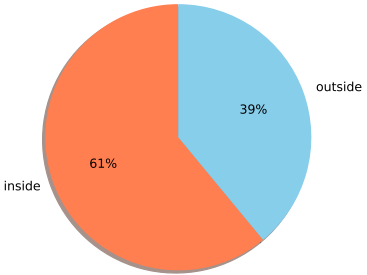
Dataset Analysis

- Claimer distribution



Dataset Analysis

- Claimer position distribution
  (whether the claimer is mentioned in the claim sentence)



Results and Analysis

Subtask 1: Claim Sentence Detection

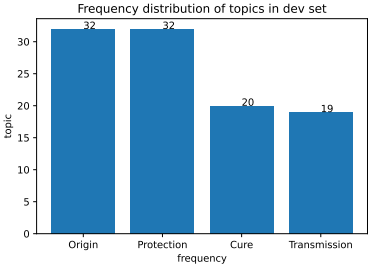
- Evaluation metrics: Presision (P), Recall (R), F1-Score (F1)
- ClaimBuster only (i.e. w/o. topic detection):
  - low precision (13.0%), high recall (86.5%)
  - F1: 22.6%
- ClaimBuster + zero-shot NLI model (i.e. w. topic detection):
  - zero-shot NLI model (i.e. BART-large) filters out claims with irrelevant topic
  - improved precision (21.8%) and F1 (31.9%)
- Conclusion: Knowledge of the claim topic helps improve the performance.

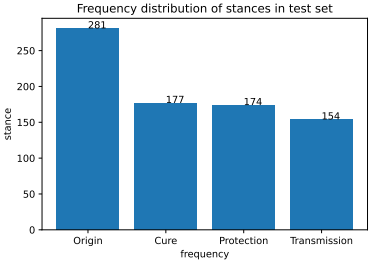
Model	P	R	F1
ClaimBuster	13.0	**86.5**	22.6
ClaimBuster + Zero-shot NLI	**21.8**	53.3	**30.9**
Human (single)	52.7	70.0	60.1
Human (3-way majority voting)	60.2	83.5	70.0

Table 2: Performance (in %) for various systems for detecting claims related to COVID-19.

Subtask 1: Claim Sentence Detection

My Analysis: Distribution of topics in *dev* and *test* sets





After fine-tuning, the model learns to correctly identify more claim sentences with the topic *Origin* and *Protection*. $\Rightarrow$ higher precision

Subtask 2: Claim Object Detection

- Evaluation metric: string-match F1 (Rajpurkar et al. 2016)
- Zero-shot setting:
  - GPT-3 $>$ T5-base
- Few-shot setting:
  - T5-base is comparable with GPT-3

Approach	Model	Type	F1
Prompting	GPT-3	Zero-shot	15.2
Prompting	T5	Zero-shot	11.4
In-context learning	GPT-3	Few-Shot	**51.9**
Prompt-based fine-tuning	T5	Few-Shot	51.6
Human	-	-	67.7

Table 3: F1 score (in %) for various zero-shot and few-shot systems for the claim object detection sub-task.

Subtask 3: Stance Detection

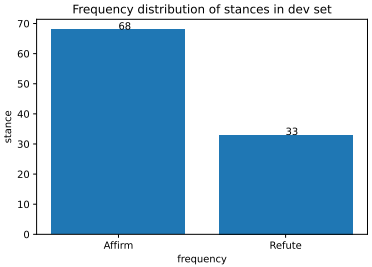
- Evaluation metrics: F1, Accuracy
- NLI (with topic) $>$ NLI (no topic) $>$ Majority Class
- The NLI model (i.e. BART-large) with access to the topic information performs the best, with significant F1 improvement for the *refute* class (68.0% $\rightarrow$ 78.8%).
- Conclusion: Access to the claim topic information helps improve the performance of stance detection.

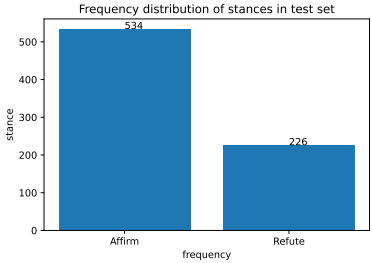
Model	Affirm F1	Refute F1	Acc.
Majority class	82.5	0.0	70.3
NLI (no topic)	89.1	68.0	83.8
NLI (with topic)	**91.1**	**78.8**	**87.5**
Human	97.0	84.2	94.9

Table 4: F1 score (in %) for *affirm* and *refute* classes along with the overall accuracy for stance detection. Zero-shot NLI is shown separately based on access the topic while constructing the hypothesis.

Subtask 3: Stance Detection

My Analysis: Distribution stance labels in *dev* and *test* sets





More data with *Affirm* in *dev* set, but does not help as much as the data with *Refute*, which are far less.

Subtask 4: Claim Span Detection

- Evaluation metric: character-span F1
- Project Debater > PolNeAR-Content (Newell, Margolin, and Ruths 2018)
- Project Debater: trained on arguments (more similar to claims); PolNeAR-Content: statements (less similar to claims).
- Conclusion: Models trained on data with similar structure to the claim perform better.

Model	Prec.	Recall	F1
PolNeAR-Content	67.0	42.8	52.3
Debater Boundary Detection	**75.7**	**77.7**	**76.7**
Human	82.7	90.9	86.6

Table 5: Performance (in %) of different systems for identifying the boundaries of the claim.

Subtask 5: Claimer Detection

- Evaluation metrics:
  - string-match F1 for the *Reported* claimer class
  - classification F1 for the *Journalist* claimer class
  - overall F1
- Reported claims: PolNeAR-Source > SRL, but both low in absolute
- Journalist: SRL > PolNeAR-Source, both higher

Model	F1	Reported	Journalist
SRL	41.7	23.5	**67.2**
PolNeAR-Source	**42.3**	**25.5**	65.9
Human	85.8	81.3	88.9

Table 6: Claimer detection scores for journalist claims and for reported claims, along with the overall F1.

Subtask 5: Claimer Detection

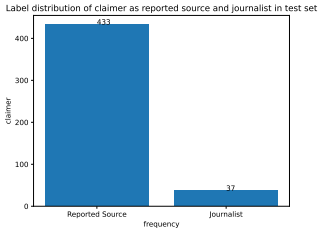
- Analysis for reported claims:
  - SRL parser only works with In-sentence claimers.
  - Both models cannot deal with cross-sentence claimer reasoning.

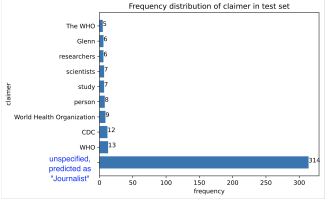
Model	In-sentence	Out-of-sentence
SRL	35.8	2.4
PolNeAR-Source	**38.9**	**2.7**

Table 7: F1 score (in %) in terms of reported claims for extracting the claimer when it is present within or outside the claim sentence.

Subtask 5: Claimer Detection

My Analysis: Claimer distribution in *test* set





- *Reported Source* contains all claimers except *Journalist*.
- Some claimer spans may be difficult to predict.
- Need analysis of the performance on each non-journalist class.
- Not enough examples of each non-journalist class in *test* set.

Conclusion

- Contribution of this paper: **NewsClaims**, a new evaluation benchmark consisting of
  - an extended task: attribute-aware claim detection in news articles
  - a dataset for zero-shot and few-shot settings
- Verifies the importance of various claim attributes in the proposed task, i.e.
  - claim span,
  - claim object,
  - claimer,
  - stance.
- Zero-shot and prompt-based few-shot approaches are promising for this task, but there is still a considerable gap between model and human performance.

Strengths and Weaknesses

Strengths:

1. Considers the problem of claim detection in the real world applications.
2. Provides a new benchmark with an annotated dataset for the research community.
3. Verifies their hypotheses with extensive experiments.
4. Wide range of application scenrios: detecting mis- & disinformation from news with unseen topics and limited amount of data.

Weaknesses:

1. Limited range of topics (i.e. 4), limited number of examples for non-journalist claimers makes the analysis difficult.
2. More attribute possible (e.g., sentiment, emotion, time etc.).
3. Bias in the news articles and of the human annotator, e.g., *Journalist* as representation of all unspecified claimer.

Future Work

- Increase the size of the test set and the range of the topic.
- Increase the number of examples for each non-journalist claimer class.
- Improve the annotation quality.
- News are not only texts, but multimodal. Images / Videos may contain more detailed, contextualized information than text alone.
- Other ways to represent the data, e.g., knowledge graph, which can better model the claim-claim relation, claimer-claimer relation, etc.

Thank you for listening!
Questions?

References I

- Brown, Tom et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.
- Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.
- Gardner, Matt et al. (2018). “Allennlp: A deep semantic natural language processing platform”. In: *arXiv preprint arXiv:1803.07640*.
- Gencheva, Pepa et al. (2017). “A context-aware approach for detecting worth-checking claims in political debates”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 267–276.
- Hassan, Naeemul et al. (2017). “Claimbuster: The first-ever end-to-end fact-checking system”. In: *Proceedings of the VLDB Endowment* 10.12, pp. 1945–1948.

References II

- Jaradat, Israa et al. (2018). “ClaimRank: Detecting check-worthy claims in Arabic and English”. In: *arXiv preprint arXiv:1804.07587*.
- Lewis, Mike et al. (2019). “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. In: *arXiv preprint arXiv:1910.13461*.
- Newell, Edward, Drew Margolin, and Derek Ruths (2018). “An attribution relations corpus for political news”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Orbach, Matan et al. (2020). “Out of the echo chamber: Detecting countering debate speeches”. In: *arXiv preprint arXiv:2005.01157*.
- Padó, Sebastian et al. (2019). “Who sides with whom? towards computational construction of discourse networks for political debates”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2841–2847.

References III

- Raffel, Colin et al. (2020). “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *The Journal of Machine Learning Research* 21.1, pp. 5485–5551.
- Rajpurkar, Pranav et al. (2016). “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250*.
- Reddy, Revanth Gangi et al. (2022). “Newsclaims: A new benchmark for claim detection from news with attribute knowledge”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6002–6018.
- Slonim, Noam et al. (2021). “An autonomous debating system”. In: *Nature* 591.7850, pp. 379–384.
- Sundriyal, Megha et al. (2022). “Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter”. In: *arXiv preprint arXiv:2210.04710*.