A Review of Fact-Checking, Fake News Detection and Argumentation

Tariq Alhindi

March 02, 2020

Outline

1. Introduction
2. Fact-Checking
3. Fake News Detection
4. Argumentation

Outline

1. Introduction
2. Fact-Checking
   a. What processes does fact-checking include and can they be automated?
   b. What sources can be used as evidence to fact-check claims?
3. Fake News Detection
4. Argumentation

Outline

1. Introduction
2. Fact-Checking
3. Fake News Detection
   a. What are the linguistic aspects of Fake News? Can it be detected without external sources?
   b. How do we build robust AI models that are resilient against false information?
4. Argumentation

Outline

1. Introduction
2. Fact-Checking
3. Fake News Detection
4. Argumentation
    a. How can we extract an argument structure from unstructured text?
    b. How can we use argumentation for misinformation detection?

Motivation for Automating Fact-Checking

Thorne et al. (2018b)

- Why the need to automate fact-checking?
  - Information readily available online with <u>no traditional editorial process</u>
  - <u>False</u> Information tend to <u>spread faster</u>
- Fact-checking in journalism, given a claim: few hours-few days
  - Evaluate previous speeches, debates, legislations, published figures or known facts **<u>Evidence Retrieval</u>**
  - Combine step 1 with reasoning to reach a verdict **<u>Textual Entailment</u>**
- Automatic fact-checking
  - Different task formulations: <u>fake news, stance, and incongruent headline</u> detection
  - Many datasets; most distinguishing factor is **<u>the use of evidence</u>**

James Thorne and Andreas Vlachos. "Automated Fact Checking: Task Formulations, Methods and Future Directions." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3346-3359. 2018.

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Stance Detection

Fake News and Fact-Checking Datasets

Dataset	Source	Size	Input	Output	Evidence
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news	74k	Claim	6 truth levels	None
FakeNewsAMT, Celebrity Pérez-Rosas et al. (2018)	News	480, 500	News article (excerpt)	ture, false	None
LIAR (Wang, 2017)	Politifact	12.8k	Claim	6 truth levels	Metadata
Community Q/A Nakov et al. (2016)	Community forums (Q/A)	88 question 880 threads	question, thread	Q: relevant, not C: good, bad	Discussion Threads
Perspective (Chen et al., 2019)	Debate websites	1k claims 10k perspect	claim	perspective, evidence, label	Debate websites
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Claim, Article headline	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50k	Headline, Article body	agree, disagree, discuss, unrelated	News Articles
FEVER (Thorne et al., 2018a)	Synthetic	185k	Claim	Sup, Ref, NEI	Wikipedia

Fact-Checking

Wikipedia as Evidence	Other Sources of Evidence
Thorne et al. (2018a)	Wang (2017)
Malon (2018)	Joty et al. (2018)
Nie et al. (2019)	Chen et al. (2019)
Zhou et al. (2019)	
Schuster et al. (2019)	

Fact-Checking

Wikipedia as Evidence	Other Sources of Evidence
Thorne et al. (2018a)	Wang (2017)
Malon (2018)	Joty et al. (2018)
Nie et al. (2019)	Chen et al. (2019)
Zhou et al. (2019)	
Schuster et al. (2019)	

Fact Extraction and VERification (FEVER)

Thorne et al. (2018a)

Goal: Provide a large-scale dataset
Data: Synthetic Claims and Wikipedia Documents
Method:

Document Retrieval	DrQA-TFIDF
Sentence Selection	TFIDF
Textual Entailment	Decomposable Attention

Supports, Refutes, NotEnoughInfo

(+) Providing a dataset for training ML models
(-) Synthetic data, does not necessarily reflect realistic fact-checked claims

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

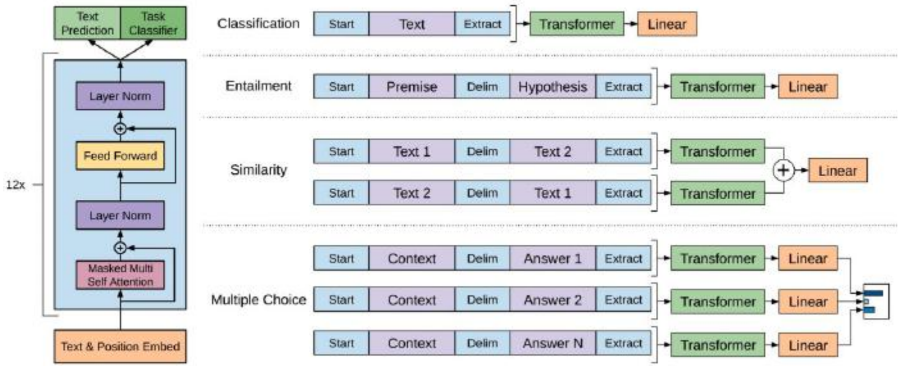
Thorne, James, et al. "FEVER: a Large-scale Dataset for Fact Extraction and VERification." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.

Transformers for Fact-Checking

Malon (2018)

Goal: Evidence Retrieval and Claim Verification
Data: FEVER
Method:
Doc. Ret. TFIDF, Named-Entities, Capitalization
Sent. Sel. TFIDF
Entailment Fine-Tuned OpenAI Transformer
Prepending with page title, individual evidence

(+) High Precision Model
(-) Imbalance towards NEI, Favoring Sup.
No handling of multi-sentence evidence



Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. Proceedings of the 1st Workshop on Fact Extraction VERification (FEVER).
Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

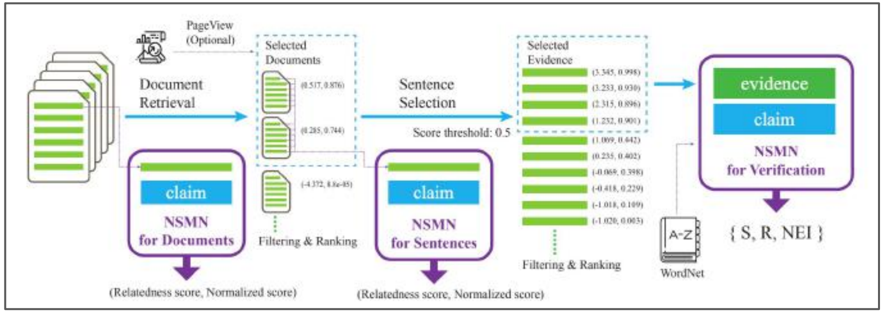
Neural Semantic Matching Networks (NSMN)

Nie et al. (2019)

Goal: Evidence Retrieval and Claim Verification
Data: FEVER
Method:

Doc. Ret.	keyword match, NSMN to filter & rank
Sent. Sel.	NSMN to filter & rank
RTE	NSMN over Glove & ELMo WordNet, numbers features

(+) Deep semantics modeling; Rich features
(-) Simple keyword match for Initial list of document candidates



Claim: Nicholas Brody is a character on Homeland.
Retrieved Evidence:
[wiki/Homeland]
Homeland is the first novel in The Dark Elf Trilogy, a prequel to The Icewind Dale Trilogy, written by R. A. Salvatore and follows the story of Drizzt Do'Urden from the time and circumstances of his birth and his upbringing amongst the drow (dark elves).

[wiki/Nicholas_Brody]
GySgt. Nicholas "Nick" Brody, played by actor Damian Lewis , is a fictional character on the American television series Homeland on Showtime, created by Alex Gansa and Howard Gordon.

Label: Support

Nie, Yixin, Haonan Chen, and Mohit Bansal. "Combining fact extraction and verification with neural semantic matching networks." In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

Modeling Evidence-Evidence Relations

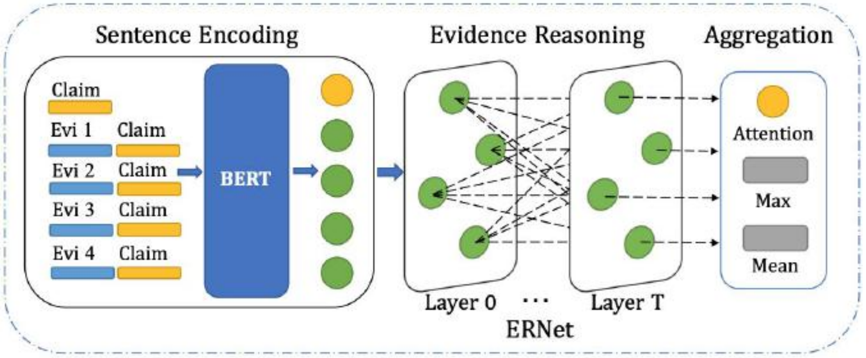
Zhou et al. (2019)

Goal: Evidence Retrieval and Claim Verification
Data: FEVER
Method:

Doc. Ret.	NPs in MediaWiki API	(UKP)
Sent. Sel.	ESIM-based Ranking	(UKP)
Entailment	Graph-based multi-evidence handling	

(+) Modeling of evidence-evidence relations
(-) No explicit modeling of evidence page info
No real effect of aggregator approaches

“REFUTED” Example	
Claim	Giada at Home was only available on DVD.
Evidence	(1) ***Giada at Home*** is a television show and first ***aired*** on October 18, 2008, ***on the Food Network***. (2) ***Food Network*** is an American ***basic cable and satellite television channel***.



Claim Verification (GEAR)

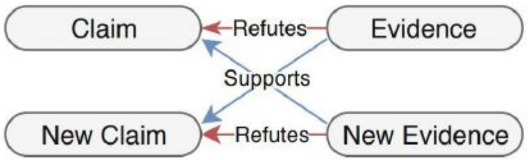
Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. "GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 892-901. 2019.

Bias in Fact-Checking Datasets

Schuster et al. (2019)

Goal: Bias Detection in fact-checking datasets
Data: FEVER + new test set
Method: Regularization to remove bias
Features: claim n-grams & labels correlation

(+) Better eval. of claim-evidence reasoning
Reweighting training objective
(-) No debiasing during training
Manual process



(A) ORIGINAL pair from the FEVER dataset

Claim:
Stanley Williams stayed in Cuba his whole life.
Evidence:
Stanley [...] was part of the West Side Crips, a street gang which has its roots in South Central Los Angeles.

(B) Manually GENERATED pair

Claim:
Stanley Williams moved from Cuba to California when he was 15 years old.
Evidence:
Stanley [...] was born in Havana and didn't leave the country until he died.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. "Towards debiasing fact verification models." Proceedings of the 2019 Conference on Empirical Methods in Natural Language.

FEVER-based models

Paper	Approach	Evidence Precision	Evidence Recall	Evidence F1	Label Accuracy	FEVER score
Malon (2018)	OpenAI Transformer Individual evidence modeling	**92.18**	50.02	**64.85**	61.08	57.36
Nei et al. (2019)	Semantic Matching Networks	42.27	70.91	52.96	68.21	64.21
Zhou et al. (2019)	Evidence-Evidence Modeling	23.61*	**85.19***	36.87	**71.60**	**67.10**

*UKP numbers

<u>Other works:</u>

Hidey et al. (2020)	BERT + Ptr Network	23.92	<u>88.39</u>	37.65	72.47	68.80
Soleimani et al. (2019)	BERT + pairwise loss	--	--	38.61	71.86	69.66
Zhong et al. (2019)	XLNet + graphs	--	--	39.45	<u>76.85</u>	<u>70.60</u>

Towards Realistic Fact-Checking

Types

Multiple propositions
- CONJUNCTION
- MULTI-HOP REASONING

Temporal reasoning
- DATE MANIPULATION
- MULTI-HOP TEMPORAL REASONING

Ambiguity and lexical variation
- ENTITY DISAMBIGUATION
- LEXICAL SUBSTITUTION

Examples

- MULTI-HOP REASONING
  - The Nice Guys is a 2016 action comedy film.
  - The Nice Guys is a 2016 action comedy film directed by a Danish screenwriter known for the 1987 action film Lethal Weapon.
- DATE MANIPULATION
  - in 2001 → in the first decade of the 21st century
  - in 2009 → 3 years before 2012
- LEXICAL SUBSTITUTION
  - filming -> shooting

FEVER-based models

Paper	Approach	Evidence Precision	Evidence Recall	Evidence F1	Label Accuracy	FEVER score	FEVER 2 adversarial
Malon (2018)	OpenAI Transformer Individual evidence modeling	**92.18**	50.02	**64.85**	61.08	57.36	**37.31**
Nei et al. (2019)	Semantic Matching Networks	42.27	70.91	52.96	68.21	64.21	30.47
Zhou et al. (2019)	Evidence-Evidence Modeling	23.61*	**85.19***	36.87	**71.60**	**67.10**	--

*UKP numbers

Other works:

Paper	Approach	Evidence Precision	Evidence Recall	Evidence F1	Label Accuracy	FEVER score	FEVER 2 adversarial
Hidey et al. (2020)	BERT + Ptr Network	23.92	<u>88.39</u>	37.65	72.47	68.80	36.61
Soleimani et al. (2019)	BERT + pairwise loss	--	--	38.61	71.86	69.66	--
Zhong et al. (2019)	XLNet + graphs	--	--	39.45	<u>76.85</u>	<u>70.60</u>	--

Fact-Checking

Wikipedia as Evidence	Other Sources of Evidence
Thorne et al. (2018a)	Wang (2017)
Malon (2018)	Joty et al. (2018)
Nie et al. (2019)	Chen et al. (2019)
Zhou et al. (2019)	
Schuster et al. (2019)	

Fact-Checking

Wikipedia as Evidence

Thorne et al. (2018a)

Malon (2018)

Nie et al. (2019)

Zhou et al. (2019)

Schuster et al. (2019)

Other Sources of Evidence

Wang (2017)
metadata

Joty et al. (2018)
community forums

Chen et al. (2019)
debates websites

LIAR LIAR

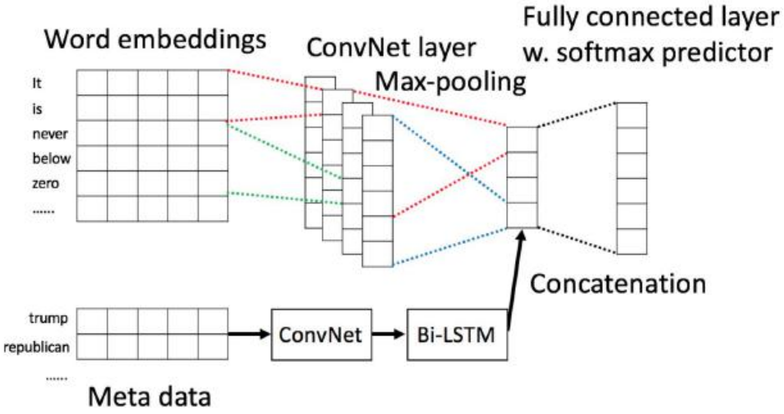
Wang (2017)

Goal: Provide a large-scale dataset
Data: Politifact.com
Method: BiLSTM + CNNs
Features: word embeddings, metadata

(+) New resource with speaker info and history
Multi-truth levels

(-) Single-domain dataset
No external evidence

Statement: *"The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero."*
Speaker: Donald Trump
Context: presidential announcement speech
Label: Pants on Fire



William Yang Wang "[“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection.]" In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422-426. 2017.

Fact-Checking in Community Q/A

Joty et al. (2018)

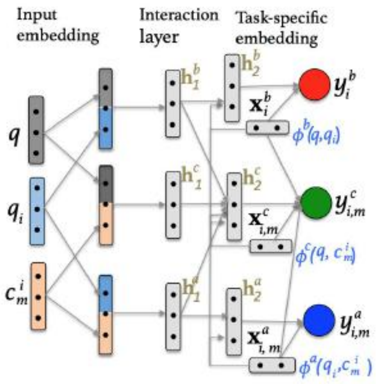
Goal: Finding relevant threads in community forums to a given question
Data: Community forums
Method: DNNs + CRF
Features: embeddings, cosine-similarity
MT features, question-comment lengths

(+) Joint modeling of all three subtasks
(-) CRF backpropagation does not update task-specific embeddings
All representations are pretrained

q: **"How can I extend a family visit visa?"**

q_i: "Dear All; I wonder if anyone knows the procedure how I can extend the family visit visa for my wife beyond 6 months. I already extended it for 5 months and is 6 months running. I would like to get it extended for couple of months more.Any suggestion is highly appreciable.Thanks"

c_m^i: "You can get just another month's extension before she completes 6 months by presenting to immigration office a confirmed booking of her return ticket which must not exceed 7 months."



Shafiq Joty, Lluís Màrquez, and Preslav Nakov. "[Joint Multitask Learning for Community Question Answering Using Task-Specific Embeddings.](#)" In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4196-4207. 2018.

Perspective

Chen et al. (2019)

Goal: “perspective” and evidence retrieval for a given claim

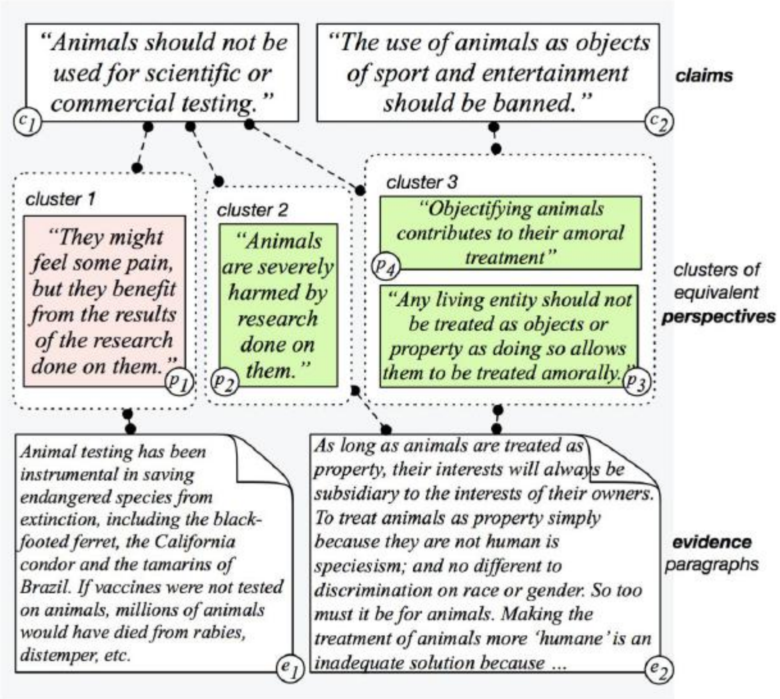
Data: debate websites

Method:

Off-the-shelf IR system + BERT

(+) Multi-level annotations: claim-perspective, perspective-perspective, and perspective-evidence

(-) Setup disconnected with the literature



Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. "Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics

Conclusion of Fact-Checking

What have we learned?

- What processes does fact-checking include and can they be automated?
  - <u>Evidence Retrieval</u> Document Retrieval, Sentence Selection
  - <u>Claim Verification</u> Textual Entailment

- What sources can be used as evidence to fact-check claims?
  - Wikipedia useful for entities with wiki-pages, and time insensitive claims
  - Metadata (speaker history) useful for some domains (e.g. politics)
  - Community Forums useful where official sources are lacking information/language
  - Debate websites useful for controversial topics

- However, fact-checking models are still not robust enough for open-domain fact-checking

Outline

1. Introduction
2. Fact-Checking
3. Fake News Detection
    a. What are the linguistic aspects of Fake News? Can it be detected without external sources?
    b. How do we build robust AI models that are resilient against false information?
4. Argumentation

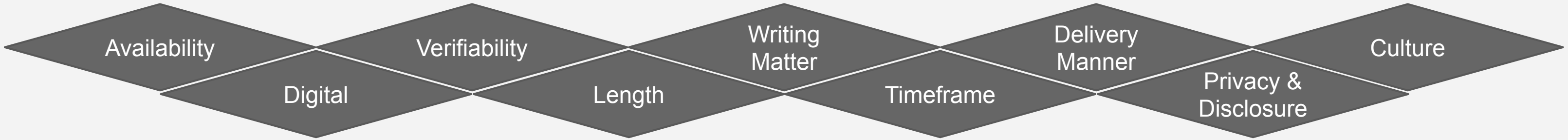
The Three Types of Fakes!

Rubin et al. (2015)

Serious Fabrications news items about false and non-existing events or information

Hoaxes providing false information via, for example, social media with the intention to be picked up by traditional news websites

Satire humorous news items that mimic genuine news but contain irony and absurdity



Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. "Deception detection for news: three types of fakes." In *Proceedings ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, p. 83. American Society for Information Science, 2015.

Fake News

Types of Fake News

Rashkin et al. (2017)

Pérez-Rosas et al. (2018)

Da San Martino et al. (2019)

Zellers et al. (2019)

Stance for Fake News Detection

Hanselowski et al. (2018)

Conforti et al. (2018)

Zhang et al. (2019)

Fake News

Types of Fake News

Rashkin et al. (2017)

Pérez-Rosas et al. (2018)

Da San Martino et al. (2019)

Zellers et al. (2019)

Stance for Fake News Detection

Hanselowski et al. (2018)

Conforti et al. (2018)

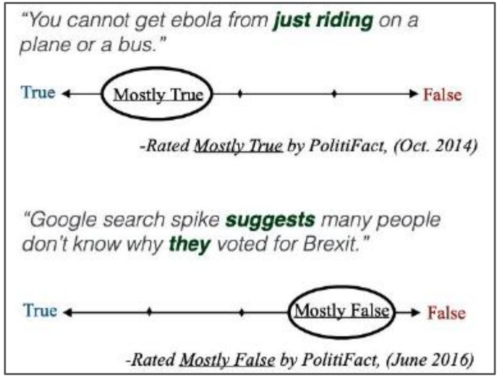
Zhang et al. (2019)

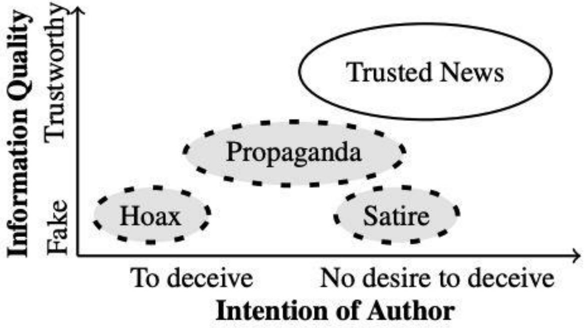
The Language of Fake News

Rashkin et al. (2017)

Goal: comparing language of real news with satire, hoaxes, and propaganda
Data: News websites and Politifact
Method: MaxEntropy, LSTM
Features: <u>TFIDF</u>, LIWC, sentiment, hedging comparative, suplaritives, adverbs. (Glove)

(+) Datasets with different types of fakes
Multiple truth levels
(-) Labeled at the publisher level
No theoretical foundation for the types





Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. "[Truth of varying shades: Analyzing language in fake news and political fact-checking.]" EMNLP 2017 (Short)

The Language of Fake News

Pérez-Rosas et al. (2018)

Goal: introducing two fake news datasets
Data: news articles
Method: SVM
Features: n-grams, LIWC, readability, syntax

(+) Corpora cover multiple domains
Cross-domain experiments

(-) No experiments with neural networks
No comparison with other existing datasets
Crawled True VS Crowdsourced Fake

FakeNewsAMT (Technology)

LEGITIMATE	FAKE
Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display.	**New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming.

Celebrity

LEGITIMATE	FAKE
Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, "Happy Holidays." In the picture, seemingly taken at Kris Jenner's annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, "It's been a very hard couple of months."	**Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West.** Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they've been getting close amid Kanye's mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn't appear to confirm or deny an affair, her reps said there is "no truth whatsoever" to the reports and labeled the situation "fabricated."

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic Detection of Fake News." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3391-3401. 2018.

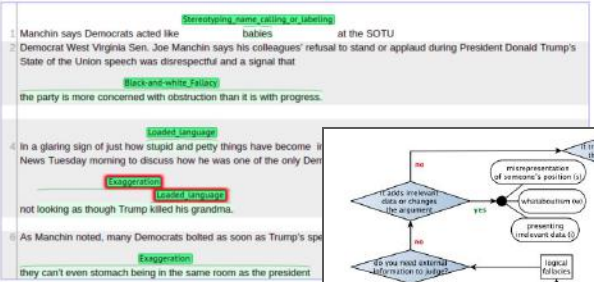
Propaganda

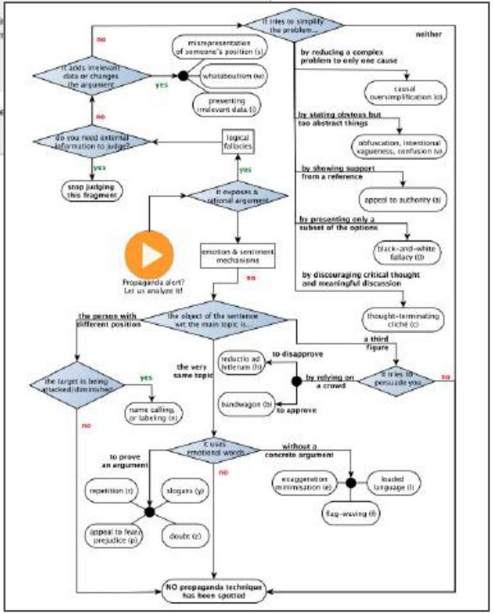
Da San Martino et al. (2019)

Goal: predict existence and type of propaganda
Data: news (450 articles)
Method: BERT fine-tuning

(+) Detailed annotation scheme
(18 techniques, compressed to 14 later)
Fine-grained annotation (fragment-level)

(-) Heavily imbalanced classes (15-2,500)





Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. "[Fine-Grained Analysis of Propaganda in News Articles.]" Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.

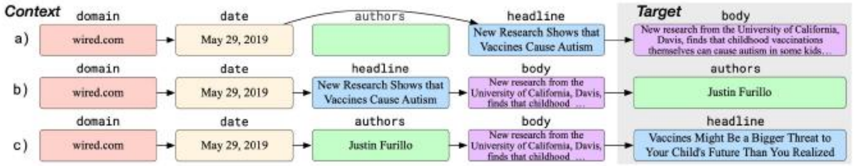
AI-Generated Fake News

Zellers et al. (2019)

Goal: Detect AI-generated fake text
Data: News articles
Method: Transformers (Generation & Detection)

(+) Large-scale model and training data
Machine text harder to detect by humans

(-) Labeled at the publisher level
Approached as Human vs Machine text
Assumes access to generative model
Less consistent with headlines



Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. "Defending Against Neural Fake News." In Advances in Neural Information Processing Systems, pp. 9051-9062. 2019.

A Second Look at Terminologies

News: verifiable information in the public interest

• Fake News	false or misleading verifiable information in the public interest
• Misinformation	information that is false but not created with the intention of causing harm.
• Disinformation	information that is false and deliberately created to harm.
• Propaganda	is a form of communication that attempts to further the desired intent of the propagandist.
○ In News	emphasizing positive features & downplaying negative ones to cast an entity in a favorable light.
• Hoax	providing false information with the intention to be picked up by traditional news websites.
• Satire	humorous news items that mimic genuine news but contain irony and absurdity.

'Fake news' is today so much more than a label for false and misleading information, disguised and disseminated as news. It has become an emotional, weaponized term used to undermine and discredit journalism. For this reason, the terms misinformation, disinformation and 'information disorder', are preferred.

Ireton, Cherilyn, and Julie Posetti. *Journalism, fake news & disinformation: Handbook for Journalism Education and Training*. UNESCO, 2018.
Jowett, Garth S., and Victoria O'Donnell. "What is propaganda, and how does it differ from persuasion." *Propaganda and Misinformation* (2006).

What are the linguistic aspects of Fake News?

- Rashkin et al. (2017)

  First-person and second-person pronouns are used more in less reliable.

  Subjectives, Superlatives, and Modal adverbs – are used more by fake news.

  Words used to offer concrete figures – comparatives, money, and numbers – appear more in truthful news.

  Trusted sources are more likely to use assertive words and less likely to use hedging words.

- Pérez-Rosas et al. (2018)

  Linguistic properties of deception in one domain *might be* structurally different from those in a second domain.

  Politics, Education, and Technology domains appear to be more robust against classifiers trained on other domains.

- Da San Martino et al. (2019)

  Propaganda has many techniques that have different lexical and structural properties.

  Reinforcing a sentence-level signal throughout the model is useful in detecting propaganda at the fragment level.

- Zellers et al. (2019)

  Humans are more vulnerable to machine-generated fakes than human-generated fakes.

  Neural models that are good fake-news generators are also good discriminators of human vs machine text.

Fake News

Types of Fake News	Stance for Fake News Detection
Rashkin et al. (2017) Pérez-Rosas et al. (2018) Da San Martino et al. (2019) Zellers et al. (2019)	Hanselowski et al. (2018) Conforti et al. (2018) Zhang et al. (2019)

Stance Detection for Fake News Detection

Types of Fake News

Rashkin et al. (2017)

Pérez-Rosas et al. (2018)

Da San Martino et al. (2019)

Zellers et al. (2019)

Stance for Fake News Detection

Hanselowski et al. (2018)

Conforti et al. (2018)

Zhang et al. (2019)

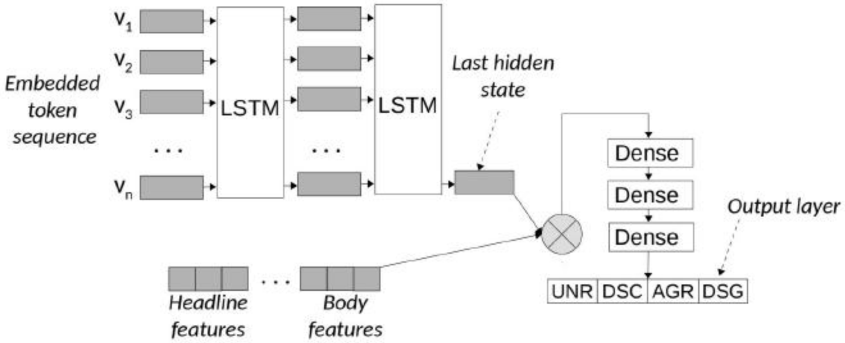
Joint Stance and Relatedness

Hanselowski et al. (2018)

Goal: Analysis of FNC-1 Results
Data: FNC-1 (News Articles)
Method: stacked LSTM
Features: structural, lexical, readability
Glove embeddings

(+) New evaluation measure that is not vulnerable to basic baselines
Testing on multiple datasets
(-) But no control for classes in cross-domain

Headline: Hundreds of Palestinians flee floods in Gaza as Israel opens dams	
Agree (AGR)	GAZA CITY (Ma'an) – Hundreds of Palestinians were evacuated from their homes Sunday morning after Israeli authorities opened a number of dams near the border, flooding the Gaza Valley in the wake of a recent severe winter storm. The Gaza Ministry of Interior said in a statement that civil defense services and teams from the Ministry of Public Works had evacuated more than 80 families from both sides of the Gaza Valley (Wadi Gaza) after their homes flooded as water levels reached more than three meters [..]
Discuss (DSC)	Palestinian officials say hundreds of Gazans were forced to evacuate after Israel opened the gates of several dams on the border with the Gaza Strip, and flooded at least 80 households. Israel has denied the claim as "entirely false". [..]
Disagree (DSG)	Israel has rejected allegations by government officials in the Gaza strip that authorities were responsible for released storm waters flooding parts of the besieged area. "The claim is entirely false, and southern Israel does not have any dams," said a statement from the Coordinator of Government Activities in the Territories (COGAT). "Due to the recent rain, streams were flooded throughout the region with no connection to actions taken by the State of Israel." At least 80 Palestinian families have been evacuated after water levels in the Gaza Valley (Wadi Gaza) rose to almost three meters. [..]
Unrelated (UNR)	Apple is continuing to experience 'Hairgate' problems but they may just be a publicity stunt [..]



Andreas Hanselowski, P. V. S. Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. "A Retrospective Analysis of the Fake News Challenge Stance-Detection Task." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1859-1874. 2018.

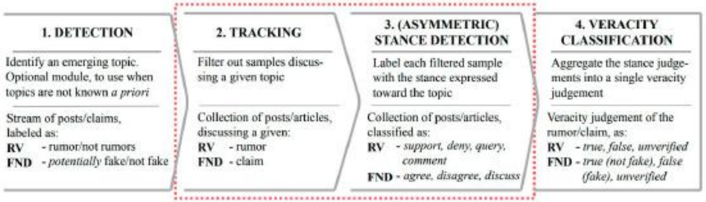
Stance (Related Classes Only)

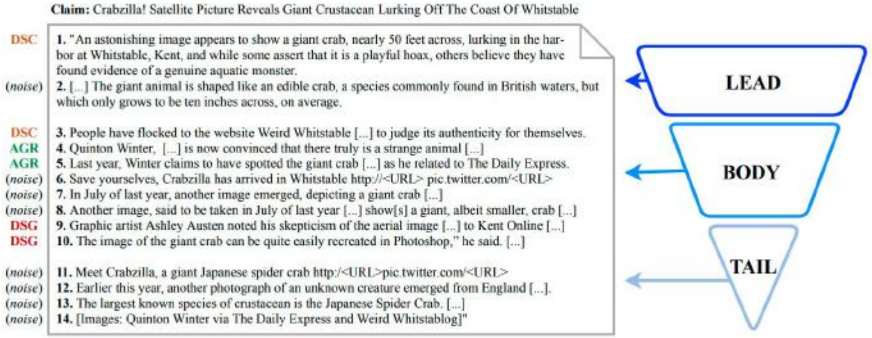
Conforti et al. (2018)

Goal: Headline-Article Stance
Data: FNC-1 (News Articles)
Method: Backward LSTM with attention
Features: word embeddings (word2vec), NEs

(+) Interpretable neural network architecture inspired by the Inverted Pyramid scheme

(-) Ignoring the ‘Unrelated’ class





Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. "[Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles.]" In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pp. 40-49. 2018.

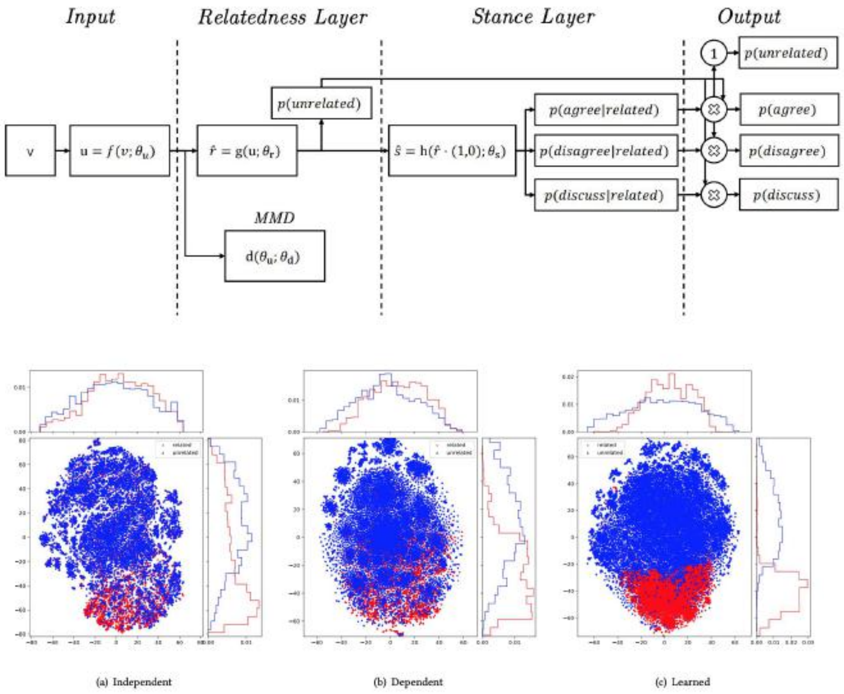
Relatedness then Stance

Zhang et al. (2019)

Goal: Claim/Headline-Article Stance
Data: FNC-1, and its seed dataset (Emergent)
Method: 2-layer Neural Network
with Maximum Mean Discrepancy
Features: TD-IDF, similarity, polarity

(+) Separate loss for relatedness and stance
Joint modeling with MMD regularization
Good performance on the minority class

(-) No use of static or contextual embeddings
Using FNC-1 original metric



Zhang, Qiang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. "From Stances' Imbalance to Their Hierarchical Representation and Detection." In *The World Wide Web Conference*, pp. 2323-2332. 2019.

Stance Detection Models

Paper	Approach	Agree	Disagree	Discuss	Unrelated	Macro F1	Weighted Accuracy
Hanselowski et al. (2018)	stacked LSTMs + handcrafted features	50.1	18.0	75.7	99.5	**60.9**	82.1
Conforti et al. (2018)	backward LSTM with attention	69.57	33.0	74.91	-	59.01*	-
Zhang et al. (2019)	2-layer NN with MMD regularization	80.61	**72.35**	77.49	99.53	-	**88.15**

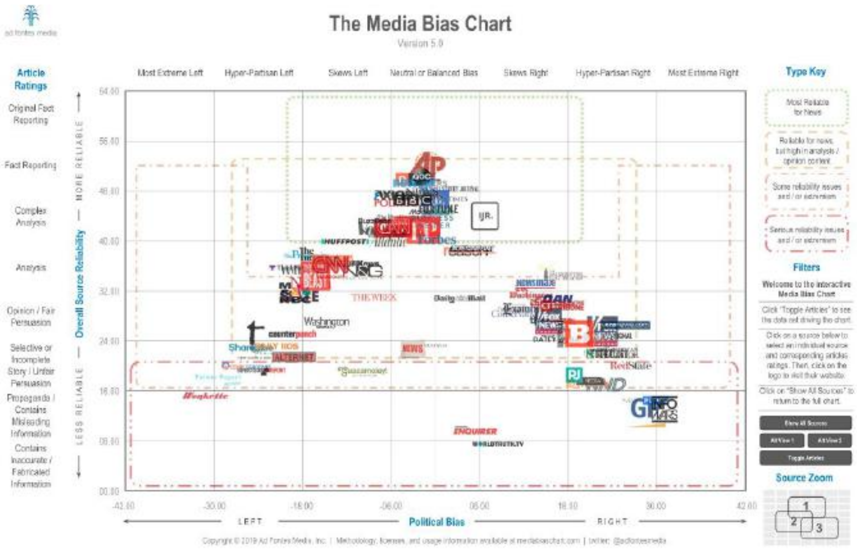
Other works:

Mohtarami et al. (2018)	Memory Networks	-	-	-	-	56.88	81.23
Dulhanty et al. (2019)	Fine-tuned RoBERTa	-	-	-	-	-	90.01
Schiller et al. (2020)	Multi-Task Deep Neural Network (MT-DNN) + BERT	-	-	-	-	76.90	88.82

Fact-Checking & Fake News Detection

How do we build robust AI models that are resilient against false information?

1. Many types of false information that have linguistic properties in some domains/genres
2. Stance Detection provides a macro-level view for Fake News Detection
3. Multi-truth levels: 6 (LIAR), 2-3 (FEVER)
4. Credibility of sources!
   Media Bias/Fact-check



Ad Fontes Media.
https://www.adfontesmedia.com/interactive-media-bias-chart/

Outline

1. Introduction
2. Fact-Checking
3. Fake News Detection
4. Argumentation
    a. How can we extract an argument structure from unstructured text?
    b. How can we use argumentation for misinformation detection?

Argumentation

Argument Structure

Peldszus and Stede (2015)
Potash et al. (2017)
Niculae et al. (2017)

Persing and Ng (2016)
Eger et al. (2017)

Claim Detection, Argument Semantics

Daxenberger et al. (2017)
Chakrabarty et al. (2019)

Hidey et al. (2017)
Wachsmuth et al. (2017)

Argumentation

Argument Structure

Peldszus and Stede (2015)
Potash et al. (2017)
Niculae et al. (2017)

Persing and Ng (2016)
Eger et al. (2017)

Claim Detection, Argument Semantics

Daxenberger et al. (2017)
Chakrabarty et al. (2019)

Hidey et al. (2017)
Wachsmuth et al. (2017)

Argumentation Pipeline

Tasks to Extract Argument Structure

- Segmentation
  - Argumentative vs Non-argumentative
  - Identification of argumentative discourse units (ADUs)
- ADU type classification: claim, premise
- Link identification
- Link type classification: support, attack

Andreas Peldszus and Manfred Stede. "Joint prediction in MST-style discourse parsing for argumentation mining."
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 938-948. 2015.

Argumentation Datasets

Dataset	Genre	Docs	Sent	Units	Relations
Peldszus and Stede (2015)	microtext (MT)	112	449	claim, premise	support, attack (rebuttal, undercut)
Stab and Gurevych (2017)	persuasive essays (PE)	402	7,116	major claim, claim, premise	support, attack
Niculae et al. (2017)	web discourse, eRuleMaking (CDCP)	731	~1.5k	policy, value, testimony, fact, reference	support (reason, evidence)
Reed et al. (2008)	AraucariaDB	507	2,842	claim, premise	-
Habernal and Gurevych (2015)	web discourse (WD)	340	3,899	claim, permise, backing, rebuttal refutation	
Biran and Rambow (2011a)	online comments (OC)	2,805	8,946	claim, justification	-
Biran and Rambow (2011b)	wiki talk pages (WTP)	1,985	9,140	claim, justification	-
Hidey et al. (2017)	reddit (CMV)	78	3,500	claim: interpret., eval., (dis)-agree; premise: logos, pathos, ethos	-
Habernal and Gurevych (2016)	debate websites (UKPConvArg)	32 topics	16k pairs	-	-

Argumentation Datasets

Dataset	Genre	Docs	Sent	Units	Relations
Peldszus and Stede (2015)	microtext (MT)	112	449	claim, premise	support, attack (rebuttal, undercut)
Stab and Gurevych (2017)	persuasive essays (PE)	402	7,116	major claim, claim, premise	support, attack
Niculae et al. (2017)	web discourse, eRuleMaking (CDCP)	731	~1.5k	policy, value, testimony, fact, reference	support (reason, evidence)
Reed et al. (2008)	AraucariaDB	507	2,842	claim, premise	-
Habernal and Gurevych (2015)	web discourse (WD)	340	3,899	claim, permise, backing, rebuttal refutation	
Biran and Rambow (2011a)	online comments (OC)	2,805	8,946	claim, justification	-
Biran and Rambow (2011b)	wiki talk pages (WTP)	1,985	9,140	claim, justification	-
Hidey et al. (2017)	reddit (CMV)	78	3,500	claim: interpret., eval., (dis)-agree; premise: logos, pathos, ethos	-
Habernal and Gurevych (2016)	debate websites (UKPConvArg)	32 topics	16k pairs	-	-

Argumentation Datasets

Dataset	Genre	Docs	Sent	Units	Relations
Peldszus and Stede (2015)	microtext (MT)	112	449	claim, premise	support, attack (rebuttal, undercut)
Stab and Gurevych (2017)	persuasive essays (PE)	402	7,116	major claim, claim, premise	support, attack
Niculae et al. (2017)	web discourse, eRuleMaking (CDCP)	731	~1.5k	policy, value, testimony, fact, reference	support (reason, evidence)
Reed et al. (2008)	AraucariaDB	507	2,842	claim, premise	-
Habernal and Gurevych (2015)	web discourse (WD)	340	3,899	claim, permise, backing, rebuttal refutation	
Biran and Rambow (2011a)	online comments (OC)	2,805	8,946	claim, justification	-
Biran and Rambow (2011b)	wiki talk pages (WTP)	1,985	9,140	claim, justification	-
Hidey et al. (2017)	reddit (CMV)	78	3,500	claim: interpret., eval., (dis)-agree; premise: logos, pathos, ethos	-
Habernal and Gurevych (2016)	debate websites (UKPConvArg)	32 topics	16k pairs	-	-

Argumentation Datasets

Dataset	Genre	Docs	Sent	Units	Relations
Peldszus and Stede (2015)	microtext (MT)	112	449	claim, premise	support, attack (rebuttal, undercut)
Stab and Gurevych (2017)	persuasive essays (PE)	402	7,116	major claim, claim, premise	support, attack
Niculae et al. (2017)	web discourse, eRuleMaking (CDCP)	731	~1.5k	policy, value, testimony, fact, reference	support (reason, evidence)
Reed et al. (2008)	AraucariaDB	507	2,842	claim, premise	-
Habernal and Gurevych (2015)	web discourse (WD)	340	3,899	claim, permise, backing, rebuttal refutation	
Biran and Rambow (2011a)	online comments (OC)	2,805	8,946	claim, justification	-
Biran and Rambow (2011b)	wiki talk pages (WTP)	1,985	9,140	claim, justification	-
Hidey et al. (2017)	reddit (CMV)	78	3,500	claim: interpret., eval., (dis)-agree; premise: logos, pathos, ethos	-
Habernal and Gurevych (2016)	debate websites (UKPConvArg)	32 topics	16k pairs	-	-

Argumentation Datasets

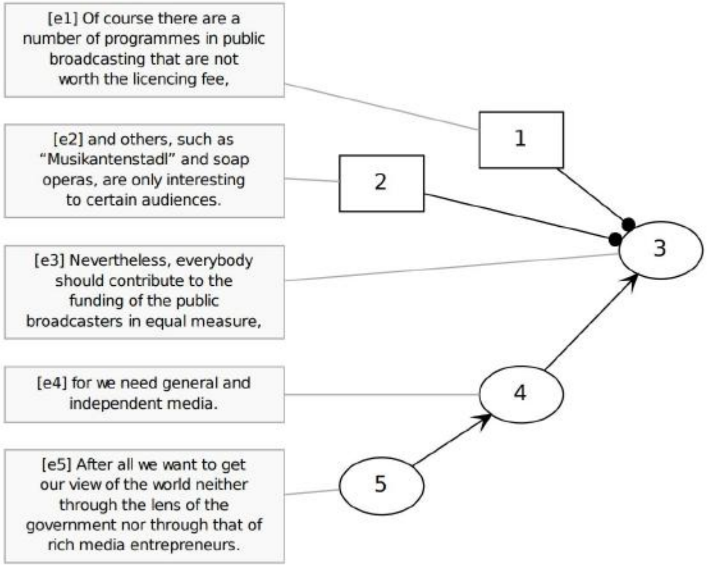
Dataset	Genre	Docs	Sent	Units	Relations
Peldszus and Stede (2015)	microtext (MT)	112	449	claim, premise	support, attack (rebuttal, undercut)
Stab and Gurevych (2017)	persuasive essays (PE)	402	7,116	major claim, claim, premise	support, attack
Niculae et al. (2017)	web discourse, eRuleMaking (CDCP)	731	~1.5k	policy, value, testimony, fact, reference	support (reason, evidence)
Reed et al. (2008)	AraucariaDB	507	2,842	claim, premise	-
Habernal and Gurevych (2015)	web discourse (WD)	340	3,899	claim, permise, backing, rebuttal refutation	
Biran and Rambow (2011a)	online comments (OC)	2,805	8,946	claim, justification	-
Biran and Rambow (2011b)	wiki talk pages (WTP)	1,985	9,140	claim, justification	-
Hidey et al. (2017)	reddit (CMV)	78	3,500	<u>claim</u>: interpret., eval., (dis)-agree; <u>premise</u>: logos, pathos, ethos	-
Habernal and Gurevych (2016)	debate websites (UKPConvArg)	32 topics	16k pairs	-	-

Argument Structure

Peldszus and Stede (2015)

Goal: unit-type, link, and link-type prediction
Data: German, English-translated micro essays
Method: Logistic Regression, MST
Features:
lemma, syntactic, discourse, structural
of segment pair (and context)

(+) Joint prediction of units and links
(-) Individual modeling of sub-tasks
English version is translated
Needs segmented text



Andreas Peldszus and Manfred Stede. "Joint prediction in MST-style discourse parsing for argumentation mining."
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 938-948. 2015.

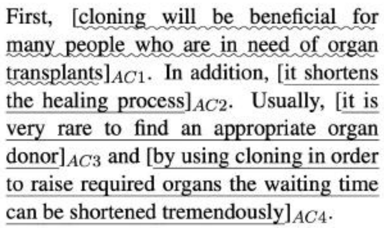
Argument Structure

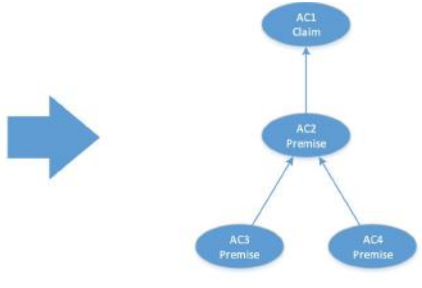
Potash et al. (2017)

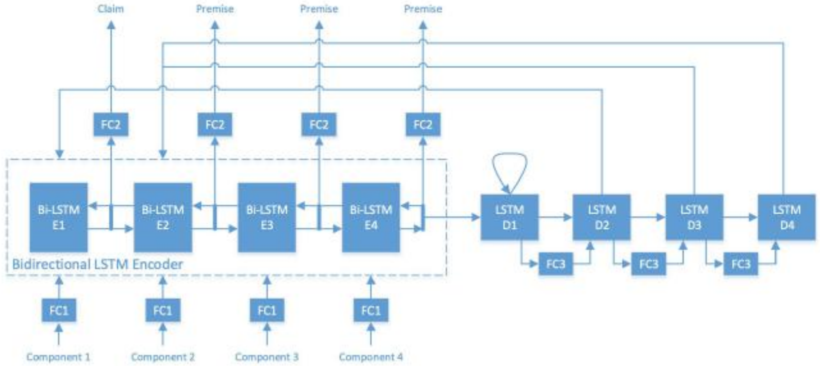
Goal: <u>unit-type</u> and <u>link</u> prediction
Data: essays (persuasive, and micro)
Method: Pointer Networks
Features: n-grams, Glove, structural

(+) Joint modeling and prediction of sub-tasks
Works well on two corpora

(-) No support for domain-specific constraints
Needs segmented text
No link-type prediction


First, [cloning will be beneficial for many people who are in need of organ transplants]$_{AC1}$. In addition, [it shortens the healing process]$_{AC2}$. Usually, [it is very rare to find an appropriate organ donor]$_{AC3}$ and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]$_{AC4}$.





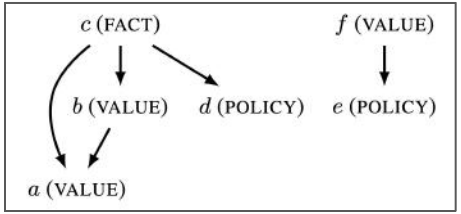
Peter Potash, Alexey Romanov, and Anna Rumshisky. "Here's My Point: Joint Pointer Architecture for Argument Mining."
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing

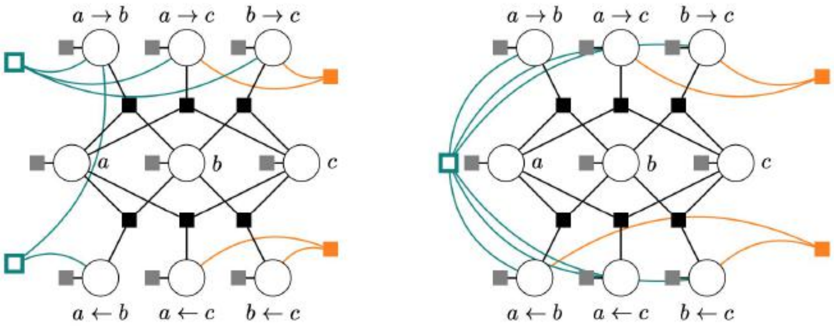
Argument Structure

Niculae et al. (2017)

Goal: unit-type and link prediction
Data: web text (user comments on proposals)
persuasive essays
Method: factor graphs in SVM and RNN

(+) Scheme has subtypes for support
(reason, evidence)
No tree-structure constraints
(-) Scheme has no attack relations
Imbalance links are difficult to handle by
SVM-overgenerates, RNN-undergenerates





(a) CDCP (b) UKP

Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured SVMs and RNNs.
In Proceedings of the 2017 Association for Computational Linguistics (Volume 1: Long Papers), pages 985– 995, 2017.

End to End Modeling of Argument

Persing and Ng (2016)

Goal: unit, unit-type, and link-type prediction
Data: persuasive essays
Method: Rules and Max Entropy classifier,
Joint prediction using ILP
Features: structural, lexical, syntactic, indicator

(+) End-to-end pipeline
Joint-inference to handle error propagation
(-) Rules, ILP constraints are corpus-specific
Tasks learned individually
Handcrafted features

(a) Potential left boundary locations

#	Rule
1	Exactly where the S node begins.
2	After an initial explicit connective, or if the connective is immediately followed by a comma, after the comma.
3	After nth comma that is an immediate child of the S node.
4	After nth comma.

(b) Potential right boundary locations

#	Rule
5	Exactly where the S node ends, or if S ends in a punctuation, immediately before the punctuation.
6	If the S node ends in a (possibly nested) SBAR node, immediately before the nth shallowest SBAR.[1]
7	If the S node ends in a (possibly nested) PP node, immediately before the nth shallowest PP.

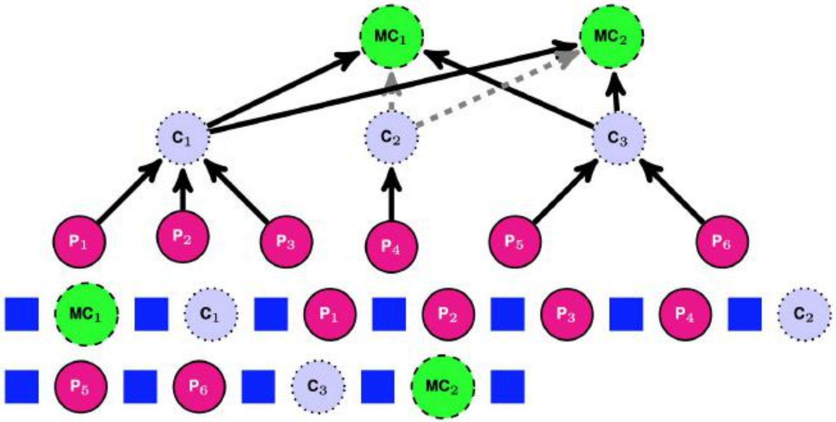
Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays.
In Proceedings of the North American Chapter of the Association for Computational Linguistics, pages 1384–1394, 2016.

End to End Modeling of Argument

Eger et al. (2017)

Goal: unit, unit-type, and link-type prediction
Data: persuasive essays
Method: BiLSTM-CRF-CNN tagger,
TreeLSTM tagger
Features: Glove embeddings, syntactic

(+) End-to-end neural tagger at the token level
Decoupling but joint learning of sub-tasks
(-) Predicts a lot of relations within a sentence
barely exists in the corpus



Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. "[Neural End-to-End Learning for Computational Argumentation Mining.]()"
In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 11-22. 2017.

Argument Structure Recap

Schemes, Genres, Tasks, and Approaches

Scheme

Units MT: claim, premise
PE: major claim, claim, premise
CDCP: policy, value, testimony, fact, reference

Links MT: support, attack (rebuttal, undercut)
PE: support, attack
CDCP: support (reason, evidence)

Genre

Essays: Peldszus and Stede (2015), Potash et al. (2017), Persing and Ng (2016), Eger et al. (2017)

Essays and Web Discourse: Niculae et al. (2017)

Task

Unit-Type, Link, Link-Type: Peldszus and Stede (2015)

Unit-Type, Link: Potash et al. (2017), Niculae et al. (2017)

End2End: Persing and Ng (2016), Eger et al. (2017)

Approach

MST: Peldszus and Stede (2015)

Pointer Network: Potash et al. (2017)

Factor Graphs: Niculae et al. (2017)

ILP: Persing and Ng (2016)

BiLSTM-CRF Tagger: Eger et al. (2017)

Argument Structure Recap

Schemes, Genres, Tasks, and Approaches

Scheme

Units MT: claim, premise
PE: major cl
CDCP: polic
Links MT: support,
PE: support,
CDCP: supp

Genre

Essays: Peldszus and Stede (2015), Potash et al. (2017), Persing and Ng (2016), Eger et al. (2017)
Essays and Web Discourse: Niculae et al. (2017)

Still infeasible to extract full argument structure automatically across domains/genres

But! Some of the sub-tasks can be extracted across domains

Task

Unit Type, Link, Link Type: Peldszus and Stede (2015)
iculae et al. (2017)
et al. (2017)

ILP: Persing and Ng (2016)
BiLSTM-CRF Tagger: Eger et al. (2017)

Argumentation

Argument Structure

Peldszus and Stede (2015)
Potash et al. (2017)
Niculae et al. (2017)

Persing and Ng (2016)
Eger et al. (2017)

Claim Detection, Argument Semantics

Daxenberger et al. (2017)
Chakrabarty et al. (2019)

Hidey et al. (2017)
Wachsmuth et al. (2017)

Argumentation

Argument Structure

Peldszus and Stede (2015)
Potash et al. (2017)
Niculae et al. (2017)

Persing and Ng (2016)
Eger et al. (2017)

Claim Detection, Argument Semantics

Daxenberger et al. (2017)
Chakrabarty et al. (2019)

Hidey et al. (2017)
Wachsmuth et al. (2017)

Claim Detection

Daxenberger et al. (2017)

Goal: Cross-domain claim detection
Data: 6 datasets (essays, web discourse)
Method: CNN, LSTM, LogReg
Features:
structural, lexical, syntactic, discourse
word2vec embeddings

(+) Extensive experiments and ablation studies
Testing generalizability on six datasets
Qualitative analysis of what a claim is
(-) Not including contextual information

OC: single word "Bastard."
emotional expressions "::hugs:: i am so sorry hon ..")

WTP: Wikipedia quality discussions
"That is why this article has NPOV issues."

MT: use of 'should'
"The death penalty should be abandoned everywhere."

PE: signaling beliefs "In my opinion, although using machines have many benefits, we cannot ignore its negative effects."

AraucariaDB: statements starting with a discourse marker, legal-specific claims, reported and direct speech claims

WD: controversy "I regard single sex education as bad."

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. "What is the Essence of a Claim? Cross-Domain Claim Identification." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2055-2066. 2017.

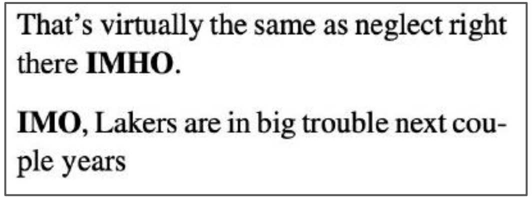
Claim Detection

Chakrabarty et al. (2019)

Goal: Cross domain claim detection
Data: 4 datasets (essays, blogs, reddit)
Method:
Fine-tuning ULMFiT on a larger unsupervised data relevant to the target corpus

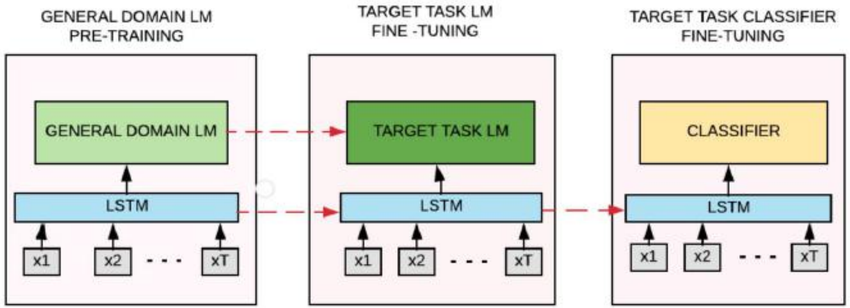
(+) Utilization of pretrained models
Utilization of self-labeled data

(-) 'IMHO' is specific to this problem



That's virtually the same as neglect right there **IMHO**.

IMO, Lakers are in big trouble next cou-ple years



Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. "IMHO Fine-Tuning Improves Claim Detection." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1 (Long and Short Papers)*, pp. 558-563. 2019.

Semantic Types of Claims and Premises

Hidey et al. (2017)

Goal: Annotation scheme for semantic types of claims and premises

Data: reddit (ChangeMyView)

Method:

Argument structure annotations (experts)

Semantic types annotations (crowdsource)

(+) A corpus with claim and premise subtypes

(-) No annotation of relation types

A CMV: Patriotism is the belief that being born on one side of a line makes you better

...

↑

B [I would define patriotism quite simply as supporting one's country, but not *necessarily* disparaging others] CLAIM$_{\text{DISAGREEMENT}}$

...

[Someone who assists another country that is in worse shape instead of assisting their own can still be a patriot, but also recognize significant need in other nations and decide to assist them as well] PREMISE$_{\text{LOGOS/PATHOS}}$

↑

A [This is true]CLAIM$_{\text{AGREEMENT}}$, but, [I think, supporting the common good is also more important than supporting your country]CLAIM$_{\text{RATIONAL EVALUATION}}$

↑

B [Yes]CLAIM$_{\text{AGREEMENT}}$, but [the two are often one the same]CLAIM$_{\text{INTERPRETATION}}$, [especially when you live in a country as large as the U.S. most acts which serve the common good generally support your country]PREMISE$_{\text{LOGOS}}$.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. "Analyzing the semantic types of claims and premises in an online persuasive forum." In *Proceedings of the 4th Workshop on Argument Mining*, pp. 11-21. 2017.

Argument Quality

Wachsmuth et al. (2017)

Goal: Theory vs Practice
of argument quality assessment
Data: Debate portals
Method:
Correlation Analysis of absolute expert ratings and crowdsourced relative ones

(+) Bridging the theory-practice gap
Evaluating the applicability of theory
Evaluating the need for expert annotators
(-) Using correlation analysis on one corpus

Quality Dimension	Short Description of Dimension
Cogency	Argument has (locally) acceptable, relevant, and sufficient premises.
Local acceptability	Premises worthy of being believed.
Local relevance	Premises support/attack conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	Argument persuades audience.
Credibility	Makes author worthy of credence.
Emotional appeal	Makes audience open to arguments.
Clarity	Avoids deviation fr[illegible] correct and unambi[illegible]
Appropriateness	Language proportic[illegible] supports credibility
Arrangement	Argues in the right [illegible]
Reasonableness	Argument is (globa[illegible] relevant, and suffici[illegible]
Global acceptability	Audience accepts u[illegible]
Global relevance	Argument helps arr[illegible]
Global sufficiency	Enough rebuttal of [illegible]
Overall quality	Argumentation qua[illegible]

Polarity	Label	Short Description of Reason
Negative properties of Argument B	5-1	*B* is attacking / abusive.
	5-2	*B* has language/grammar issues, or uses humour or sarcasm.
	5-3	*B* is unclear / hard to follow.
	6-1	*B* has no credible evidence / no facts.
	6-2	*B* has less or insufficient reasoning.
	6-3	*B* uses irrelevant reasons.
	7-1	*B* is only an opinion / a rant.
	7-2	*B* is non-sense / confusing.
	7-3	*B* does not address the topic.
	7-4	*B* is generally weak / vague.
Positive properties of Argument A	8-1	*A* has more details/facts/examples, has better reasoning / is deeper.
	8-4	*A* is objective / discusses other views.
	8-5	*A* is more credible / confident.
	9-1	*A* is clear / crisp / well-written.
	9-2	*A* sticks to the topic.
	9-3	*A* makes you think.
	9-4	*A* is well thought through / smart.
Overall	**Conv**	*A* is more convincing than *B*.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. "Argumentation quality assessment: Theory vs. practice." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 250-255. 2017.

Conclusions

Claim Detection

Daxenberger et al. (2017)

1. 'Claim' conceptualization is different, but, has some shared lexical properties
2. Choice of training data is crucial especially when target is unknown

Chakrabarty et al. (2019)

Fine-tuning language models on relevant unlabeled data is important for cross-domain claim detection

Semantics of an Argument

Hidey et al. (2017)

1. Semantic types of claims are premises can be annotated by non-experts
2. Analyzing semantic types is useful in modeling argument persuasion

Wachsmuth et al. (2017)

1. Comparison metrics are easier in practice
2. Simplifying theory to capture the most important reasons in practice improves its applicability

Argumentation for Fact-Checking (Micro)

How can we use argumentation for misinformation detection?

- **<u>Given a claim find supportive/opposing sentences in the text.</u>**
  This could be used for evidence retrieval in Fact-checking
  - Rather than selecting sentences first then modeling entailment
  - Current joint models do not look at context

- **<u>Factual Claim Detection (what to fact-check)</u>**
  - Looking at sentence alone to decide whether they should be fact-checked
  - Looking at argument structure to find dangling claims

Argumentation for Fake News & Stance Detection

How can we use argumentation for misinformation detection?

Argumentative search is used for Stance Retrieval of debates given a topic. (e.g. args.me)

A similar setup for Stance Detection in news?

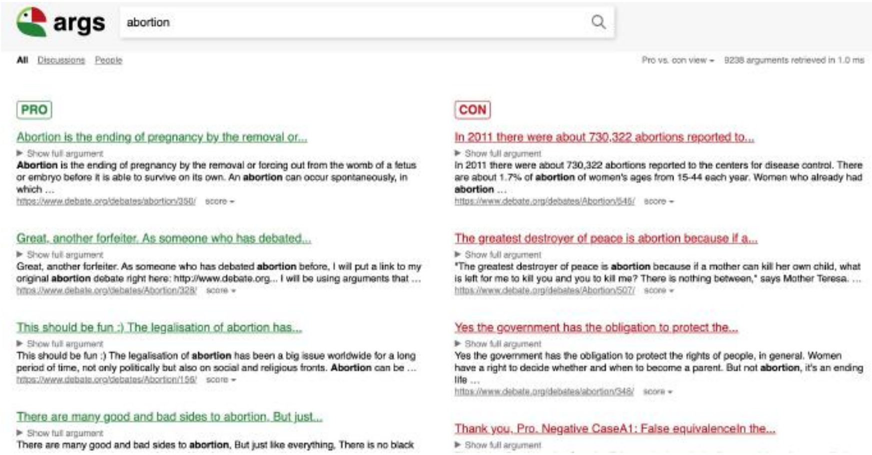
Can argumentation help in the task of predicting truthfulness of a sentence (claim)?

Distinguishes opinion claims vs factual claims

CDCP (Policy, Value) vs (Testimony, Fact)

CMV Evaluation-Emotional vs Evaluation-Rational

Logos vs Pathos



Outline

1. Introduction
2. Fact-Checking
    a. What processes does fact-checking include and can they be automated?
    b. What sources can be used as evidence to fact-check claims?
3. Fake News Detection
    a. What are the linguistic aspects of Fake News? Can it be detected without external sources?
        i. Fake News, Misinformation, Disinformation, Hoax, Satire and Propaganda.
    b. How do we build robust AI models that are resilient against false information?
4. Argumentation
    a. How can we extract an argument structure from unstructured text?
        i. End2end, sub-tasks, claim detection
    b. Semantics of argument units; Argument quality assessment
    c. How can we use argumentation for misinformation detection?

Thank You